Geographical Analysis ISSN 0016-7363

Statistical Perspectives on Geographic Information Science

Michael F. Goodchild

National Center for Geographic Information and Analysis, and Department of Geography, University of California, Santa Barbara, CA

Statistical geometry applies probabilistic methods to geometric forms. In the early days of the quantitative revolution statistical geometry appeared to provide a useful framework for geographic research, but its value appeared to decline in the 1970s and 1980s. Geographic information science (GIScience) addresses the fundamental issues underlying the geographic information technologies, and statistical geometry has proven valuable in a number of respects. Several classical results from statistical geometry are useful in the design of geographic information systems, and in understanding and modeling uncertainty in geographic information, and several statistical principles are observed to be generally applicable to geographic information. Modeling uncertainty in area-class maps has proven particularly difficult, and seven possible models are discussed. Statistical geometry provides an important link between the early work of the quantitative revolution in geography and modern research in GIScience.

Introduction

In the history of quantitative geography there have been many instances of techniques and theoretical frameworks being discovered in other disciplines, and adapted, expanded, and enhanced to support the analysis of phenomena distributed on the Earth's surface. This process peaked during the 1960s, in the heyday of the quantitative revolution, but has never been entirely absent. Multidimensional scaling, for example, was developed largely in mathematical psychology; its applications to geographic space were recognized by some of the early behavioral geographers (Bunge 1966); and it is now an important part of the toolkit of spatial analysis. More recently, geostatistics, developed initially in France in the 1960s

Correspondence: Michael F. Goodchild, National Center for Geographic Information and Analysis, and Department of Geography, University of California, Santa Barbara, CA 93106-4060
e-mail: good@geog.ucsb.edu

Submitted: May 19, 2007. Revised version accepted: October 23, 2007.

(Matheron 1971; Goovaerts 1997), is now an important part of our understanding of accuracy in geographic data (Zhang and Goodchild 2002).

In this article I hope to demonstrate that the goals that motivated the quantitative geographers of the 1960s—the legends of this special issue—are alive and well in at least one branch of modern geography. I focus on statistical geometry, or the application of probabilistic methods to geometric forms. The roots of the field go back at least as far as the 18th Century, and later major contributions were made by the likes of M.G. Kendall (1961), W.W. Rouse Ball and H.S.M. Coxeter (1987). Bunge (1966) clearly saw many geographic forms as resulting from stochastic processes, and believed that similar forms would arise in very different areas of geography as a result of similarity of process. For example, similarity of form between river meanders and alpine roads suggested some degree of isomorphism in their generating processes that might be uncovered by trying to identify their underlying stochastic processes. In the late 1960s and early 1970s an extensive literature developed along these lines, with geographers examining the spatial distributions of phenomena as diverse as settlements, businesses, trees, and instances of disease against simple stochastic models. That literature is now increasingly relevant to the field of geographic information science (GIScience), which studies the fundamental scientific issues behind the use of geographic information technologies (Goodchild 1992).

Later, of course, this line of research lost much of its momentum, for two compelling reasons. First, the same spatial distribution will often arise from several distinct processes, and while the principle of Occam's Razor might allow one to pick the simplest of the alternatives, it is hard to argue in many cases that one process is any simpler than the others. Variation in the density of a point pattern, for example, can always arise from either a first-order process in which density responds to some other spatial variable, or from a second-order process in which the presence of one point makes others more likely in the immediate vicinity; thus concluding that a pattern is clustered, and rejecting the null hypothesis of complete spatial randomness, does very little to advance understanding. Second, researchers grew frustrated with the broader quality of explanation provided by such models. Sack (1980) argued that geometric theories of pattern could never be truly satisfying, and with others led geography's long retreat from the summit of the quantitative revolution.

But in one area of geography statistical geometry remains a very significant framework, and it continues to dominate much of my own work. Geographic information systems (GIS) arose in the 1960s in response to several overlapping needs. In Canada, Roger Tomlinson and others confronted the massive demands for numerical analysis created by the Canada Land Inventory, and recognized that by far the most cost-effective means of measuring vast numbers of irregularly shaped areas on maps was to digitize them into computers, despite the enormous cost of computing at the time. At Harvard, the Laboratory for Computer Graphics founded by Howard Fisher was developing some of the first software for computer-based

mapping, and later evolved under the direction of William Warntz and Brian Berry into one of the most influential centers of GIS development as the Laboratory for Computer Graphics and Spatial Analysis (Chrisman 2006). In Chicago, Duane Marble needed to organize the numerous large data sets being assembled for studies of metropolitan transportation; while in Washington, DC, the Bureau of the Census was anxious to employ computers to manage and analyze the geographic dimensions of the 1970 census. All of these efforts and others eventually converged into the software products of the late 1970s, and into the GIS software industry that we know today (Coppock and Rhind 1991; Foresman 1998).

The Canada Geographic Information System (CGIS) built by Tomlinson and others was vector-based, recording the boundaries between adjacent areas as sequences of vertices connected by straight lines. In order to estimate the amounts of land that were *both* classified as suitable for agriculture on the map of Soil Capability for Agriculture *and* classified as currently not used for agriculture on the map of Current Land Use it was necessary to compute the intersection of the respective areas, a task that turned out to be quite daunting from an algorithmic perspective. Moreover, it was inevitable that some boundaries should coincide on both maps, because they followed identical features on the Earth's surface; but equally inevitable that the two digital versions would not coincide. The result was an enormous number of small slivers, and when many maps were overlaid the numbers of slivers could easily swamp the software. Table 1 shows an example drawn from CGIS. Various combinations of five layers have been overlaid, and the resulting areas tabulated by their area measure in acres. The largest numbers of slivers are generated by overlaying maps of land use at different times, because of the frequency with which the same lines on the ground appear repeatedly.

Table 1 Numbers of Polygons by Area for Five Canada Geographic Information System (CGIS) Layers and Overlays

Acres	1	2	3	4	5	1+2+5	1+2+3+5	1+2+3+4+5
0–1	0	0	0	1	2	2,640	27,566	77,346
1–5	0	165	182	131	31	2,195	7,521	7,330
5–10	5	498	515	408	10	1,421	2,108	2,201
10–25	1	784	775	688	38	1,590	2,106	2,129
25–50	4	353	373	382	61	801	853	827
50–100	9	238	249	232	64	462	462	413
100–200	12	155	152	158	72	248	208	197
200–500	21	71	83	89	92	133	105	99
500–1,000	9	32	31	33	56	39	34	34
1,000–5,000	19	25	27	21	50	27	24	22
>5000	8	6	7	6	11	2	1	1
Totals	88	2,327	2,394	2,149	487	9,558	39,188	90,599

Layer 1: soil capability for agriculture; Layers 2–4: land use at three different dates; Layer 5: recreational capability.

This problem is clearly amenable to statistical analysis, and in 1977 I published the first of the many articles I have written on statistical aspects of GIS (Goodchild 1977). I was able to fit mixtures of stochastic models to the distributions of polygon measures, and thus to estimate the probability that a given polygon was a sliver. Polygon area for real polygons tends to follow a lognormal distribution, while for slivers its square root tends to follow an exponential distribution. I was much influenced by the work of D. H. Maling, a cartographer interested in statistical applications and author of a text on the accuracy of measurements from maps (Maling 1989). In order to avoid some of the bottlenecks of the vector approach, a former student J. H. Ross and I had argued that rasterizing the digital maps of CGIS would permit a much more user-friendly and personalized approach to analysis, but the all-important question of the amount of information loss by rasterizing at a given spatial resolution needed to be addressed. Frolov and Maling (1969) had obtained results regarding the accuracy of area estimates from rasterized maps, and I was able to extend these (Goodchild 1980a) using Mandelbrot's recently developed theory of fractals (Mandelbrot 1977).

Since the 1970s many useful results have been obtained that help us to understand the nature of uncertainty in geographic data. In the next section I briefly review some applications of classic problems in statistical geometry that have recently found interesting new applications in GIScience. This is followed by a discussion of some general principles, and then by a final section that examines the problem of uncertainty in area-class maps, which is in many ways the outstanding problem of this area of GIScience.

Classic problems in statistical geometry

The problem now known as Buffon's Needle was first posed by Georges-Louis Leclerc, Compte de Buffon, in 1777. Given a floor of parallel strips of wood of equal width, what is the probability that a needle dropped on the floor will lie across the boundary between two adjacent strips (or conversely, will fall entirely within one strip)? It is easy to show that if the needle length is equal to the strip width, the probability is $2/\pi$ or 0.6366—or more generally, that if the needle is of length l and the strips of width s, the probability is $2l/\pi s$. Part of the interest in the problem stems from the fact that it provides the basis for an empirical determination of π, if the needle is dropped a sufficient number of times. The experiment is hardly efficient, however—to determine π to five decimal places, for example, and using the variance of the binomial distribution, it would be necessary to drop the needle on the order of 10^{10} times.

The relevance of this problem to GIScience becomes clear when it is generalized from linear strips to a raster (Shortridge and Goodchild 2002; and see earlier work by Okabe and Sadahiro 1997). Let b represent the linear dimension of each square cell in the raster, and as before let l represent the length of the needle. When $l \geq s\sqrt{2}$ the needle must intersect more than one square, and the expected number

of squares intersected is given by $4l/\pi s$. On the other hand for short needles and $l < s$ one might be interested in the probability that the two ends fall in different cells, which is given by

$$\left(4ls - l^2\right)/\pi s^2.$$

Both of these results have several interesting applications in the design and evaluation of software. The long-needle case provides useful results for several tasks in raster GIS, for example in determining the expected number of operations in computing whether one cell is visible to an observer located in another cell, based on the distance between them. This is the fundamental operation of the so-called *intervisibility* or *viewshed* problem, which is frequently used in siting cellphone transmitters, or in evaluating the visual impact of new developments.

Two examples illustrate the applications of the short-needle result. In managing large spatial databases it is common to invoke systems of *tiling*, dividing the database into smaller partitions in order to increase efficiency in data creation, management, and use. For example, the U.S. Geological Survey's 1:24,000 series of national topographic maps, covering the 48 coterminous states, is tiled into over 50,000 separate map sheets based on 7.5-min intervals of latitude and longitude. A tiled database is clearly efficient for operations that are confined to a single tile, but becomes much less efficient for operations that must access data in more than one tile. For example, a query about the distance between two points will be handled much faster if both points are located on the same tile. The short-needle case of Buffon's Needle provides the result needed to determine the probability that two points a distance l apart will lie in different tiles, and thus incur this access penalty.

The second example concerns the design of rasters, and particularly the design of quadrat sampling schemes in ecology. Experiments based on quadrats typically confine their observations to single quadrats, and tend to miss interactions that occur across quadrat boundaries. The Buffon's Needle result can be used to determine the probability that interactions over a distance l will cross at least one quadrat boundary, and thus can provide useful guidance in the selection of the quadrat dimensions.

Another interesting application of classical results concerns the statistics of features on various types of map. Consider a county boundary map, such as that shown in Fig. 1. It partitions the space into irregular areas, forming a network of faces, edges, and nodes. In the figure, which is part of the boundary network of the counties of Georgia, there are 12 faces, 32 edges, and 21 nodes (the boundary of the area is included in the count of edges). Area-class maps, which show land classified by such properties as soil class, land-cover class, or land-use class, are similar in dividing the world into an irregular *tesselation* of faces. We observe that virtually all nodes in such maps are 3-valent, with the occasional exception on maps of political boundaries (there is one possible such node in the figure, and one in the U.S. state boundary network).

This result has value in the design of databases, particularly those that use the *coverage* data model and include linked tables of faces, edges, and nodes. A the-

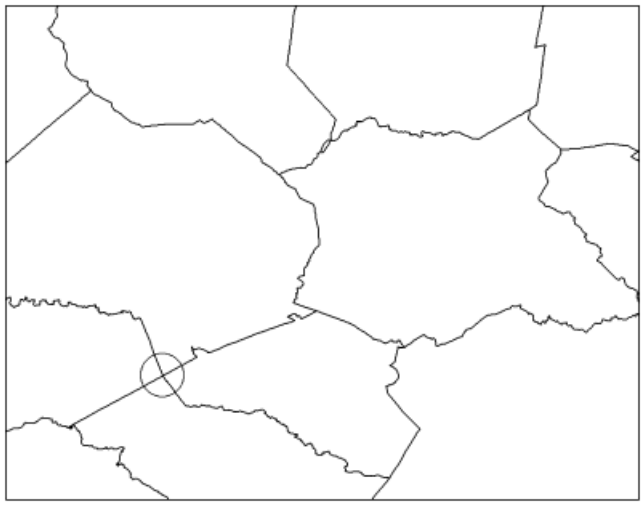

Figure 1. Part of the county boundary network of the state of Georgia, USA. One possible 4-valent node is highlighted with a circle.

orem of Euler states that the numbers of faces F, edges E, and nodes N in such a map must obey the relationship:

$$F - E + N = 1$$

and one can readily confirm this with the figure. If all nodes are 3-valent, it follows that the ratio E/F must be slightly <3 (the deficit is due to the fact that edges on the perimeter of the map are counted only once; Okabe et al. 2000), and that each face will have slightly <6 neighbors (the respective values for Fig. 1 are 2.67 and 4.33).

This result also provides an interesting illustration of the difficulty of inferring process from the study of geographic pattern or form. In classical Central Place Theory (Christaller 1966) it is predicted that settlements on a landscape that is uniform with respect to agricultural productivity will adopt a hexagonal pattern, and similarly that the hinterlands of each settlement will be hexagonal. Experiments in the 1960s (Haggett and Chorley 1969) showed that the average number of neighbors of each settlement's hinterland was very close to six, which appears to lend empirical weight to the theory. But we know from the previous analysis that the average number will be very close to six *whatever* the formative processes operating on the landscape (Getis and Boots 1978). Okabe and Sadahiro (1996) show the implications of this point for Christaller's settlement hierarchy.

Uncertainty in geographic information

Early interest in the accuracy of geographic data and of the analytic methods of GIS focused on the issue of error, based on the normal scientific model of error in measurement. For example, several studies focused on the errors introduced when

points, lines, or areas are digitized as vector point, polyline, or polygon objects. Goodchild and Gopal (1989) published an edited collection of articles from this era, almost all of them using the terms *accuracy* and *error*, and implying that the contents of a GIS database were versions of some true set of contents, distorted by various measurement mechanisms. The recorded position x^* of a point, for example, might be distorted from the true position x by a simple additive error δx.

During the 1990s it became clear that the differences between the contents of a GIS database and the real world that the database purported to represent were much more complex than this simple model suggested. In some cases the concept of a *true* real world was problematic, because mapping involved some degree of subjectivity; or because the definitions of key terms were themselves uncertain. These ideas were particularly applicable to data sets that arose as a result of classification, of phenomena such as soil, land use, or land cover, where the definitions of classes were open to different interpretations, and where it was expected that two observers mapping the same area would produce maps that differed in virtually all aspects: positions of boundaries between areas of homogeneous class, the numbers of such areas, and the numbers of associated edges and nodes. One might still believe that the combined maps of two observers were more reliable than each individual's map, but it was not at all clear how such maps should be combined. Moreover the role of scale was not defined, except to the extent that classification schemes varied with scale.

Because of these arguments the preferred term shifted from error to uncertainty (Zhang and Goodchild 2002), and the literature grew by leaps and bounds. We now have models of most of the more obvious cases of uncertainty in geographic data, and of the processes by which uncertainty in data propagates into uncertainties in the results of spatial analysis (Heuvelink 1998). Monte Carlo simulation has proven to be a very general, powerful, and accessible approach, and several authors have described extensive applications (Aerts, Goodchild, and Heuvelink 2003).

Some general principles can be distilled from this work, and in the next few paragraphs I outline some of them. All derive from the notion that maps and geographic data can be regarded as the outcomes of stochastic processes: that a map is a realization of a stochastic process. In this framework uncertainty leads to a population of possible maps, any of which might be the truth, with the variation between realizations representing uncertainty. This is similar to the traditional theory of measurement error, in which a stochastic process leads to a population of possible measurements. For reasons to be discussed below, however, it is essential that the entire map be treated as a realization, because parts of the map, including individual point, line, or area features, are rarely if ever statistically independent.

First, *all geographic data leave the user uncertain to some degree about the exact nature of the real world*. Whether one believes in a truth or not, no representation of the real world can be exact, because all representations involve some combination of approximation, measurement error, or generalization. The real

world is infinitely complex, but the capacity of any digital system is strictly limited, so some amount of detail must be lost in the creation of any GIS database.

One of the most general principles of geographic data is *spatial dependence*, which is often expressed in the form of Tobler's First Law (TFL; Tobler 1970; Sui 2004), that ''all things are related, but nearby things are more related than distant things.'' This is a statement about spatial autocorrelation, and an informal version of the theory that underlies the fields of geostatistics and spatial statistics. We observe that TFL applies also to uncertainty in geographic data; that differences between database and real world tend to persist over substantial distances, or in other words that differences exhibit positive spatial autocorrelation. There are many reasons for this in the processes of geographic data production—for example, databases created from aerial photography will inherit the positional errors of each photograph, leading to positional errors that persist over the footprint of each image. It follows that *relative errors over short distances will be much less than absolute errors*. In other words, while the positions of objects in a GIS may be distorted in absolute terms, the shapes of objects will by and large be preserved. In the case of elevation data, while elevations may be in error in absolute terms by substantial amounts, the presence of strong spatial autocorrelation of errors ensures that slopes can be estimated with acceptable accuracy (Hunter and Goodchild 1997).

In the final section I focus on one particular type of geographic data, the area-class map, and on the problems of modeling uncertainty in this case. Both of the principles identified above turn out to be of critical importance to what in many ways remains an unsolved problem in GIScience, but one that clearly falls within the broad context of statistical geometry.

Uncertainty in area-class maps

An area-class map can be defined as a mapping $c = f(\mathbf{x})$ where $\mathbf{x}$ denotes a location in the plane and c denotes a class. Area-class maps might be termed nominal fields, in that location maps to a nominal variable, and arise in numerous contexts: as the outcomes of the classification of remotely sensed images; and in the mapping of soils, land cover, or land use. Figure 2 shows a section of a typical area-class map.

Area-class maps are typically made by a long and complex process that involves many stages. Soil maps may be made by characterizing soils at carefully chosen sample sites, and then using aerial photographs and other sources to partition the plane into areas of homogeneous class; other maps may be made by supervised or unsupervised classification of remotely sensed images. Some of the stages are undoubtedly subjective, and as a result maps of the same theme for the same area will not generally be the same. Maps may differ by level of detail and degree of generalization; as a result of vagueness in the definitions of classes; because of variations between observers; and in the case of remote sensing because of errors in measuring instruments, differences in classifiers and training sites, or differences in sensors. Typically, any information that might shed light on the un-

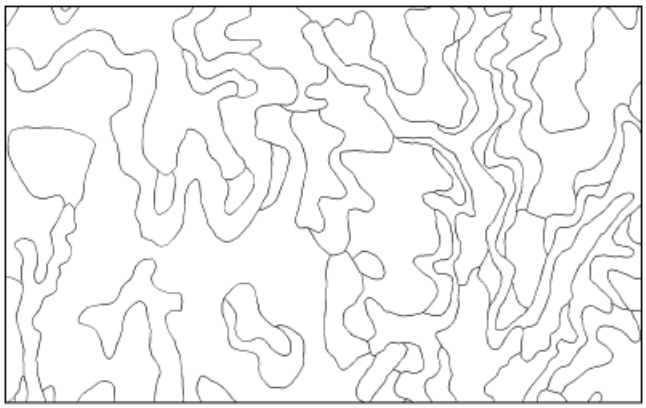

Figure 2. An example area-class map: part of a SSURGO soil map of Osage County, KS.

certainties involved in the mapping process is lost, though the detailed descriptions of each class may imply vagueness of definition.

When represented in the traditional form of a coverage, area-class maps are collections of faces, edges, and nodes, as discussed earlier. In repeated mappings, perhaps by different individuals, the positions of the edges will vary, depending on the clarity with which these boundaries between classes are apparent on the ground. The numbers of objects will also vary, because different observers may perceive different arrangements of faces and their boundary networks. Some concept of a *minimum mapping unit* will have been employed, and faces that are smaller than the minimum area will have been merged into larger faces. Thus the assumption that each face is in reality homogeneous is unlikely to be true.

Any stochastic model of uncertainty for such maps should satisfy a series of requirements, as follows:

1. It should address confusion at every point, between the class c^* recorded on one map, and the classes recorded at the same point on other maps, or where appropriate the true class c at that point.
2. The variation between realizations of the model should emulate the variation between repeated mappings, particularly with respect to variation in the numbers of faces, edges, and nodes.
3. Outcomes at nearby points should exhibit positive spatial autocorrelation, in accordance with TFL.
4. The model should emulate the effects of generalization, both cartographically through changes in the minimum mapping unit or the smoothness of edges and thematically through aggregation of classes.
5. Realizations should be invariant under changes in the underlying spatial representation. For example, if the map is rasterized then realizations should be invariant under changes in the raster cell size, provided that cell area is much less than the minimum mapping unit.

6. Results should be invariant under a re-ordering of the classes, because the recorded class c^* and the true class c are nominal variables.

In the remainder of this section I review several possible models against these criteria, and then provide a brief summary.

The confusion matrix

A common way of assessing the accuracy of area-class maps is to compare the recorded class c^* with some reference source that is believed to be of greater accuracy. For example, classified scenes from remote sensing are often compared with the results of field survey, or *ground truth*. The comparisons are conducted at a sample of locations, and tabulated in a matrix, often termed the *confusion matrix*. Thus $p(c^*|c)$ might denote the probability that the map records class c^* at a location where field survey indicates class c. The kappa index provides a convenient measure of accuracy, scaling from 0 (agreement is what would be expected by chance) to 1 (perfect agreement) (Longley et al. 2005).

The confusion matrix can be applied in two ways. In the *per-polygon* case entire faces are checked against ground truth, but this provides no opportunity to record within-face variation, violating Requirement 1. No variation in the numbers of faces, edges, and nodes is possible, violating Requirement 2. In the *per-point* case individual points are checked, satisfying both of these objections to the per-polygon approach. But as a stochastic model of uncertainty the confusion matrix provides no information on spatial autocorrelation of nearby outcomes, so simulations based on it will not satisfy Requirement 3. Moreover when applied on a raster basis the outcomes in individual cells will be independent, violating Requirement 5. Thus while the confusion matrix provides a useful descriptive measure of accuracy, it fails as a basis for a stochastic model of uncertainty.

The epsilon band

If the transitions across edges are not instantaneous, then one might think instead of transition zones of finite width (Mark and Csillag 1989). Perkal (1966) described a band of width epsilon that represented this zone of uncertainty, and more recently Cohn and Gotts (1996) and others have discussed an *egg-yolk* model in which a band of uncertainty exists around the perimeter of a face, rather like the white of an egg. In terms of the requirements above, the epsilon band addresses only the issue of positional accuracy in a fixed boundary network, and moreover assumes a uniform degree of positional accuracy. Thus it fails to meet Requirements 1 and 2. The epsilon band does not address the process of generalization, violating Requirement 4.

Convolution

Suppose that at every point $\mathbf{x}$ a vector of probabilities P exists, giving the probability that the point belongs in each of the n classes:

$$P(\mathbf{x}) = \{p_1(\mathbf{x}), p_2(\mathbf{x}), \ldots, p_n(\mathbf{x})\},$$

where $p_i(\mathbf{x})$ denotes the probability that point $\mathbf{x}$ belongs to class i. We adopt a raster representation to render the number of locations finite, and assume that P is constant over each raster cell.

If cells were assigned randomly based on these probabilities the map would show no spatial autocorrelation in outcomes, and its ''salt and pepper'' appearance would immediately distinguish it from any real map. Requirement 3 would be violated, and also Requirement 5 because the individual cells would be evident in the outcomes.

A simple way to induce spatial dependence is through a convolution operation, in which a filter is passed over the map. A modal filter that replaces each cell by the commonest class in its neighborhood would be suitable, with the neighborhood size set by some notion of the range of spatial autocorrelation. The results would now satisfy Requirements 3 and 5. However the proportions of outcomes of each class would not equal the prior probabilities, because modal convolution tends to favor the more probable classes at the expense of the less probable ones.

Sequential assignment

Goodchild, Sun, and Yang (1992) propose a model that addresses this problem. A random field $z(\mathbf{x})$ is generated with a controlled pattern of spatial autocorrelation, and rescaled to a uniform distribution in the range $\{0, 1\}$. The value of $z(\mathbf{x})$ is then compared with the vector of probabilities $P(\mathbf{x})$ in every cell, and the class c assigned for which:

$$\sum_{i=0}^{c-1} p_i(\mathbf{x}) < z(\mathbf{x}) < \sum_{i=0}^{c} p_i(\mathbf{x}),$$

where $p_0(\mathbf{x}) = 0$. For example, given probabilities $\{0.2, 0.3, 0.5\}$ Class 1 will be assigned for $z(\mathbf{x})$ in the interval $\{0.0, 0.2\}$, Class 2 for $z(\mathbf{x})$ in the interval $\{0.2, 0.5\}$, and Class 3 for $z(\mathbf{x})$ in the interval $\{0.5, 1.0\}$.

The outcomes of this model satisfy Requirement 1 (within-face variation is modeled), Requirement 2 (realizations vary in the counts of faces, edges, and nodes), Requirement 3 (outcomes exhibit spatial autocorrelation), and Requirement 5 (results are invariant under changes of cell size, given an appropriate resampling mechanism to derive P for each new raster cell). The model also satisfies Requirement 4 concerning generalization, because this can be handled by changes in cell size, smoothing of $z(\mathbf{x})$, or smoothing of $P(\mathbf{x})$. However, the results are not invariant under reordering of the classes, and thus Requirement 6 is violated. Class boundaries will follow the isolines of the $z(\mathbf{x})$ surface, and 3-valent nodes will occur only at the map boundary. Thus the kind of map shown in Fig. 2 will be impossible as an outcome of this process. The problem is moot, of course, in the case of $n = 2$.

Indicator kriging

Indicator kriging (Goovaerts 1997) is used to address problems in which fields have been reduced to binary through the application of some threshold z^*. For example, one might assign $c = 0$, where $z(\mathbf{x}) \leq z^*$, and $c = 1$, where $z(\mathbf{x}) > z^*$. Area-class maps might be simulated by first assigning cells to Class 1 and not Class 1, then among the not Class 1 cells assigning Class 2 and not Class2, and so on. All cells assigned not Class $n-1$ would be assigned Class n. Because $n-1$ random fields would be used, the results would not be subject to the previous objection. Nevertheless the resulting map would be distinct. Figure 3 shows a simple example for $n = 3$. The faces of Class 1 are bounded by isolines of the first random field, but the faces of all subsequent classes are confined to the remaining area. Thus the angles of incidence at the two 3-valent nodes shown in Fig. 3 have unequal probability distributions.

Despite this objection, outcomes such as Fig. 3 make sense in some situations, particularly when ordering relationships exist between some pairs of classes. Suppose that Class i is historically antecedent to Class $i-1$, as it would be for example if it represents agriculture being invaded by urban land use, or grassland being invaded by forest land cover. In such cases the shape of the boundary between Class i and Class $i-1$ is determined by Class i. Thus for some applications the set of classes constitutes a partially or completely ordered set.

Shuffling across realizations

Goodchild (1980b) describes a method of generating maps with specified patterns of spatial autocorrelation. Each face is first independently assigned a random value from a specified distribution. Pairs of faces are then considered at random, and their values swapped if by doing so the map as a whole moves closer to the target spatial autocorrelation.

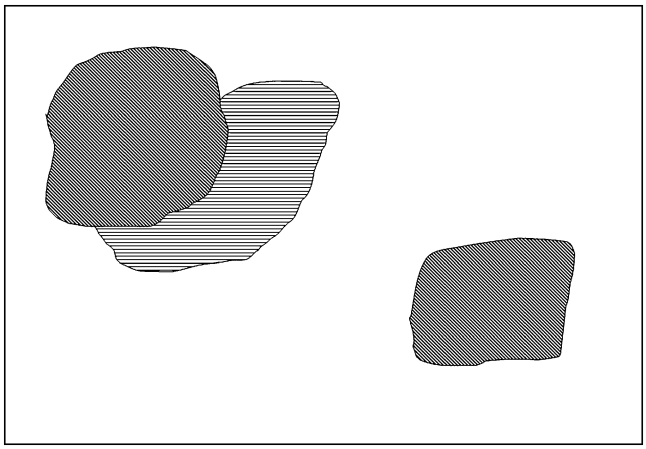

Figure 3. An example area-class map for $n = 3$, showing the effects of sequential assignment on the incidence angles of edges at nodes.

Methods such as this would be unacceptable as models of area-class map uncertainty because they assume homogeneity over the map. But consider a variant in which N realizations are first generated over a raster based on the probabilities $P(\mathbf{x})$. Now pick a random cell and a random pair of realizations, and swap the values in the realizations if both realizations move closer to the target spatial autocorrelation. The process is repeated until no further improvement occurs. It has no justifying interpretation, as the process has no analogs in the real world. Nevertheless all requirements are satisfied, with one exception. Requirement 4, which addresses generalization, could be satisfied if spatial autocorrelation were defined based on distance rather than cell size, and if suitable methods were employed for generalizing $P(\mathbf{x})$ to different cell sizes. Fohl (1998) has implemented the method successfully.

A phase-space model

Goodchild and Dubuc (1987) describe a method for simulating area-class maps (see also Goodchild, Yuan, and Cova 2007). A set of m interval/ratio variables $\{z_1(\mathbf{x}), z_2(\mathbf{x}), \ldots, z_m(\mathbf{x})\}$ is first defined, and random fields generated for each variable covering the area of the map. An m-dimensional ''phase'' space is defined with axes representing each of the variables, and partitioned into n regions, each associated with one of the n classes. Finally, every point $\mathbf{x}$ in the plane is mapped into the phase space, and the corresponding region identified as the class $c(\mathbf{x})$. The entire process is analogous to that used in classifying multiple spectral bands into classes in remote sensing.

This model does not require the specification of P, because the probabilities of each class are established by the regions of phase space and by the random fields. As a model of uncertainty, one might employ it by fixing initial random fields, and then distorting them differently for each realization. The amount of distortion would determine the amount of variation between realizations, and distortions would have to be spatially autocorrelated.

The model has the interesting property that only classes adjacent in phase space can be adjacent geographically, because of the spatial autocorrelation induced in the random fields. This matches experience, in that edges between certain pairs of classes are observed to be much more common than between other pairs.

The model satisfies all of the criteria listed earlier. Requirement 4 can be satisfied in a spatial sense by smoothing the random fields, and in a thematic sense by coarsening the partitioning of phase space. It bears a strong relationship to ecological models of land cover and habitat (e.g., Holdridge 1971). However it is clear that the model is greatly over-specified. It would be very difficult if not impossible in practice to calibrate the variances and correlations of the random fields, or the partitioning of phase space.

Summary

This section has focused on one particularly problematic but nevertheless very common form of geographic data, the area-class map. Such maps clearly defy the normal theories of measurement in science, because they are in part subjective and based on definitions that are necessarily vague. As a result, we have as yet no generally accepted theory of how uncertainty in such data can be modeled, or of the nature of scale effects and generalization. Of the models presented in the previous seven sections, only one, the phase-space model, satisfies all of the specified requirements. Yet unlike such simple models as the Gaussian, it contains far too many parameters to be practical. The shuffling model presented in ''Shuffling across realizations'' satisfies almost all of the requirements, yet lacks any satisfactory empirical basis. At this point, then, one must conclude that the problem of modeling uncertainty in area-class maps remains largely unsolved.

Conclusions

My intent in writing this article was to draw parallels between the contemporary science of geographic information and the work of the early leaders of geography's quantitative revolution, and to argue that the spirit of the 1960s was in many ways alive and well. Much of that early work focused on reasoning from spatial form to spatial process, and thus matches GIScience's traditional focus on spatial form. Much of it focused on stochastic models, and again has proven useful in understanding many aspects of the nature of geographic data, particularly the nature of spatial uncertainty.

Currently, GIScience is experiencing something of a paradigm shift as it places increasing emphasis on time, and the dynamic nature of geographic data. This is driven in part by a new abundance of spatiotemporal data, in part by the development of improved methods of analysis and improved software tools, and in part by the realization that the dynamic aspects of the Earth's surface are in many ways more interesting and important than the static aspects. Fundamental principles, such as TFL, need to be reexamined in this light, as do our models of uncertainty. The comments made in ''Sequential assignment'' and ''Indicator kriging'' about the need for models of uncertainty to be process-based become even more compelling from this perspective.

Many comments have been made over the years about the ability of GIS and GIScience to remotivate interest in quantitative geography. There is no doubt that current interest in TFL is at least in part due to the growth of GIS and GIScience, and to its importance as an underlying principle of these fields. Taylor (1990) has called GIS the ''positivists' revenge'', but the notion that it represents the last resort of the early geographic quantifiers is hardly consistent with the very broad interest that has developed during the past decade or so in almost all scientific disciplines, in public policy and private corporations, and in society at large.

Acknowledgments

This article draws on work supported by many agencies, most recently by the National Geospatial-Intelligence Agency, the Army Research Office, and Award BCS 0417131 from the National Science Foundation.

References

Aerts, J. C. J. H., M. F. Goodchild, and G. B. M. Heuvelink. (2003). ''Accounting For Spatial Uncertainty in Optimization with Spatial Decision Support Systems.'' *Transactions in GIS* 7, 211–30.

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

Chrisman, N. (2006). *Charting the Unknown: How Computer Mapping at Harvard Became GIS*. Redlands, CA: ESRI Press.

Christaller, W. (1966). *Central Places in Southern Germany*. Englewood Cliffs, NJ: Prentice Hall.

Cohn, A. G., and N. M. Gotts. (1996). ''The 'Egg-Yolk' Representation of Regions with Indeterminate Boundaries.'' In *Geographic Objects with Indeterminate Boundaries*, 171–88, edited by P. A. Burrough and A. U. Frank. London: Taylor and Francis.

Coppock, J. T., and D. W. Rhind. (1991). ''The History of GIS.'' In *Geographical Information Systems: Principles and Applications*, 1: 21–43, edited by D. J. Maguire, M. F. Goodchild, and D. W. Rhind. Harlow, UK: Longman Scientific and Technical.

Fohl, P. (1998). Simulating Nominal Fields. Unpublished MA thesis, Department of Geography, University of California, Santa Barbara.

Foresman, T. W. (ed.). (1998). *The History of Geographic Information Systems: Perspectives from the Pioneers*. Upper Saddle River, NJ: Prentice Hall PTR.

Frolov, Y. S., and D. H. Maling. (1969). ''The Accuracy of Area Measurements by Point Counting Techniques.'' *Cartographic Journal* 6, 21–35.

Getis, A., and B. N. Boots. (1978). *Models of Spatial Processes*. New York: Cambridge University Press.

Goodchild, M. F. (1977). ''Statistical Aspects of the Polygon Overlay Problem.'' In *Harvard Papers on Geographic Information Systems*, Vol. 6, 1–30. Reading, MA: Addison-Wesley.

Goodchild, M. F. (1980a). ''Fractals and the Accuracy of Geographical Measures.'' *Mathematical Geology* 12, 85–98.

Goodchild, M. F. (1980b). ''Algorithm 9: Simulation of Autocorrelation for Aggregate Data.'' *Environment and Planning A* 12, 1073–81.

Goodchild, M. F. (1992). ''Geographical Information Science.'' *International Journal of Geographical Information Systems* 6(1), 31–45.

Goodchild, M. F., and O. Dubuc. (1987). A Model of Error for Choropleth Maps with Applications to Geographic Information Systems. Proceedings, Auto Carto 8. Falls Church, VA: ASPRS/ACSM, 165–174.

Goodchild, M. F., and S. Gopal. (1989). *Accuracy of Spatial Databases*. Basingstoke, UK: Taylor and Francis.

Goodchild, M. F., G. Sun, and S. Yang. (1992). ''Development and Test of an Error Model for Categorical Data.'' *International Journal of Geographical Information Systems* 6(2), 87–104.

Goodchild, M. F., M. Yuan, and T. Cova. (2007). ''Towards a General Theory of Geographic Representation in GIS.'' *International Journal of Geographical Information Science* 21, 239–60.

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*. New York: Oxford University Press.

Haggett, P., and R. J. Chorley. (1969). *Network Analysis in Geography*. London: Edward Arnold.

Heuvelink, G. B. M. (1998). *Error Propagation in Environmental Modelling with GIS*. London: Taylor and Francis.

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

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

Kendall, M. G. (1961). *A Course in the Geometry of n Dimensions*. London: Charles Griffin.

Longley, P. A., M. F. Goodchild, D. J. Maguire, and D. W. Rhind. (2005). *Geographic Information Systems and Science*, 2nd ed. New York: Wiley.

Maling, D. H. (1989). *Measurements from Maps: Principles and Methods of Cartometry*. New York: Pergamon.

Mandelbrot, B. B. (1977). *Fractals: Form, Chance, and Dimension*. San Francisco: Freeman.

Mark, D. M., and F. Csillag. (1989). ''The Nature of Boundaries on 'Area-Class' Maps.'' *Cartographica* 26, 65–77.

Matheron, G. (1971). *The Theory of Regionalized Variables and Its Applications*. Paris: École National Supérieure des Mines.

Okabe, A., B. Boots, S. Sugihara, and S. N. Chiu. (2000). *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. Chichester, UK: Wiley.

Okabe, A., and Y. Sadahiro. (1996). ''An Illusion of Spatial Hierarchy: Spatial Hierarchy in a Random Configuration.'' *Environment and Planning A* 28, 1533–52.

Okabe, A., and Y. Sadahiro. (1997). ''Variation in Count Data Transferred from a Set of Irregular Zones to a Set of Regular Zones Through the Point-In-Polygon Method.'' *International Journal of Geographical Information Science* 11(1), 93–106.

Perkal, J. (1966). ''An Attempt at Objective Generalization.'' Translated by W. Jakowski from Julian Perkal, Proba obiektywnej generalizacji. Geodezia es Kartografia, Tom VII, Zeszyt 2, 1958, 130–142. In *Michigan Inter-University Community of Mathematical Geographers, Discussion Paper 10,* edited by J. Nystuen. Ann Arbor, MI: University of Michigan.

Rouse Ball, W. W., and H. S. M. Coxeter. (1987). *Mathematical Recreations and Essays*. New York: Dover.

Sack, R. D. (1980). *Conceptions of Space in Social Thought: A Geographic Perspective*. Minneapolis, MN: University of Minnesota Press.

Shortridge, A. M., and M. F. Goodchild. (2002). ''Geometric Probability and GIS: Some Applications for the Statistics of Intersections.'' *International Journal of Geographical Information Science* 16, 227–43.

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, 269–77.

Taylor, P. J. (1990). ''Editorial Comment: GKS.'' *Political Geography Quarterly* 9, 211–12.

Tobler, W. R. (1970). ''A Computer Movie Simulating Urban Growth in the Detroit Region.'' *Economic Geography* 46, 234–40.

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