JOS 2015—27/6/2007——278526

Journal of Official Statistics, Vol. 23, No. 3, 2007, pp. 1–15

The Morris Hansen Lecture 2006 Statistical Perspectives on Spatial Social Science

Michael F. Goodchild

Recent commentators have drawn attention to what appears to be a "spatial turn" in several disciplines, including some of the social sciences, driven in part by advances in the geographic information technologies – geographic information systems, the Global Positioning System, and satellite remote sensing – and in part by an increasing emphasis on place-based analysis and policy formulation. It is possible to identify several general characteristics of geographic data, each of which presents problems in the application of traditional statistical methods. Spatial dependence and spatial heterogeneity both run counter to standard assumptions of statistical methods, yet both are potentially useful properties of geographic data. There are interesting applications of classic problems in statistical geometry, and much attention over the past two decades has been devoted to modeling the uncertainties that are inevitably present in geographic data. The presentation ends with comments and speculation on future directions for the field, including an increasing emphasis on the temporal dimension.

Key words: Geographic information system; place-based analysis; spatial dependence; spatial heterogeneity; statistical geometry.

1. Introduction

The most powerful discoveries of science – from the Second Law of Thermodynamics to the structure of DNA or Quantum Theory – are powerful in part because they apply anywhere in space and time. There would be little value, for example, in a Periodic Table that varied from one place to another, or from one day to another. In this sense science is fundamentally *nomothetic*, concerned with the discovery of general principles that can be applied anywhere, at any time. The study of the unique has its place, of course, and in recent years the discovery of planets outside the Solar System, the liquid lakes of Antarctica, and the genes responsible for certain inherited traits are all important advances. But in a broader context *idiographic* science, or the use of scientific methods to study the unique, is regarded as distinctly second-rate, and terms such as *descriptive* or *anecdotal* can even have pejorative significance.

This clear picture is more nuanced in the social sciences, however. In geography, where the very name of the discipline implies description, the debate between idiographic and nomothetic waxes and wanes, is never entirely absent from the syllabi of courses in

National Center for Geographic Information and Analysis, and Department of Geography, University of California, Santa Barbara, CA 93106-4060, U.S.A. Email: good@geog.ucsb.edu
Acknowledgments: The author thanks the Morris Hansen Lecture Committee for the invitation to give the 2006 Lecture. The Center for Spatially Integrated Social Science is supported by grants from the National Science Foundation.

methodology. As early as the 17th Century the Dutch geographer Varenius clearly distinguished between what he called *general geography*, which included general principles applicable everywhere, and what he called *special geography*, which had to do with the unique properties of places (Goodchild et al. 1999). The nomothetic star probably reached its zenith in the 1960 s with the work of Bunge (1966) and other members of the University of Washington-school, and their search for general principles that might account for patterns of settlements and other phenomena on the Earth's surface. But while the dream of an ideal social science modeled on the physical sciences is still alive, today few social scientists would argue that a purely nomothetic social science is a realistic goal – and others would attack the very notion that general principles can be abstracted from human behavior.

In geography the focus of the debate is naturally on space, and more particularly on the space of the Earth's surface. Idiographic geography is interpreted as implying a focus on the unique properties of *places*, while nomothetic geography searches for principles that apply everywhere *in space*. In general one can imagine an idiographic science based on any set of organizing dimensions, including time or the individual, but the focus of this discussion is on space, for reasons that will become apparent. In short, this article is about the role played by space (and to some extent time) in the current evolution of those sciences that deal with the surface of the Earth, with particular emphasis on the social sciences and on statistical methods.

There is evidence that space has become more important in many social sciences in the past few years – that these sciences are undergoing what has been called a *spatial turn*. In anthropology, for example, Moran, Brondizio, and McCracken (2002) have used spatial analysis and spatial data to study household behavior in relation to land cover changes in the Amazon Basin. In criminology, Cohen and Tita (1999) illustrate the integration of spatial diffusion modeling in the analysis of homicide patterns in Pittsburgh. Economists have invoked space to add a robust theoretical context to the "new economic geography" (Fujita, Krugman, and Venables 1999), and advances in spatial econometrics have provided new insight into the ways in which economic processes operate in geographic space (Anselin, Florax, and Rey 2004). In political science, spatial methods and place-based thinking have invigorated the work of Huckfeldt and Sprague (1995) on the role of communication in electoral processes, of King (1997) on ecological inference, and of Gimple and Schuknecht (2003) on accessibility to voting stations. In public health, researchers have drawn on the spatial theory embedded in methods of cluster detection (Cromley and McLafferty 2002), while in sociology neighborhood effects on social processes have been analyzed by Sampson, Raudenbush, and Earls (1997). Janelle and I (Goodchild and Janelle 2004a) recently published a multidisciplinary collection of such spatially oriented studies under the title *Spatially Integrated Social Science*.

This spatial turn is reflected in changing patterns of publication. My colleagues at the Center for Spatially Integrated Social Science at University of California, Santa Barbara analyzed almost 8,900 articles appearing in social science journals between 1990 and 2001, looking for terms such as "spatial analysis" that indicated a quantitative, spatial approach. The data show that about 1.3% of articles published in 1990 mentioned one or more of the key terms, using as a denominator the total number of articles indexed by six abstracting services. The percentage remained at about 2% until 1998, when it began a

comparatively rapid increase, reaching 3.7% by 2001 (Figure 1; full details of this analysis are available at www.csiss.org/research/litsearch.html).

There appear to be three distinct components to this spatial turn. First, it recognizes the key role that spatial concepts, such as distance, location, proximity, neighborhood, and region, play in human society. Second, it emphasizes research that advances understanding through the analysis of spatial patterns and processes. Third, it invokes powerful principles of spatial thinking and reasoning, recognizing for example that visualization is often an effective way of capturing, developing, and conveying ideas. In this article I explore those aspects of this *spatial social science* that might be of particular interest to the statistical community, and that have interested me in my own research. First, however, the next section addresses the drivers that appear to be responsible at least in part for the spatial turn: rapid developments in spatial technologies, a new range of *place-based* analytic techniques, and increasing interest in turning evidence and theory into policy. This is followed by a review of the distinctive characteristics of geographic data and the problems they pose for the statistical sciences. The third major section looks at some interesting applications of statistical geometry to geographic space, and the final section discusses issues of particular concern to the federal statistical community, and speculates on future directions.

In much of the relevant literature the terms spatial and geographic are used virtually interchangeably. In principle, one might define geographic as a special case of spatial, applying specifically to the geographic domain of the Earth's surface, while spatial might refer to all spaces, including those for example of other planets, the human brain, or the digits of π. Normal practice is more ambiguous, however, since it is common for example to refer to analyses as spatial even when they address the geographic domain specifically.

2. Drivers

The notion of applying digital technology to information about the surface of the Earth dates only from the 1960 s, when the first experiments were made at representing and analyzing maps in computers. Two lines of development emerged: the geographic

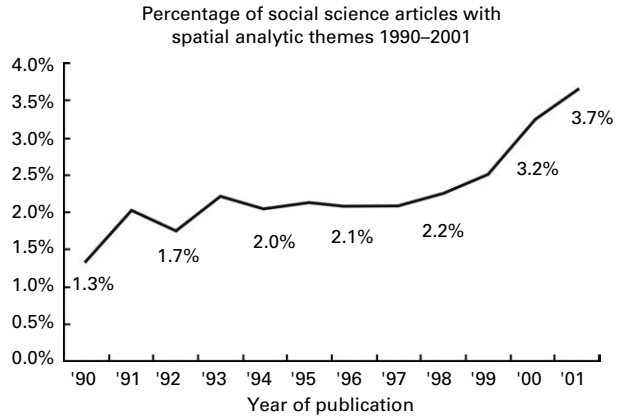

Fig. 1. The growth in the number of published articles in the social sciences that utilize spatial methods

information system (GIS), a class of software designed to manipulate geographic information; and automated cartography, the use of computers to facilitate the compilation of maps. In time these merged, and today a GIS can be defined as a system for performing virtually any conceivable operation – acquisition, editing, analysis, storage, visualization – on geographic data (for a recent introduction to GIS see Longley et al. 2005). Information technology also came to form the basis of remote sensing, which refers to the acquisition of detailed images of the Earth from above, and many other specialized subfields. The Global Positioning System launched in the 1980s offered the first simple means of direct measurement of geographic position, and has become an integral part of in-vehicle navigation systems, cell-phones, and many other devices. Radio-frequency identification (RFID) offers another, cheaper technology for local tracking of the real-time positions of goods, animals, construction components, pets, and even people. Together, these developments mean that it is now almost routine to add geographic coordinates to records on events, transactions, and other forms of human activity, thereby opening up the possibility of mapping, analysis, and record linkage through common geographic location. At the extreme, the concept of a *spatial web* imagines a time when many of the objects on the Earth's surface will be capable of continuously reporting their positions, along with other useful information – a technological marvel but at the same time a nightmarish scenario from many social perspectives.

In recent years the Internet and World Wide Web have revolutionized the role of geographic information in society, and the general public is now increasingly familiar with digital geographic data and remote sensing through sites such as Google Earth. Sir Tim Berners-Lee, the inventor of the Web, recently wrote that "Geographic information is stimulating new uses of the WWW, evolving existing applications and underpinning the creation of new ones to adapt to global trends" (www.ordnancesurvey.co.uk/oswebsite/-media/news/2006/aug/terrafuture.html). Geo-portals such as the U.S. Geospatial One-Stop (www.geodata.gov) offer terabytes of downloadable geographic information free to anyone (Figure 2); geocoding sites allow for the rapid conversion of street addresses to geographic coordinates; Wikipedia is being adapted to allow anyone to post the locations and descriptions of features in their own neighborhood; other sites offer their users the possibility of linking photographs to geographic positions; and Google Earth "mashups" are being created to add three-dimensional, time-dependent geographic context to virtually any information. Google Earth was recently described (in a quote attributed to me) as representing "the democratization of GIS, just as the PC democratized computing in the early 1980s" (*Nature*, February 13, 2006).

These technological developments have made it far easier than ever before for individuals to create maps, to acquire and analyze geographic data, to add geographic coordinates to observations, and to test hypotheses about pattern. There is no doubt that they have an inherent attraction, and not only to the technically minded. Courses in GIS and related technologies are available on almost all campuses, and student interest is high. Bradburn (2004) has written that "The advent of GIS has enabled an explosion of interest in and ability to study the spatial patterns of behavior. GIS not only makes it possible to store in digital form vast amounts of spatial data, it makes possible statistical analysis, modeling, and visual display of geographical data. It provides a powerful new tool that has stimulated new and exciting social science research using geographical concepts and data.

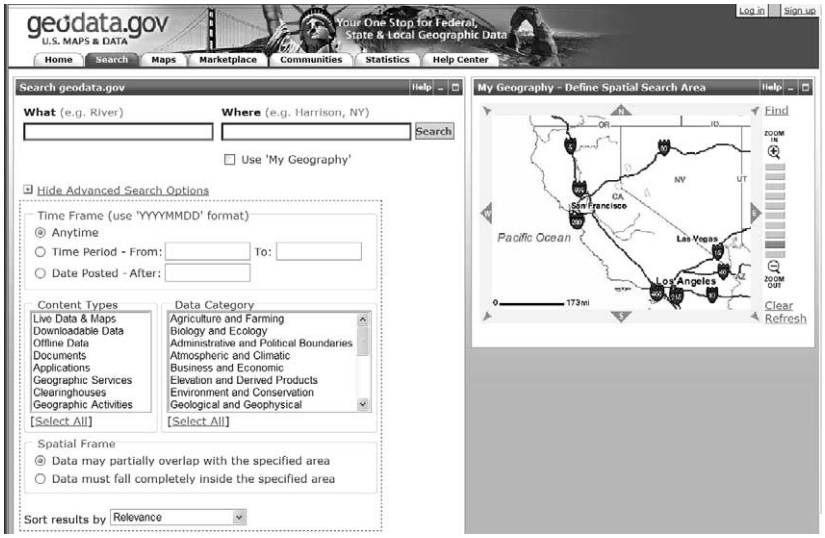

Fig. 2. The Geospatial One-Stop (www.geodata.gov) offers terabytes of geographic information to anyone with an Internet connection

At last, long-held but unverified hypotheses about the importance of locational and spatial variation can be tested. We are at the dawn of a revolution in a spatially oriented social science."

But technology and data are surely not the only drivers of the spatial turn, which is also motivated by more substantive considerations. Traditionally, methods of spatial analysis have been designed to extract knowledge by testing hypotheses (in a deductive context) or by computing readily interpreted indices (in an inductive context). While it is essential that all data submitted to spatial analysis be georeferenced, and techniques only qualify as spatial if their products are not invariant under relocation, the results of such analyses are traditionally abstracted from space and time, satisfying the expectations of nomothetic science. But recently a new suite of techniques has been devised that are variously described as *place-based* or *local*. For example, the Geographically Weighted Regression (GWR) of Fotheringham, Brunsdon, and Charlton (2002) allows the parameters of a model to vary spatially, holding only the structure of the model constant. Instead of a single, universal estimate of each parameter, the technique produces maps showing each parameter's spatial variation. The results can be interpreted in several ways. One might, for example, hypothesize that the process represented by the model is simply nonstationary, and accept something less than the ideals of nomothetic science. On the other hand one might argue that in the social sciences the goal of a fully specified model and an R^2 of 1 is in principle unattainable, and that the unspecified variables will almost always display spatial variation. Either way, GWR represents something of a retreat from nomothetic science to a new middle ground in which the structure of the model is held constant but the parameters vary over space. Figure 3 shows a typical output from GWR analysis, displayed in a GIS. The percentage of the population with bachelor degrees has been regressed against the percentage living in poverty, by sequentially centering the

JOS 2015—27/6/2007——278526

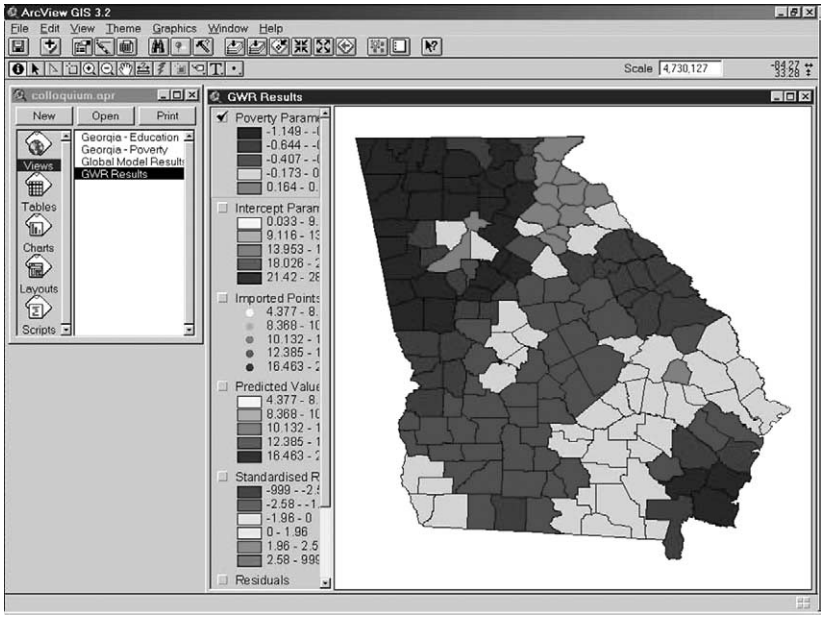

Fig. 3. A Geographically Weighted Regression of percent with bachelor degrees against percent in poverty for the counties of Georgia. The display shows the spatial variation in the regression coefficient

analysis on each county and weighting all other counties by an inverse function of distance, leading to a map of the spatial variation in the regression coefficient.

Another substantive driver comes from interest in the application of knowledge to policy. If knowledge in the social sciences is nomothetic and abstracted from space and time, then its application in the development of policy clearly requires that it be reinserted into a spatial context. Solving social problems in Los Angeles, for example, requires that general knowledge be combined with knowledge of the specific conditions of that city. This duality is mirrored precisely in the structure of a GIS, which applies the general procedures of its algorithms and analytic scripts to the special local conditions represented in its database. Tools such as GIS play an essential role in the complete cycle of science, in which results are first generalized away from their spatial context in the pursuit of nomothetic knowledge, and then reinserted into it in the development of policy and the making of decisions.

These are only two of the substantive arguments for the spatial turn, and Janelle and I list a further four (Goodchild and Janelle 2004b). Combined with the technological arguments, they present a powerful case for a transformative movement in the social sciences, and to a lesser extent in the environmental sciences and in any discipline that values visualization as a basis of interpretation or communication.

3. The Characteristics of Geographic Data

Geographic data can be defined simply as data about the locations of phenomena on the Earth's surface. More formally, they can be expressed as tuples of the form $< \mathbf{x},\mathbf{z} >$ where **x** denotes one or a set of locations and **z** denotes properties that are observed at those

locations. To qualify as geographic, therefore, any record must have associated with it a georeference, expressed in some recognized system of coordinates. Given this comparatively narrow definition of a class of information, it seems reasonable to ask whether geographic data possess any distinguishing characteristics, beyond those that are inherent in this definition. Are there ways in which "spatial is special?" Such properties would be important in the design of technologies for handling geographic data, for example, as well as in analysis. Over the past few decades a series of properties have been identified, some of them more significant than others (Anselin 1989; Goodchild 2004a). In this section I will review four, all of which present particular problems for the statistical analysis of spatial data.

First, and perhaps best-known, is the property of spatial dependence, which geographers often know as Tobler's First Law of Geography (Tobler 1970; Sui 2004). Simply stated, "all things are similar, but nearby things are more similar than distant things" – in other words, geographic series tend to be positively[1] spatially autocorrelated, and correlation decreases with distance. While the statement appears innocent enough, it underlies all concepts of region, and all methods of spatial interpolation, which attempt to estimate the value of a spatial variable at a location **x** where it was not measured, based on measured values at other locations. Many methods of spatial interpolation are weighted averages, in which nearby measurements are given more weight than distant ones. More formally, the law is the basis for geostatistics, otherwise known as the theory of regionalized variables, which is based on the observation that for most geographic data spatial autocorrelation declines in a regular fashion with distance, normally up to a distance termed the *range*.

Endemic positive spatial dependence has obvious consequences for the application of methods of statistical inference to geographic data. If observations are not spaced sufficiently far apart, degrees of freedom will be artificially inflated, leading to exaggerated significance and Type I errors. Culling data to increase separation to at least the range of the phenomenon in order to achieve independence is never popular, so the only satisfactory option is to model spatial dependence explicitly, through the use of spatially autoregressive or spatially lagged models (Haining 2003).

Second, the Earth's surface is characterized by strong nonstationarity or what is often termed spatial heterogeneity. It is difficult to conceive of an average place on the Earth's surface, and the full extent of geographic variation was only known when all parts of the surface had been explored. Spatial heterogeneity has important corollaries for statistical analysis. It implies, for example, that the results of any analysis will depend explicitly on its geographic bounds, and ensures that statistical inference about the entire Earth's surface from samples will be highly problematic. It presents a strong case for the kinds of place-based or local analysis discussed previously. It implies also that standards and practices devised for any region of the globe will differ from those devised for other regions, and that analysts studying extended areas will therefore almost always have to contend with differences in classification schemes, geographic coordinate systems, and other aspects of geographic practice.

[1] Positive in the sense of indices such as Moran's *I*, which behave similarly to bivariate correlation coefficients, but not in the sense of Geary's *c*. Note also that the expected value of *I* in a random pattern is slightly negative.

Third, many geographic phenomena exhibit fractal behavior. Examining many phenomena more closely reveals additional detail, almost *ad infinitum*, and often at a predictable rate. Measures of many phenomena such as coastlines are strictly functions of scale, as are counts of such features as islands or lakes. Geographic surfaces are often nondifferentiable, and as a result properties such as slope are also scale-dependent.

Finally, and in part as a consequence of the second principle of spatial heterogeneity, it is rare to encounter geographic projects for which the assumption of random sampling is tenable. Many projects use all of the cases available within a defined project area, such as all of the census tracts of Los Angeles, or all of the lakes of Minnesota, or all of the rivers of Kentucky. In such instances there is no concept of a larger population from which the cases of the study were drawn randomly and independently, and about which inferences might be made. One might take the tracts of Los Angeles as a sample of all of the tracts in the U.S., or one might suppose that the census tracts of Los Angeles represent a sample of all of the tracts that might have existed if the universe's clock had been restarted, but in neither instance is the random sampling assumption acceptable.

In summary, conventional statistical inference is rarely possible with the "natural experiments" captured in geographic data. One can subsample to avoid the spatial dependence problem, but can then make inferences only about the larger sample, not some more general universe. One can construct null hypotheses that do not require assumptions of random, independent sampling, such as the randomization null commonly used to assess the significance of tests of spatial autocorrelation (Cliff and Ord 1981). One can simply avoid statistical inference altogether, arguing that it is sufficient to make descriptive statements about the data at hand. But it is hard to defend this position against journal editors and others who feel, rightly or wrongly, that every scientific result should be accompanied by some measure of significance.

4. Applications of Statistical Geometry

One of the lasting joys of working in this field derives from its ability to motivate classical problems in statistical geometry. In a recent paper, Shortridge and I (Shortridge and Goodchild 2002) show how the problem of Buffon's Needle can be used to estimate useful properties of spatial databases, based on knowing the probability that two points distance d apart will lie in the same cell of a square grid. As another example, consider that many maps, including those of soils, land cover, or political administration, consist of irregular tesselations in which each face is assigned to one of a number of classes (or administrative units in the political case) – the *area-class* map. Digital representations of such maps often use the common boundary between each pair of adjacent faces as the fundamental unit – the *arc*, edge, or link of the boundary network. We observe that almost all nodes in such networks are three-valent (the U.S. state boundary network famously contains one four-valent node), another general characteristic of geographic data. It follows from a theorem due to Euler that the average number of arcs per face will be slightly less than six, whatever the process responsible for creating the boundary network (Okabe, Boots, and Sugihara 1992) – an extreme example of equifinality and the inability of spatial pattern to identify process uniquely, and yet a result of some utility in system design.

Maps ultimately derive from measurements, and thus must be subject to measurement error. But in practice several other sources of error and uncertainty affect the content of geographic databases, including lack of precise definitions of boundaries (where is the edge of a hill, for example?), lack of replicability in the definitions of classes (two observers will almost always not agree on the details of a soil map), and uncertainty in the coordinate frame itself (the Earth's axis wobbles measurably, and tectonic and tidal movements constantly affect locations). No model can provide a perfect replica, and geographic models are inevitably subject to levels of generalization and approximation appropriate to their designed scales. In short, all geographic data are subject to uncertainty, in the sense that they leave the user uncertain about what actually exists on the Earth's surface.

Confidence regarding the *marginal* properties of single points can be addressed using simple measures of uncertainty such as the standard error or the probability of misclassification. One can, for example, replicate the assignment of classes to points on the landscape using a sample of observers, or rely on the known properties of measuring instruments such as GPS. But many of the useful products of GIS processing require knowledge of *joint* characteristics. For example, in order to determine uncertainty in slope estimates it is necessary to know the covariances between errors in point estimates of elevation (Hunter and Goodchild 1997), and we know from Tobler's First Law that such covariances will generally be positive, decreasing functions of distance. Similarly covariances are needed to determine uncertainties in estimates of distance or area, and in the products of virtually all useful GIS operations. In another example, suppose a model has been used to assign probabilities p of landslide to each pixel in an area, based on such variables as surficial geology, slope, and aspect (Chung and Fabbri 2005). But landslides do not confine their impacts to single pixels, so it would be useful to know the probability of a slippage involving all of the n pixels in some contiguous area of uniform probability. Because of unknown but inevitable spatial dependence, we know only that the probability of all pixels slipping lies somewhere between p^n and p.

Certain useful conclusions result from this reality. First, it is common to treat an entire map as a single realization of a stochastic process, since its individual component parts cannot be modeled independently. In the past few years a wealth of models have been devised for the simulation of uncertainty in geographic data, each realization representing one possible and equally likely version of the truth (Zhang and Goodchild 2002). Second, *relative* properties such as distance and slope will be much more accurately represented in geographic data than *absolute* properties, and indeed it is surprising perhaps to realize that it is possible to track the movement of faults to mm accuracy, and to estimate slopes to single degrees, despite absolute errors of position or elevation that are commonly in the meter range.

This last point leads to an interesting debate over the appropriate way to design a geographic database. Traditionally, every point in such databases has been positioned in absolute terms, using latitude and longitude or some similarly universal coordinate system. Somewhat paradoxically, then, in order to determine the distance between two points a few meters apart, a GIS must first look up their respective positions relative to the Equator and the Greenwich Meridian, and then compute distance from these coordinates. Precisions of better than 1 part in 10^7 are required in such a calculation to achieve even

meter accuracy, and software designers are often forced to resort to double precision. Yet a simple tape measure with an accuracy of 1 part in 10^3 would produce results good to mm. This suggests an alternative to the traditional *coordinate-based* GIS that one might term a *measurement-based* GIS, in which positions are obtained on the fly from knowledge of the measurements on which they were based – and knowledge of the accuracies of those measurements can then be readily propagated into estimates of product uncertainty. A series of recent papers (Goodchild 2004b) provides a comprehensive comparison of these approaches.

Models of uncertainty in many types of geographic data are now available (Zhang and Goodchild 2002), and slowly making their way into software products. Much is known about the visualization of uncertainty, and several interesting approaches have been explored. Standards have been adopted for the description of data quality as part of data documentation or *metadata*. In one respect, however, the picture remains tantalizingly incomplete – we have as yet no acceptable model of uncertainty in area-class maps.

As noted earlier, such maps partition the plane into areas of uniform class (the case of administrative units requires a somewhat different approach to uncertainty), and as such constitute mappings from location $\mathbf{x}$ to a nominal variable c. The marginal uncertainty at a point can be characterized as a vector of probabilities, and summarized as a table comparing recorded and actual classes that is commonly known as the *confusion* or *error* matrix. But in order to place confidence limits on properties such as area, it is necessary to know the joint probabilities, and these are subject to covariance.

Area-class maps are assemblages of nodes, arcs, and faces, so one might approach the problem by specifying error models for each entity. But this approach would fail to address the fact that multiple replications will differ not only in the positions of nodes and arcs, but also in the numbers of nodes and arcs and the topology of the boundary network. Moreover the homogeneity of a face must itself be subject to uncertainty.

Thus a more acceptable approach is to represent the map as a raster, and to address the contents of each raster cell and the covariances between them. Two approaches among the many that have been suggested appear to be worth pursuing. First, assume that we know the probability that cell j belongs to class i. Generate n realizations by assigning classes independently to each cell according to these probabilities. Then taking each cell in turn, consider a random pair of realizations, and evaluate a swap of their contents. If swapping brings both realizations closer to a prescribed covariance goal, make the swap, and continue until no further improvement occurs.

The second method has the advantage of being more firmly grounded in a theory about how area-class maps might arise. Consider a set of m properties $z_i(\mathbf{x})$, $i = 1, m$ that are each functions of location $\mathbf{x}$. In the example of vegetation cover these might represent rainfall, temperature, and other physical variables. Create a partition of an m-dimensional space with axes defined by the properties z_1 to z_m that assigns every point in the space to one of k classes. Now simulate the variables z with specified spatial autocorrelation properties using one of a number of suitable random-field models. By looking up each location in the m-dimensional space we produce a simulated area-class map, with appropriate covariances. This model matches well to theories about the spatial distribution of vegetation (see, for example Holdridge 1971), and is also attractive as a model of soil

maps. But its main disadvantage is its over-specification – it is difficult to imagine how it might be calibrated in practice.

5. Challenges and Future Directions

In this final section I would like to address four topics that to me represent challenges and opportunities in the application of statistical methods in spatial social science. There are many others, of course, and it is clear that the technological drivers of the field will continue to stimulate rapid evolution.

The roots of GIS are in maps, and maps are inherently static, favoring aspects of the world that remain approximately constant. Phenomena such as flows, events, and transactions are not normally associated with maps. Yet digital technology has no such aversion to dynamics, and many authors have remarked on the extent to which the handling of time is inadequate in today's geographic information technologies. In part this is the result of a lack of data, and the difficulty of creating longitudinal records of human societies. In part it is due to a lack of conceptual and theoretical frameworks, and appropriate null hypotheses. For example, we make frequent use of the Complete Spatial Randomness (CSR) null in the analysis of cross-sectional point patterns, but have no equivalent for the tracks of those points through time. Moreover, the supply of spatiotemporal data is expanding rapidly, as a result of tracking (Figure 4), the geocoding of events and transactions, the increasing temporal frequency of imaging systems, and analyses of online activity. The STARS (Space Time Analysis of Regional Systems; regal.sdsu.edu/index.php/Main/STARS) open-source software developed by Rey and others makes many techniques of spatio-temporal analysis accessible, and there have been significant advances recently in the theory of tracks (Miller 2005), as well as interesting

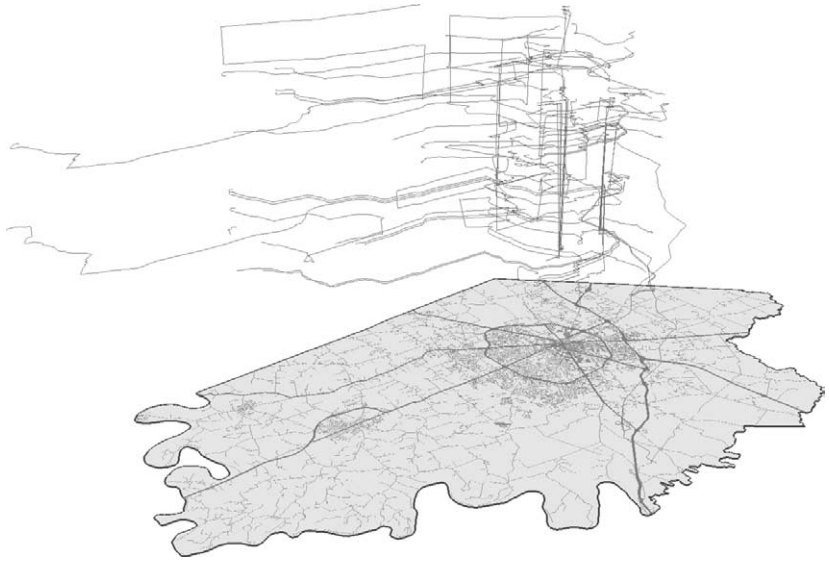

Fig. 4. Tracks of a sample of individuals in Lexington, KY. Each track is shown as a line in three dimensions, with the vertical dimension representing time. (Courtesy MeiPo Kwan)

developments in the extension of GIS packages to dynamic modeling (Maguire, Batty, and Goodchild 2005).

Advances in software engineering practice are also promising new opportunities in spatial social science. In the past, proprietary packages dominated the GIS software industry, resulting in endless problems of interoperability and frustration on the part of the research community, which saw software vendors as interested primarily in satisfying the needs of their largest customers, rather than in advancing the analytic power and intellectual integrity of their products. Today object-oriented software design and the widespread adoption of standards for software and date interoperability have made it far easier to combine the functions of different packages (Ungerer and Goodchild 2002), and to exchange data and procedures. There is growing acceptance of the need for standards to support the exchange and archiving of complex analysis scripts (Crosier et al. 2003), and growing interest in the open-source paradigm as a reliable method of developing and sharing research-oriented software. On the other hand, the growing power of Google seems set to ensure that many essential aspects of the computing environment will continue to be dictated by the private sector.

Longley and I (Goodchild and Longley 1999) have written about the many new sources of data that are becoming available for spatial social science. These include data from the private sector that are often associated with the term *geodemographics*, and comprise detailed records on individuals and small areas that have been culled from the census, and from retail transactions and buying habits (see, for example www.spatial-literacy.org). While such data are rarely assembled with the kinds of rigor expected of scientific research, including publication of sufficient detail to allow replication, they are nevertheless useful in a number of contexts, particularly in providing the basis for stratified sampling. In addition to these, the search engines are now making detailed data available on search and communication patterns, and the spatial structure of online connectivity has already yielded interesting insights (Dodge and Kitchin 2001). Other major new data sources include the results of large-scale tracking studies conducted by various transportation agencies (Kwan and Lee 2004), which allow detailed visualization and examination of the spatio-temporal behavior of individuals.

Many of these new sources raise important issues of confidentiality and research ethics, and are likely to create interesting discussions in the context of the IRB process. Geographic location is an important key to individual identity and in the form of street address can be used to link records that would by themselves be acceptable. Several vendors now offer very extensive databases on individuals that have been assembled through this process, and augmented by additional data on small areas that have been imputed to the area's individual residents. A recent story in the *New York Times* (Barbaro and Zeller 2006) drew attention to the potential of data being collected and published by the Internet search services (in this case AOL) to pry into the personal interests of individuals, and to identify the individuals uniquely through a process of geographic reasoning. In another recent example, Curtis (2006) has shown the power of current technologies to breach confidentiality based on geographic location. Shortly after Hurricane Katrina, the New Orleans *Times-Picayune* published a map showing the locations where bodies were found during the initial response. Although the map was at a coarse scale, and although streets were not shown on the map, it was possible to

geo-register it using the census tract boundaries shown (readily available in digital form from many sources), and to use free web-based geocoding services to identify correctly roughly 70% of the houses and apartments involved. Curry (1998) provides a more general discussion of the issue of privacy in the context of current geographic information technologies.

In summary, spatial social science appears to have caught the imagination of a significant number of researchers, and to be generating widespread interest. It is driven in part by advances in geographic information technologies and by new sources of data, both of which are likely to continue to provide new opportunities in the future – and in part by the substantive merits of a spatial perspective. At the same time it raises issues of confidentiality that are likely both to effect the work of social scientists and to provide grist for those social scientists interested in the broader social effects of technology.

6. References

Anselin, L. (1989). What Is Special about Spatial Data? Alternative Perspectives on Spatial Data Analysis. Technical Paper No. 89-4. Santa Barbara, CA: National Center for Geographic Information and Analysis.

Anselin, L., Florax, J.G.M., Rey, S. (eds) (2004). Advances in Spatial Econometric Modeling: Methodology, Tools, and Applications. Heidelberg: Springer-Verlag.

Barbaro, M., Zeller Jr., T. (2006). A Face is Exposed for AOL Searcher No. 4417749. New York Times (August 9). Available: http://www.nytimes.com/

Bradburn, N.M. (2004). Foreword. In Spatially Integrated Social Science, M.F. Goodchild and D.J. Janelle (eds). New York: Oxford University Press, v-vi.

Bunge, W. (1966). Theoretical Geography. Lund: Gleerup.

Chung, C.F., Fabbri, A.G. (2005). Systematic Procedures of Landslide Hazard Mapping for Risk Assessment Using Spatial Prediction Models. In Landslide Hazard and Risk, T. Glade, M.G. Anderson, and B.J. Crozier (eds). New York: John Wiley and Sons, 139–174.

Cliff, A.D. and Ord, J.K. (1981). Spatial Processes: Models and Applications. London: Pion.

Cohen, J. and Tita, G. (1999). Diffusion in Homicide: Exploring a General Method for Detecting Spatial Diffusion Processes. Journal of Quantitative Criminology, 15, 451–494.

Cromley, E.K. and McLafferty, S.L. (2002). GIS and Public Health. New York: The Guilford Press.

Crosier, S.J., Goodchild, M.F., Hill, L.L., and Smith, T.R. (2003). Developing an Infrastructure for Sharing Environmental Models. Environment and Planning B: Planning and Design, 30, 487–501.

Curry, M.R. (1998). Digital Places: Living with Geographic Information Technologies. New York: Routledge.

Curtis, A. (2006). GIS and Emergency Management. Presentation, GIS in Practice. National Centre for Geocomputation, National University of Ireland, Maynooth, January 25. Available: http://ncg.nuim.ie/ncg/events/20060125/

Dodge, M. and Kitchin, R. (2001). Mapping Cyberspace. New York: Routledge.

Fotheringham, A.S., Brunsdon, C., and Charlton, M. (2002). Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Hoboken, NJ: Wiley.

Fujita, M., Krugman, P., and Venables, A. (1999). The Spatial Economy. Cities, Regions and International Trade. Cambridge: The MIT Press.

Goodchild, M.F. (2004a). The Validity and Usefulness of Laws in Geographic Information Science and Geography. Annals of the Association of American Geographers, 94, 300–303.

Goodchild, M.F. (2004b). A General Framework for Error Analysis in Measurement-based GIS. Journal of Geographical Systems, 6, 323–324.

Goodchild, M.F., Egenhofer, M.J., Kemp, K.K., Mark, D.M., Sheppard, E. (1999). Introduction to the Varenius project. International Journal of Geographical Information Science, 13, 731–745.

Goodchild, M.F. and Janelle, D.G. (eds) (2004a). Spatially Integrated Social Science. New York: Oxford University Press.

Goodchild, M.F. and Janelle, D.G. (2004b). Thinking Spatially in the Social Sciences. In Spatially Integrated Social Science, M.F. Goodchild and D.G. Janelle (eds). New York: Oxford University Press, 3–21.

Goodchild, M.F. and Longley, P.A. (1999). The Future of GIS and Spatial Analysis. In Geographical Information Systems: Principles, Techniques, Applications and Management, P.A. Longley, M.F. Goodchild, D.J. Maguire, and D.W. Rhind (eds). New York: Wiley, 567–580.

Haining, R.P. (2003). Spatial Data Analysis: Theory and Practice. New York: Cambridge University Press.

Holdridge, L.R. (1971). Forest Environments in Tropical Life Zones: A Pilot Study. New York: Pergamon Press.

Hunter, G.J. and Goodchild, M.F. (1997). Modeling the Uncertainty in Slope and Aspect Estimates Derived from Spatial Databases. Geographical Analysis, 29, 35–49

Kwan, M.-P. and Lee, J.-Y. (2004). Geovisualization of Human Activity Patterns Using 3D GIS. In Spatially Integrated Social Science, M.F. Goodchild and D.G. Janelle (eds). New York: Oxford University Press, 48–66.

Longley, P.A., Goodchild, M.F., Maguire, D.J., and Rhind, D.W. (2005). Geographic Information Systems and Science. New York: Wiley.

Maguire, D.J., Batty, M., and Goodchild, M.F. (eds) (2005). GIS, Spatial Analysis, and Modeling. Redlands, CA: ESRI Press.

Miller, H.J. (2005). A Measurement Theory for Time Geography. Geographical Analysis, 37, 17–45.

Moran, E.F., Brondizio, E.S., and McCracken, S. (2002). Trajectories of Land Use: Soils, Succession, and Crop Choice. In Deforestation and Land Use in the Amazon, C. Wood and R. Porro (eds). Gainesville: University of Florida Press, 193–217.

Okabe, A., Boots, B.N., and Sugihara, K. (1992). Spatial Tessellations: Concepts and Applications of Voronoi Diagrams. New York: Wiley.

Sampson, R.J., Raudenbush, S., and Earls, F. (1997). Neighborhoods and Violent Crime: A Multilevel Study of Collective Efficacy. Science, 277, 918–924.

Shortridge, A.M. and Goodchild, M.F. (2002). Geometric Probability and GIS: Some Applications for the Statistics of Intersections. International Journal of Geographical Information Science, 16, 227–243.

Sui, D.Z. (2004). Tobler's First Law of Geography: A Big Idea for a Small World? Annals of the Association of American Geographers, 94, 269–277.

Tobler, W.R. (1970). A Computer Movie: Simulation of Population Change in the Detroit Region. Economic Geography, 46, 234–240.

Ungerer, M.J. and Goodchild, M.F. (2002). Integrating Spatial Data Analysis and GIS: A New Implementation Using the Component Object Model (COM). International Journal of Geographical Information Science, 16, 41–54.

Zhang, J.-X., and Goodchild, M.F. (2002). Uncertainty in Geographical Information. New York: Taylor and Francis.

Received April 2007