GEOCOMPUTATION AND GISCIENCE

Michael F. Goodchild[1] and Paul A. Longley[2]

Abstract

This chapter begins with definitions of geographic information science (GIScience), of geocomputation, and of spatial analysis. We then discuss how these research areas have been influenced by recent developments in computing and data-intensive analysis, before setting out their core organizing principles from a practical perspective. The following section reflects on the key characteristics of geographic information, the problems posed by large data volumes, the relevance of geographic scale, the remit of geographic simulation, and the key achievements of GIScience and geocomputation to date. Our subsequent review of changing scientific practices and the changing problems facing scientists addresses developments in high-performance computing; heightened awareness of the social context of GIS; and the importance of neogeography in providing new data sources, in driving the need for new techniques, and in heightening a human-centric perspective.

1. INTRODUCTION

Geographic information science (GIScience) addresses fundamental issues associated with geographic information and the use of geographic information systems to perform spatial analysis, using a scientific approach (for detailed discussions of the nature of geographic information science see Duckham, Goodchild, and Worboys 2003). The issues may be practical, as in the question of how to address uncertainty in geographic information; they may be empirical, as in the observation generally known as Tobler's First Law of Geography[3] (Tobler 1970); or they may be theoretical, as in the fundamental contribution known as the 9-intersection of topology[4] (Egenhofer and Franzosa 1991). To some, the term implies the use of geographic information systems (GIS) as a scientific tool in research and decision-making, and as such it has been widely applied to the solution of virtually any problem that is embedded in geographic space, from global warming to crime and water pollution. Much progress has been made in GIScience in the two decades since the term was coined (Goodchild 1992), through the efforts of a growing scientific community. It is also important to note that other terms convey similar meaning, including geomatics, geoinformatics, spatial data science, and spatial information science; and that GIScience plays an important role in the practice of regional science, as both a technology that can support research, and as an approach to problem-solving.

[1] Center for Spatial Studies and Department of Geography, University of California, Santa Barbara, CA 93106-4060, USA. good@geog.ucsb.edu

[2] Consumer Data Research Centre and Department of Geography, University College London, Gower Street, London WC1E 6BT, UK. p.longley@ucl.ac.uk

[3] "All things are related, but nearby things are more related than distant things"

[4] Briefly, the set of topologically distinct relationships that can exist between two areas in the plane.

Geocomputation is also fundamentally concerned with geographic information, in other words information about features and phenomena and their locations on or near the Earth's surface. Coined a little later by Openshaw and Abrahart (1996), the term is often used in cross-sectional analysis to describe the repeated analysis and simulation of spatial distributions, in order to explore spatial distributions and to draw inferences about the processes that created and govern them. More specifically, the term is often taken to imply simulation of processes operating in the geographic domain, and thus with geographic information that captures evidence of those processes and is thus primarily dynamic. The major issues in geocomputation often center on the computational problems that arise in simulating complex systems with massive numbers of features, data items, or agents. In this sense geocomputation develops an application-led focus upon the way the world *works*, founded upon rich digital representations of the way that the world *looks*, and makes prediction a central goal. The main contribution of geocomputation may thus lie in the development of better tools for dealing with complex, dynamic systems and for predicting their future states.

From these definitions it is clear that GIScience and geocomputation have much in common, that their interests overlap substantially, and that it may even be helpful to think of geocomputation as a computationally intensive, application-led component of GIScience. Accordingly, the focus of this chapter is on the common ground between them, using the terms somewhat interchangeably. The term GIScience is used wherever the context seems to demand it, and similarly with the term geocomputation. Both terms are fundamentally concerned with spatial analysis (or recently spatial analytics), defined as the set of methods whose results change in response to changes in the locations of the objects being analyzed, and we sometimes use this umbrella term. The remainder of this section elaborates on the basic definition of GIScience and the research conducted under its banner. This is followed by a discussion of the basic principles of GIScience; in a nod to geocomputation, the discussion emphasizes those areas where GIScience has been successful at solving computationally intensive problems. Major methods of analysis are reviewed.

The third section of the chapter addresses changing practices in GIScience, focusing on the increasing importance of collaboration, on novel and diverse data sources, on the availability of massive computational resources, and on the problems of dealing with uncertainty. Science generally is changing in response to the need to study complex systems and the use of simulation, and this trend is certainly affecting GIScience. The concept of data-intensive science, the so-called *Fourth Paradigm* (Hey, Tansley, and Tolle 2009), has a natural fit to geographic problems and their massive volumes of data, while the meta-issues of generalisation, documentation and provenance are beginning to loom large in a science that is no longer dominated by the individual investigator. The very rapid growth of the discipline of data science in recent years is also challenging many of the traditional approaches to science, and its relationship to GIScience is addressed in this third section.

Finally, the fourth major section speculates on the future, and discusses the co-evolution of GIScience and geocomputation. Future developments are likely to be driven, as in the

past, by trends in data, in computation, and in the society that forms the context for both fields.

While debates about the nature and meaning of science have raged for centuries and will probably never end, the core ideas are clear. First, science seeks laws and principles that can be shown to be valid in the observable world, and are generalizable in the sense that they apply everywhere and at all times. Both of the examples cited earlier – Tobler's First Law and the 9-intersection – are clearly of this nature, and as a theoretical conclusion the 9-intersection not only applies everywhere at all times, but also applies in any imaginable space. Second, science is founded on definitions of terms that are rigorously stated and understood by all scientists. Third, scientific experiments and their results are replicable, being stated in sufficient detail that someone else could expect to obtain them by carrying out an identical experiment. In this context the term *black box* is pejorative, since procedures that are hidden inside a box cannot be described and therefore cannot be replicated. Well-understood principles also apply to the details of reporting, as in the rule that any measurement or numerical result be stated to a precision (number of significant digits) that reflects the accuracy of the measuring device or model. Principles such as these help to define GIScience and geocomputation, and to distinguish them from less rigorous applications of GIS and related technologies.

A distinction is often drawn between *pure* science, or science for the sake of curiosity and the quest for general discoveries, and *applied* science, or science that aims to solve problems in the observable world using scientific methods. The geo- prefix reminds us that the Earth provides a unique laboratory for scientific investigation, and the uniqueness of the places on it often limits the scope for the kinds of controlled experiments that characterize scientific activity in other disciplines. Geographic space is the space of human activity, and most of the problems facing human society are embedded in it, from poverty and hunger to health. Indeed, it is hard sometimes to avoid application in GIScience because the field is inevitably close to the real world, a fact that perhaps accounts for at least some of the passion displayed by its practitioners. Moreover curiosity has often provided the motivation to explore, characterize, and map the geographic world, though the results of such exploration are rarely generalizable in the sense that Newton's Laws of Motion or the Mendeleev periodic table are generalizable.

This pure/applied distinction explains how progress in spatial analysis is measured. On the one hand, the refereed journals in which much successful GIScience research is published, and the presentations at conferences such as the biennial International Symposia on Geographic Information Science, emphasize the purer forms of science, while the emphasis at other conferences, such as the biennial International Conferences on Geocomputation, emphasize how the core organizing principles and concepts of GIScience can be brought to bear on solving practical problems. A large industry, valued according to some estimates at $20 billion annually (Longley et al. 2015), has sprung up around the data acquisitions and tools needed in such practical problem-solving. Clearly the metrics of success here are much more diverse than in pure science.

2. PRINCIPLES OF GISCIENCE

In this section we describe some of the major achievements of GIScience in its first two decades. The selection includes advances that closely resemble geocomputation in the sense of being concerned with large, complex systems and with large volumes of data. We begin with a discussion of the characteristics that distinguish geographic information and geographic problem solving from data-driven science in other domains. We then discuss the strategies that have been adopted in GIScience for avoiding or successfully dealing with the problems of large data volumes, including aggregation, divide-and-conquer, and compression. We discuss some of the unintended consequences of such strategies, in the form of uncertainty, the ecological fallacy, and the modifiable areal unit problem. We elaborate on the nature of simulation in geographic space, on some of the more successful research conducted in this area, and on some of the issues it raises. Finally, we present a brief summary of progress in GIScience in the past 20 years.

2.1 The characteristics of geographic information

One of the first attempts to identify the special characteristics of geographic information, or "What is special about spatial?", was made by Anselin (1989). He argued that two characteristics were universal: spatial dependence and spatial heterogeneity. Reference has already been made to the first, in the form of Tobler's First Law of Geography: "All things are similar, but nearby things are more similar than distant things." The statement may appear informal and vague, but it is readily formalized in the principles of regionalized variables that underlie the science of geostatistics (Chilès and Delfiner 2012), and in the models widely used in spatial statistics (Cressie 1993). While we can argue about whether the statement meets the criteria for a law as that term is normally understood by philosophers of science, and whether exceptions should be allowed, it is clear that the vast majority of phenomena distributed over the Earth's surface and near-surface adhere to it, while differing in precisely how similarity decays with distance. Moreover there is no doubt of the law's efficacy in GIS.

The principle is essentially one of context, since it requires a phenomenon at one point to be consistent with the same phenomenon at nearby points. It appears to apply well in three-dimensional space and also to apply in four-dimensional space-time. Perhaps the easiest way to demonstrate its validity is by a thought experiment in which it is not true, where a minute displacement on the Earth's surface produces a completely independent environment – clearly this does not happen and cannot happen, though there are many examples where a displacement of at least a finite amount produces an apparently independent environment (such a minimal displacement is known in geostatistics as the *range* of the phenomenon, and synonyms exist in many domains of science).

As a cornerstone of GIScience the principle has two major implications. First, similarity over short distances allows the Earth's surface to be divided into regions within which phenomena are approximately homogeneous, achieving great economies in data volume by expressing attributes as properties of entire areas rather than of individual points. In short, the principle enables the assumed-homogeneous polygons that dominate many representations in GIS. Similarly, it allows reasonable guesses to be made of the

properties of places that have not been visited or measured, in a process known as spatial interpolation. The principle thus justifies the techniques that are used, for example, to create weather maps from scattered point observations.

Unfortunately the principle of spatial dependence also provides a major headache for researchers working with geographic information, since it runs counter to the assumption made in many statistical tests that the data were acquired through a process of random and independent sampling from a parent population. An analysis of the 58 counties of California, for example, cannot make that assumption since the principle implies that conditions in neighboring counties will be similar. Moreover there is no larger universe of which the set of all counties of California constitute a random sample.

Anselin's second principle addresses spatial heterogeneity, or the tendency for parts of the Earth's surface to be distinct from one another. This also has profound implications. Consider, for example, a local agency seeking to define a taxonomy of local land use. The result will inevitably be different depending on the agency's location and the local conditions in its jurisdiction, and every jurisdiction will argue that its scheme is better than any global or national standard. In early geodesy, the figure of the Earth (the mathematical function used to approximate the Earth's shape and thus define latitude and longitude) was unique to each jurisdiction or region, and it was not until the 1960s that pressure for a single standard prevailed, driven by the growing importance of air travel and the targeting of intercontinental ballistic missiles. Unfortunately any universal standard will inevitably be sub-optimal for any local jurisdiction, whether it be over land-use classification or the shape of the Earth, so there will always be tension between the desire to be locally optimal and the desire to be globally universal.

2.2 Dealing with large data volumes

The previous section was concerned with principles that can be demonstrated to be empirically true. We now move to a discussion of some of the principles that guide the design of GIS technology, and allow GIS to deal with problems that might otherwise be overwhelmingly voluminous, a key issue in geocomputation given its goal of addressing large problems. The Earth's surface has approximately 500 million sq km, and a description of it at a resolution of 1 sq m would therefore create 500 trillion data elements if no strategy were adopted to reduce the volume. Even allocating a single byte to each data element would create half a petabyte of data.

In the previous section we discussed Tobler's First Law, the basis for aggregating data elements into statements about entire polygons[5]. California's land area amounts to 403,800 sq km, and describing each sq m with a two-byte designation CA would produce roughly 0.8 terabytes of data. But capturing the coordinates of its boundary and adding a single attribute CA to the polygon could clearly compress this to only a few kilobytes, even with precise coordinates; and by recording only a single attribute, would avoid the

[5] Because the Earth's surface is curved the shortest path between two points is never a straight line. Thus the use of the term "polygon" implies that some method has been used to project the Earth onto a flat surface.

potential for error in the vast number of identical attributes that would have to be recorded in a cell-by-cell approach. Alternatively, a variety of compression techniques can also be used to replace cells of individual data elements with a series of <run length, value> pairs. Many other methods of compression, generalization, and abstraction have been devised to deal with the volume problem, some of them *lossy* in the sense that the result is only approximately identical and the original data cannot be recovered from the compressed version, and some of them *loss-less*.

In a *divide-and-conquer* strategy a geographic area is partitioned, and analysis or modeling proceeds one partition at a time. The term *tile* is often used for partition, especially where the partitions are rectangular[6]. Instead of solving a problem for the whole of California, for example, one might solve it separately for each of its counties. Interactions exist between counties in almost every application: in analyzing water pollution, for example, the actions of a county will influence the water quality in any downstream county, and air pollution will travel to any counties downwind. Thus a successful divide-and-conquer strategy must also consider the degree to which counties interact, and include this in the model, often by iterating between modeling within-county effects and modeling between-county effects. Nevertheless the overall computational efficiency of the modeling will probably be improved by adopting this strategy. Many GIS algorithms make explicit use of divide-and-conquer, as an approach to handling the vast amounts of data provided by satellite-based remote sensing, and implicit divide-and-conquer has been an intrinsic part of human problem-solving from time immemorial.

Despite these traditional strategies, we note in the next major section that massive improvements in computing capacity, along with the increasing availability of fine-resolution data on both environmental and social phenomena, is opening a host of new possibilities. It is increasingly possible to avoid the constraints of divide-and-conquer and to study processes at previously unheard-of resolution.

2.3 Scale-related issues

The term *scale* is often used in GIScience in the sense of spatial resolution, to distinguish between fine-scale or detailed data and coarse-scale or generalized data. Some of the techniques described in the previous section essentially sacrifice scale in the interests of reducing data volume. To a cartographer, reducing a map's *representative fraction*, the ratio of distance on the map to distance on the Earth[7], is similarly a sacrifice of scale, often in the interests of visual clarity. To a compiler of social statistics, reporting counts of people based on large, aggregated reporting zones may also be a means of reducing data volume.

All of these techniques have consequences that are well recognized in GIScience. The *modifiable areal unit problem* refers to the effects that changes in reporting zone

[6] As before, note that tiles cannot be rectangular on a curved surface, and that the Earth must first be projected to a plane

[7] Note that no flat map can have a precisely constant representative fraction relative to the curved surface of the Earth

boundaries will have on the results of any geographic analysis. The term was first formally characterized by Openshaw (1983), who demonstrated that changing reporting-zone boundaries could produce dramatic swings in results, even when holding scale constant. His solution, which became a fundamental tenet of geocomputation, was to recommend exploring the aggregation effect in any specific case, by repeated analysis using different zones. Unfortunately in most cases this can only be done by aggregating predefined zones, producing different results but at a still coarser level of aggregation, since data compiled for different zonal arrangements at the same level of aggregation will usually not be available. Many studies have documented the problem, while others have argued that it results not from a failure of analytic method but from a failure on the part of the investigator to be explicit about the scale at which the hypothesized effects occur. For example, in Openshaw's original case study, the 99 counties of Iowa were used to explore the relationship between percent of the population over 65 and percent registered Republican voters. Aggregating the counties in various ways did indeed produce different results, but at coarser scale. What is missing in this case is a well-defined hypothesis as to why this correlation should appear, and at what scale. Perhaps the process works at the individual level, and older people are more likely to vote Republican, in which case the hypothesis is best tested at the individual level. Or perhaps the process is ecological: a neighborhood with a large percent of people over 65 also attracts a large percent of Republican voters, whether or not they are over 65. In the latter case the appropriate scale of analysis is that of the neighborhood, requiring a formal definition of that concept and an aggregation of finer-scale data, such as block-group data, to the neighborhood level. The general point is relevant to the definition of spatial analysis in Section 1, and is that we should not be looking for statistics that are invariant to the phenomenon that we wish to study. As such, the MAUP is not an empirical problem but rather is a theoretical requirement to hone statistics to the geographic context in which they are applied.

A closely related problem, also well-recognized in GIScience, is the ecological fallacy, the fallacy of reasoning from the aggregate to the individual. The fallacy already appeared in the previous paragraph, since it would be wrong to infer from a county-level correlation that individuals over 65 tend to vote Republican – in fact, in the extreme, Openshaw's correlations could exist in Iowa at the county level even though no person over 65 was a registered Republican. King (1997) reviews the problem in greater detail and suggests ways of addressing it. Other approaches to down-scaling, or replacement of coarse-scale data by fine-scale data, can be found, such as the work of Boucher and Kyriakidis (2006) in the context of remote sensing.

2.4 Simulation in GIScience

Many processes that operate on the Earth's surface can be abstracted in the form of simple rules. One might hypothesize, for example, that consumers always purchase groceries from the store that can be reached in minimum time from their homes. Exactly how such hypotheses play out in the real world can be difficult to predict, however, because of the basic heterogeneity and complexity of the Earth's surface. Christaller was able to show that such simple assumptions about behavior led to simple patterns of settlements in areas dominated by agriculture, but only by assuming a perfectly uniform

plane. Similarly, Davis was able to theorize about the development of topography through the process of erosion, but only by assuming a starting condition of a flat, uplifted block. Research in both areas has clearly demonstrated that the perfect theoretical patterns predicted never arise in practice.

One strategy for addressing such issues is to assume that in the infinite complexity of the real world, all patterns are equally likely to emerge; and that the properties we will observe will be those that are most likely. This strategy enabled Wilson (1970) to show that the most likely form of distance decay in human interaction was the negative exponential; and Shreve (1966) was able to show that the effect of random development of stream networks would be the laws previously observed by Horton. Similar approaches have been applied to the statistical distribution of city size, or the patterning of urban form (Batty and Longley 1994).

Nevertheless, while they yield results that are often strikingly in agreement with reality, such approaches lack the practical value that real-world decision-making demands. Instead, GIScience and geocomputation are increasingly being used to simulate the effects of simple hypotheses about behavior on the complex landscapes presented by the geographic world. The generality of such approaches lies in the hypotheses they make about behavior; the landscapes they address, and the patterns they produce, are essentially unique.

Such approaches fall into two major categories, depending on how the hypotheses about behavior are expressed. The approach of *cellular automata* begins with a representation of the landscape as a raster, and implements a set of rules about the conditions in any cell of the raster. The approach was originally popularized by Conway in his Game of Life, in which he was able to show that distinct patterns emerged through the playing out of simple rules on a uniform landscape. Such patterns are known as *emergent properties*, since they would be virtually impossible to predict through mathematical analysis. The cellular-automata approach has been used by Clarke (e.g., Clarke and Gaydos 1998) and others to simulate urban growth, based on simple rules that govern whether or not a cell will change state from undeveloped to developed. Such approaches allow for the testing of policy options, expressed in the form of modifications to the rules or to the landscape, and have been widely adopted by urban planners.

The alternative approach centers on the concept of *agent*, an entity that is able to move across the geographic landscape and behave according to specified rules. This *agent-based* approach is thus somewhat distinct from the cell-based approach of cellular automata. Agent-based models have been widely implemented in GIScience and geocomputation. For example, Torrens, Li, and Griffin (2011) have studied the behavior of crowds using simple rules of individual behavior, with applications in the management of large crowds with their potential for panic and mass injury. Evans and Kelley (2004) have studied the behavior of decision-makers in their role in the evolution of rural landscapes, and examined policies that may lead to less fragmentation of land cover, and thus greater sustainability of wildlife. Maguire, Batty, and Goodchild (2005) discuss

several other examples of cellular automata and agent-based models in GIScience and geocomputation.

Both approaches raise a number of issues (for a general discussion of these issues see, for example, Parker et al. 2003). From an epistemological perspective, several authors have explored the role of such modeling efforts in advancing scientific knowledge. On the one hand, a model is only as good as the rules and hypotheses about behavior on which it is based. It is unlikely that the results of simulation will lead directly to a modification of the rules, and more likely that rules will be improved through controlled experiments outside the context of the modeling. If patterns emerge that were unexpected, one might argue that scientific knowledge has advanced, but on the other hand such patterns may be due to the specific details of the modeling, and may not replicate anything that actually happens in the real world.

Validation and verification of simulation models are always problematic, since the results purport to represent a future that is still to come. *Hindcasting* is a useful technique, in which the model is used to predict what is already part of the historic record, usually by working forward from some time in the past. But the predictions of the model will never replicate reality perfectly, forcing the investigator to ask what level of error in prediction is acceptable, and what unacceptable. Moreover it is possible and indeed likely that rules and hypotheses about social behavior that drive the model will change in the future. In that regard models of physical processes may be more reliable than models of social processes.

2.5 Achievements of GIScience

As we noted earlier, the term *GIScience* was coined in a 1992 paper (Goodchild 1992). In some ways the paper was a reaction to comments being made in the literature about the significance of GIS: that it was little more than a tool and did not therefore deserve a place in the academy. The funding of the US National Center for Geographic Information and Analysis (NCGIA) in 1988 by the National Science Foundation seemed to indicate a willingness in some quarters to see more in GIS than technique. Nevertheless the tool/science debate continued for some time, and is summarized by Wright, Goodchild, and Proctor (1997).

Two decades later several efforts were made to look back and assess progress. A meeting for that purpose was convened in Santa Barbara in December 2008 (http://ncgia.ucsb.edu/projects/isgis/), and a paper summarizing its results and offering a personal perspective has been published by Goodchild (2010). It draws on the assessments of several individuals, and on a bibliographic analysis performed by Skupin. While any level of consensus is inevitably difficult to achieve, the following might be argued to be the major achievements of two-and-a-half decades of GIScience:

- Clarification and specification of the basic data model, including recognition of the fundamental significance of discrete-object and continuous-field

conceptualizations, the emergence of object-oriented data modeling, and the specification of spatial relations.

- The development of place-based techniques of spatial analysis, including local indicators of spatial association (Anselin 1995, Ord and Getis 1995), spatial regression models (LeSage and Pace 2009), and geographically weighted regression (Fotheringham, Brunsdon, and Charlton 2002).
- The specification of standards for simple features, metadata, real-time interaction across the Internet, and many other aspects of GIS practice, led by the Open Geospatial Consortium and the US Federal Geographic Data Committee.
- The development of digital globes such as Google Earth that allow real-time interaction with three-dimensional models of the Earth.
- Recognition of the importance of ontology, as the key to interoperability across communities, languages, and cultures.
- Search and retrieval based on geographic location, through mechanisms such as the geoportal (Maguire and Longley 2005).
- Advances in geovisualization, going far beyond the capabilities of conventional cartography to include animation, the third spatial dimension, reduction of high-dimensional data sets, and many other topics.
- Achievement of a new level of understanding of uncertainty in geographic information, its handling, and its effects, together with a fundamental shift of focus from accuracy to uncertainty.
- Focus on the source and operation of bias in geographic representation, particularly where Big Data sources are repurposed for research applications that were not anticipated or intended when data were created (Longley, Cheshire, and Singleton 2018). This can be seen as retaining focus upon data collection methods and their suitability for spatial analysis in the Big Data era.

Perhaps more important are the institutional achievements, which can be seen as the indirect result of such advances. GIScience is now widely recognized in the titles of journals and the names of departments and programs. In recent years several GIScientists have been elected to prestigious institutions such as the US National Academy of Sciences and the UK's Royal Society. GIScience conferences have proliferated, and the GIScience bookshelf now contains an impressive array of titles.

3. CHANGING PRACTICE AND CHANGING PROBLEMS

In this section we examine the changing nature of GIScience, and speculate on its future. GIS has always been driven by competing factors. On the one hand, it has been at the mercy of trends and changes within the larger computing industry, including new technologies that may or may not offer significant benefits for GIS. For example, the relational database management systems of the 1970s led to a major breakthrough in data modeling in GIS. GIS has also been driven by the need to solve problems of importance to society, from the resource management that provided the initial applications of GIS in the 1980s to the military applications that have always been important but half-hidden, and new applications in public health that are as yet only partially developed. GIS as a

tool for science is subject to the winds of change that are currently blowing through the scientific community, pushing it towards a more collaborative, multi-disciplinary, and data-centric paradigm. Finally, GIS exists in a social context of concerns about privacy, about scientific practices, and about the role that a sometimes expensive technology can play in empowering the already empowered; and GIS is being influenced by the importance of the average citizen as both a consumer and producer of geographic information.

This section is structured as follows. We begin with a discussion of high-performance computing, and its importance for the kinds of massive simulation models discussed previously. We then move to a discussion of the social context of GIS, and the social critique that emerged in the 1990s and now drives the research of many GIScientists. This is followed by a discussion of the relationship between GIScience and data science. Finally, we examine the phenomenon of *neogeography* and the importance it may hold in providing new data sources and in driving the need for new techniques.

3.1 CyberGIS and parallel processing

A major report of the US National Science Foundation (NSF 2003) proposed the term *cyberinfrastructure* to describe the kinds of computing infrastructure that would be needed to support science in the future. Instead of the lone investigator and the desktop system, the report envisioned a distributed infrastructure that would support widespread collaboration across a range of disciplines, following the notion that science in the future would address complex problems with complementary teams of scientists of varied expertise. The solution of complex, large-scale problems would also require a heavy level of investment in high-performance computing (HPC) with its massively parallel architectures. Parallel architectures have an inherently good fit to the nature of geographic space and its somewhat independent individual and community agents, all of which can be seen as semi-independent decision-makers acting in parallel rather than serially.

A number of authors have argued that geographic research and problem-solving requires a specific form of cyberinfrastructure that addresses several key issues, and have coined the term *cyberGIS* (Wang 2016, Wang and Goodchild 2018). How exactly should the geographic world be partitioned across processors? How should one measure computational intensity as a geographic variable? How should the user interface of an integrated cyberGIS be designed? What types of problems, models, and analyses best justify these new approaches? What incentives will persuade the average GIScientist to engage with cyberGIS, given the initial impression of complexity and inaccessibility, and a high level of personal investment in conventional GIS?

Efforts to parallelize GIS date from the 1990s but were not successful for several reasons. First, parallel computing was expensive at the time, and it was difficult for investigators to justify the cost. Second, parallel computing was rendered inaccessible by the need to reprogram in specialized languages. Third, while it was easy to find examples of geographic problems that involved massive volumes of data, it was harder to find ones

that involved massive computation. Finally, collaborative technologies had not yet advanced to the point where it was possible for widely distributed research teams to work together productively.

Many of these arguments are now moot, however. HPC is widely available, and Cloud and Grid technologies are making the transition from conventional computing almost transparent. The need for collaboration is much stronger, and the kinds of problems that used to be solved by individual investigators are now hard to find. Finally, geocomputation has opened the doors to the kinds of massive computation that HPC is designed to address. Indeed, the most compelling examples of the need for HPC lie in the kinds of agent-based and cellular simulations reviewed in the previous section.

In recent years it has also become possible to parallelize processing on the desktop, following the addition of graphical processing units (GPUs) to graphics boards in order to improve the quality and speed of image rendering. Although an innovation of the computer games market, GPU chips were subsequently adapted to more general-purpose computing: today, Nvidia (which, along with AMD, is the world's largest graphics-card manufacturer) produces chips designed specifically for non-graphics applications, and provides a specialized programming-language architecture for use with them. GPUs outperform traditional computation on a central processing unit (CPU) because a GPU has a higher density of cores and uses a process called streaming to handle a number of operations simultaneously. The result is increased processing speed of computationally intensive algorithms. General-purpose computing on graphics processing units (GPGPU) describes the exploitation of the resources of the GPU for various tasks which might previously have been conducted on a CPU. It has particular advantages for real-time systems where the speed of return of results is fundamental to usability and interaction. Adnan, Longley, and Singleton (2014) describe an application in geocomputational geodemographics, in which *k*-means (a frequently used algorithm in the creation of geodemographic classifications) is enhanced to run in parallel over a GPU. This work exploits the parallel-computing Computer Unified Device Architecture (CUDA), which allows code written in standard C or C++ to be used in GPU processing.

3.2 The social context of GIS

Although the GIS technology that underpins GIScience and geocomputation is an established part of the IT mainstream, there is enduring unease in some academic quarters about the social implications of this technology. Early statements were contained in Pickles' (1993) edited volume *Ground Truth: the Social Implications of Geographic Information Systems*, which remains an enduring statement of concerns built around four principal issues. First, there is the view that GIS technology is used to portray homogeneity rather than representing the needs and views of minorities, and that this arises in part because systems are created and maintained by vested interests in society. The roots to this critique can be traced to a wider debate as to whether the umbrella term GIS is best conceived as a tool or as a science, and is something that can be addressed through clarifying the ontologies and epistemologies of GIScience and geocomputation. Second, there is the view that use of a technological tool such as GIS can never be

inherently neutral, and that GIS is used for ethically questionable purposes, such as surveillance and the gathering of military and industrial intelligence. Web 2.0, discussed below, has begun to address this criticism, since it has gone some way to level the playing field in terms of data access, and enabled participation of a wider cross-section of society in the use of this technology of problem-solving. Moreover, it is difficult to construe the views of the Earth promulgated through services such as Google and Bing as intrinsically privileged, not least if they are open to anyone with access to an Internet browser. Third, there has been a dearth of applications of GIS in *critical* research, and a preoccupation with the quest for analytical solutions rather than establishing the impacts of human agency and social structures upon unique places. The rise of mixed-method approaches to GIS (Cope and Elwood 2009) has gone some way towards addressing these concerns. Finally, there is still a view in some quarters that GI systems and science are inextricably bound to the philosophy and assumptions of the approach to science known as *logical positivism*. This implies that GIScience in particular, and science in general, can never be more than a positivist tool and a normative instrument, and cannot enrich other more critical perspectives in geography. Although still featured in many introductory courses on social science methodologies, this critique is something of a caricature of the positivist methods that pervade scientific investigation more generally.

3.3 GIScience and data science

Data science has become an academic growth industry in recent years, fuelled in part by massive increases in the volume and variety of data that are now available via the Internet, in part by growing practical interest in prediction using the techniques of artificial intelligence, and in part by the belief that generic approaches are available to many of the issues of data handling, among them data mining, data search, data modeling, data curation, data sharing, and data description (Kelleher and Tierney 2018). Academic programs have been instituted in response to what is perceived as a rapidly growing market for data skills.

Traditionally, information has been regarded as inherently more useful than data, based on the understanding that information can be defined as data that are “fit for purpose” (Longley et al. 2015). In that sense a geographic information science is implicitly more sophisticated than a geographic data science. But such semantic quibbles aside, it is clear that data science and geographic information science have much in common and much to learn from each other.

A quick review of the syllabi of courses in data science will reveal that few give much attention to geographic information or to techniques of spatial analysis. In part this appears to be a consequence of the belief that there is nothing “special about spatial”, despite the arguments put forth in Section 2.

More fundamentally, however, there are strong arguments for adopting the approach of data science, and specifically in adopting the mantra “let the data speak for themselves” in addressing problems framed in space and time. It is impossible to measure location perfectly, and many of the attributes commonly processed in GIScience and

geocomputation, such as soil class or vegetation cover type, have an inherent degree of subjectivity and are thus non-replicable. In short, uncertainty is present in all geographic information (Zhang and Goodchild 2002), and few if any geographic data sets give the researcher objective knowledge of the differences between the data set and the real world (Janelle and Goodchild 2018). Instead the user must rely on indirect measures such as map scale to understand the limitations of the data.

Additional data issues arise from the repurposing of data that were never collected with the interests and concerns of researchers and analysts in mind. This is perhaps most evident in the analysis of consumer data, which arise essentially as a by-product of an interaction between a consumer and a consumer-facing organization in the course of supply of goods or services, such as social media. Such data account for a large and increasing real share of all of the data collected about citizens today, yet the absence of monopoly providers of most goods and services means that issues of self-selection and bias plague re-use of such data for research purposes (Longley, Cheshire, and Singleton 2018). Triangulation of such sources to existing framework data sources such as censuses presents one route beyond this impasse – such sources may be less rich, granular, or frequently collected, but may suggest ways in which bias can be accommodated or research refocused.

Prediction has been a major driving force in the growth of data science, and the investments being made in this area by major corporations. Tools of machine learning, including artificial neural nets and deep learning, have been shown to be very effective in making successful predictions from highly voluminous and diverse data sets. Yet despite its practical value, prediction has always taken a back seat in science to explanation and understanding. A trained neural network is difficult to interpret within the kinds of hypothesis testing and theory confirmation that have characterized much of science to date. Moreover it is difficult to see how successful these techniques can be in achieving generalizability across study areas, a key requirement of GIScience and geocomputation.

3.4 Neogeography, wikification, open data, and consumer data

Recent years have seen the re-use of the term *neogeography* to describe the developments in Web mapping technology and spatial data infrastructures that have greatly enhanced our abilities to assemble, share, and interact with geographic information on-line. Allied to this is the increased crowd-sourcing by online communities of *volunteered geographic information* (VGI: Goodchild 2007) and *user generated content* (UGC). As such, neogeography is founded upon the two-way, many-to-many interactions between users and websites that have emerged under Web 2.0, as embodied in projects such as Wikimapia (www.wikimapia.org) and OpenStreetMap (www.openstreetmap.org). Today, Wikimapia contains user-generated entries for more places than are available in any official list of place names, and the term *vernacular region* is used to describe regions which emerge from geocomputational analysis of feeds from social networking sites. OpenStreetMap remains well on the way to creating a free-to-use global map database through assimilation of digitized satellite photographs with GPS tracks supplied by volunteers.

This has converted many new users to the benefits of creating, sharing, and using geographic information, often through *ad hoc* collectives and interest groups. Such sites go some way to alleviating concerns about the social implications of GIS, insofar as participation in the creation and use of GIS databases is not restricted, and the contested nature of place names and other characteristics can be tagged in publicly editable databases. As such, Web 2.0 simultaneously facilitates crowd-sourcing of VGI while making basic GIS functions increasingly accessible to an ever-broader community of users. This creation, maintenance, and distribution of databases has been described as a "wikification of GIS" (Sui 2008).

Geographers have long recognized the importance of *place* in humans' understanding of geography. The average person is familiar with thousands or even tens of thousands of named places and their associations. Yet GIS technology requires named places to be represented either as precisely located points, polylines, or polygons. It is a rare individual who knows even roughly the latitude and longitude of his or her home, yet everyone knows their home's street address and postal code. Recently there has been much interest in a *platial* approach to geographic knowledge, as a more human-centric alternative to the familiar *spatial* approach. New data sources, including social media, provide a rich basis for exploring associations of place.

Official data are also becoming available through renewed pressures for government accountability, and the broader realization that wide availability of data collected by government and pertaining to citizens can lubricate economic growth. The result has been a plethora of open-data initiatives in many developed countries, leading to Web-based dissemination of data relating to many areas of public concern, such as personal health, transport, property prices, and even the weather. The previous section described the increasing re-use of consumer data for research purposes, arising out of initiatives that develop common ground in spatial analysis from shared problems of commerce, government, and academia. Conventional official sources such as censuses of population today account for a very much smaller proportion of the data that are collected about citizens, and there is a sense in which open and consumer data initiatives are playing catch-up – providing researchers and analysts with some facility with which to understand the increasingly diverse and complex social, economic, and demographic milieu that characterizes advanced societies. Despite the hubris that has been generated around open and consumer data initiatives, however, most of the data sources that have been released present extremely partial and disconnected representations of the world. For reasons set out in the discussion of modifiable areal unit effects above, the much more holistic concerns with issues of choice and service delivery, or the localism agenda in general, require linked characteristics at the level of the individual citizen, or at the very least small neighbourhood units. These initiatives bring new analytical focus to anonymisation and data disclosure prevention procedures, which in principle may lead to release of data pertaining to individuals rather than geographic aggregations.

All of this requires clear thinking of issues of spatial resolution (level of detail) and disclosure control that are central to the wider spatial literacy agenda (Janelle and

Goodchild 2011). One consideration that is likely to reignite aspects of the social critique of GIS is that it is unlikely that privacy strictures can ever be absolute. Open and consumer data initiatives are creating the need for a broader policy framework for data that responds to concerns of citizen privacy and confidentiality, while remaining cognizant of the benefits that can accrue though opening up, integrating, and using the contents of government data silos. What level of data degradation is an informed public likely to be happy with, if it can be shown to bring benefits in terms of efficient and effective provision of public and private goods?

A related issue is that empowerment of the many to perform basic (and even advanced) GIS operations brings new challenges to ensure that tools are used efficiently, effectively, and safely. Whether using official statistics or VGI, Web 2.0 can never be more than a partial and technological substitute for understanding of the core organizing principles and concepts of GIScience. These highlight the need to know and specify the basis of inference from the partial representations that are used in GIS to the world at large; yet such information is conspicuous by its absence from many VGI sources.

4. CONCLUSION

In undertaking a wide-ranging review of the achievements of GIScience and geocomputation, this chapter has also set out the principal issues and challenges that face these fields today. Improved computation and the facility to create, concatenate and conflate large datasets will undoubtedly guide the future trajectories of the fields in the short to medium term. Ultimately, though, our focus in this chapter has been upon changes in scientific practice that may appear mundane but are nonetheless profound and far-reaching. Good science is relative to what we have now, and improved understanding of data and their provenance is a necessary precursor to better analysis of spatial distributions in today's data- and computation-rich world.

Ultimately, GIScience and geocomputation are applied sciences of the real world, and in large part will be judged upon the success of their applications. Improved methods and techniques can certainly help, as can ever-greater processing power. Yet the experience of the last 20 years suggests that there are rather few purely technical solutions to substantial real-world problems. The broader challenge is to address the ontologies that govern our conception of real-world phenomena, and to undertake robust appraisal of the provenance of data that are used to represent the world using GIS.

This argues that the practice of GIScience and geocomputation poses fundamental empirical questions that require place or context to be understood as much more than location. Scientific approaches to representing places will undoubtedly benefit from the availability of new data sources and novel applications of existing ones, as well as citizen participation in their creation and maintenance. Yet a further quest for GIScience is to develop explicitly geographical representations of the accumulated effects of historical and cultural processes upon unique places.

References

Adnan M, Longley PA, Singleton AD (2014) Parallel processing architectures of GPU: Applications in geocomputational geodemographics. In Abrahart R, See L (ed) GeoComputation, 2nd edn. Taylor and Francis, London.

Anselin L (1989) What is special about spatial data? Alternative perspectives on spatial data analysis. Technical Paper 89-4. National Center for Georaphic Information and Analysis, Santa Barbara, CA.

Anselin L (1995) Local indicators of spatial association – LISA. Geographical Analysis 27(2): 93–115.

Batty MJ, Longley PA (1994) Fractal cities: A geometry of form and function. Academic Press, San Diego, CA.

Boucher A, Kyriakidis PC (2006) Super-resolution land cover mapping with indicator geostatistics. Remote Sensing of Environment 104(3):264–282.

Chilès JP, Delfiner P (2012) Geostatistics: Modeling spatial uncertainty. 2nd edn. Wiley, Hoboken, NJ.

Clarke KC, Gaydos L (1998) Loose coupling a cellular automaton model and GIS: long-term growth prediction for San Francisco and Washington/Baltimore. International Journal of Geographical Information Science 12(7):699–714.

Cope M, Elwood S (2009) Qualitative GIS: A mixed methods approach. SAGE, Thousand Oaks, CA.

Duckham M, Goodchild MF, Worboys MF (2003) Foundations of geographic information science. Taylor and Francis, New York.

Egenhofer MJ, Franzosa RD (1991) Point-set topological spatial relations. International Journal of Geographical Information Systems 5(2):161–174.

Evans TP, Kelley H (2004) Multi-scale analysis of a household level agent-based model of landcover change. Journal of Environmental Management 72(1-2):57–72.

Fotheringham AS, Brunsdon C, Charlton M (2002) Geographically weighted regression: The analysis of spatially varying relationships. Wiley, Hoboken, NJ.

Goodchild MF (1992) Geographical information science. International Journal of Geographical Information Systems 6(1):31–45.

Goodchild MF (2007) Citizens as sensors: the world of volunteered geography. GeoJournal 69(4):211–221.

Goodchild MF (2010) Twenty years of progress: GIScience in 2010. Journal of Spatial Information Science 1(1):3–20.

Hey AJG, Tansley S, Tolle KM (2009) The fourth paradigm: Data-intensive scientific discovery. Microsoft Research, Redmond, WA.

Janelle DG, Goodchild MF (2011) Concepts, principles, tools, and challenges in spatially integrated social science. In Nyerges TL, McMaster R, Couclelis H (eds) The SAGE handbook of GIS and society. SAGE, Thousand Oaks, CA, pp. 27–45.

Janelle DG, Goodchild MF (2018) Territory, geographic information, and the map. In Wuppuluri S, Doria FA (eds) The Map and the Territory: Exploring the Foundations of Science, Thought and Reality. Springer, Dordrecht, pp. 609–628.

Kelleher JD, Tierney B (2018) Data science. MIT Press, Cambridge, MA.

King G (1997) A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data. Princeton University Press, Princeton, NJ.

LeSage J, Pace RK (2009) Introduction to spatial econometrics. CRC Press, Boca Raton, London and New York.

Longley PA, Cheshire JA, Singleton AD (2018) Consumer data research. UCL Press, London.

Longley PA, Goodchild MF, Maguire DJ, Rhind DW (2015) Geographic information science and systems, 4th edn. Wiley, Hoboken, NJ.

Maguire DJ, Batty MJ, Goodchild MF (eds) (2005) GIS, spatial analysis, and modeling. ESRI Press, Redlands, CA.

Maguire DJ, Longley PA (2005) The emergence of geoportals and their role in spatial data infrastructures. Computers, Environment and Urban Systems 29(1):3–14.

National Science Foundation (2003) Report of the Blue-Ribbon Advisory Panel on Cyberinfrastructure. National Science Foundation, Washington, DC.

Openshaw S (1983) The modifiable areal unit problem. GeoBooks, Norwich, UK.

Openshaw S, Abrahart RJ (1996) Geocomputation. In Abrahart RJ (ed) Proceedings, First International Conference on GeoComputation, University of Leeds, pp. 665–666.

Ord JK, Getis A (1995) Local spatial autocorrelation statistics: Distributional issues and applications. Geographical Analysis 27(4):286-306

Parker DC, Manson SM, Janssen MA, Hoffmann MJ, Deadman P (2003) Multi-agent systems for the simulation of land-use and land-cover change: A review. Annals of the Association of American Geographers 93(2):314–337.

Pickles J (ed) (1993) Ground truth: The social implications of geographic information systems. Guilford Press, New York.

Shreve RL (1966) Statistical law of stream numbers. Journal of Geology 74:17–37.

Sui D (2008) The wikification of GIS and its consequences: Or Angelina Jolie's new tattoo and the future of GIS. Computers, Environment and Urban Systems 32:1–5.

Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. Economic Geography 46(2):234–240.

Torrens PM, Li X, Griffin WA (2011) Building agent-based walking models by machine learning on diverse databases of space-time trajectory samples. Transactions in Geographic Information Science 15(s1):67–94.

Wang S (2016) CyberGIS and spatial data science. GeoJournal 81(6):965-968.

Wang S, Goodchild MF (eds) (2018) CyberGIS for geospatial discovery and innovation. Springer, Dordrecht.Wilson AG (1970) Entropy in urban and regional modelling. Pion, London.

Wright DJ, Goodchild MF, Proctor JD (1997) Demystifying the persistent ambiguity of GIS as 'tool' versus 'science'. Annals of the Association of American Geographers 87(2):346–362.

Zhang JX, Goodchild MF (2002) Uncertainty in Geographical Information. Taylor and Francis, New York.