A MODEL OF ERROR FOR CHOROPLETH MAPS, WITH APPLICATIONS TO GEOGRAPHIC INFORMATION SYSTEMS

Michael F. Goodchild
and
Odette Dubuc
Department of Geography
The University of Western Ontario
London, Ontario, Canada N6A 5C2

ABSTRACT

The precision of geographic information systems is in sharp contrast to the accuracy of much spatial data, and requires a more objective approach than is conventional in cartography. Existing models of the error of cartographic lines are inappropriate for topological data for various reasons. We propose a model of error in choropleth data, with specific application to the data types found in natural resource inventories. One or more spatially autocorrelated continuous variables are generated, and mapped through a number of domains into a choropleth map with nominal attributes. Fractional Brownian surfaces are convenient sources of the continuous variables. The choropleth boundaries are subject to additional smoothing. Although the model is probably too complex to calibrate, it can be used to simulate choropleth images under a wide range of conditions, in order to investigate effects of error and accuracy in a variety of GIS functions.

INTRODUCTION

One of the more striking results of the introduction of digital data handling methods to cartography has been an increased interest in the interrelated issues of accuracy, precision, error and generalization. A digital system operates with a level of precision which is generally much higher than comparable manual methods, and often much higher than the accuracy of the data. For example, a point in a geographic information system might be represented by a pair of coordinates with a precision determined by the machine's floating point arithmetic, perhaps ten significant digits, whereas its location on a printed map might be accurate to no more than four digits, and might approximate a real feature on the ground to no more than three. The precision of various digital operations may also be far higher than is justified by the accuracy of the data or the conceptual basis of analysis. Poiker (1982, p.241) has compared the high precision of spatial data handling systems to "a person with the body of an athlete in his prime time and the mind of a child".

Statistical theory provides satisfactory methods for describing and dealing with error in scientific measurement, including surveying, but not to the same extent in cartography. Perkal's epsilon band (Perkal, 1956, 1966; Blakemore, 1984; Chrisman, 1982) has been used as an error model of cartographic lines in several recent studies (see also Honeycutt, 1986). Suppose there exists some abstract, true version of a line. Then the model proposes that all real representations of the line will lie within a band of error of width epsilon on either side of this true line. Blakemore (1984) has shown how this

model can be used as the basis for a modified version of the point in polygon problem which explicitly recognizes the uncertainty in the location of a polygon boundary. Honeycutt (1986) has investigated the use of the model for distinguishing between spurious and real sliver polygons in topological overlay algorithms (see also Goodchild, 1978).

Despite the simplicity of the epsilon band concept, there are several reasons for believing that it is not completely satisfactory as a model of cartographic line error. First, although the model proposes that every line lies entirely within the epsilon band, we would expect intuitively that no such deterministic upper limit to error exists: instead, it would seem that larger errors are simply less likely. Error models of simple measurements, such as the Gaussian distribution, place no upper limit on the sizes of errors. Second, the model provides no distribution of error within the epsilon band. Although intuition might suggest that the most likely position for the real line is the centre of the epsilon band, in other words the true position, Honeycutt (1986) has found evidence that digitizing tends to produce a bimodal distribution, such that error on either side of the true line of some measureable amount less than epsilon is more likely than no error. These points suggest that a more suitable model would be some continuous distribution with asymptotic tails centred on the true line, the deterministic epsilon distance being replaced by a standard deviation parameter. The most suitable candidate would be a Gaussian distribution, or following Honeycutt (1986) an equal mixture of two Gaussians, one centred a distance to the left and one the same distance to the right.

Third, while the epsilon band and the modifications suggested above provide a model of deviation for a point on the line, they fail to model the line itself. The locations of two nearby points on the line are not chosen independently, but instead show a strong degree of autocorrelation. Furthermore, it is not clear which points on the line are modelled: Honeycutt (1986) analyzed the positions of digitized points, which are clearly not randomly and independently sampled from the set of all possible points on the line. So the error model cannot provide useful results about spurious sliver polygons, since these are formed not by the deviation of single points but by runs of autocorrelated points on both overlaid lines. A satisfactory error model would have to deal with the line as a continuum with strong autocorrelation.

The final objection to these methods concerns the nature of the data itself. Although it is convenient from a cartographic perspective to regard a line as an independently located feature with a true position, in reality many types of lines are subject to toplogical constraints, and are not independent of the areal features which they bound. For example a contour's position is not independent of other contours, since a large error in location may result in one contour crossing another. Contours are cartographic expressions of the value of some variable, often elevation, which is continuously distributed over the area. Problems with topological constraints on contour positions can be overcome if one regards error in contour position as an outcome of error in elevation, and concentrates on developing suitable models of elevation error instead. Fractional Brownian motion has been proposed as a suitable stochastic model of elevation (Mandelbrot, 1975, 1977, 1982; Goodchild, 1982; Goodchild et al., 1985; Mark and Aronson, 1984), in part because simulations using this

stochastic process bear striking resemblance to some types of real terrain.

For the purposes of this discussion we can divide choropleth data into two types. The first, which we will refer to as socioeconomic, arises in fields such as the Census when a continuous variable is summarized using defined reporting units. The "cookie cutters" or unit boundaries are located in most cases independently of the variable being reported; in fact they may be used to report several hundred different and possibly unrelated variables. Error modelling is likely to be difficult since the process leading to error in each boundary depends on the nature of the boundary; lines which follow streets are likely to have very different errors from lines which are defined to follow rivers, for example. For this type of data it seems appropriate to separate error in attributes from error in feature location, as several authors have done (MacDougall, 1975; Chrisman, 1982), and to attempt to model each separately.

The boundaries of a choropleth map form an irregular tesselation of the plane. The literature contains a number of methods for generating random tesselations which might form useful models of error in choropleth boundaries (Boots, 1973; Getis and Boots, 1978; Miles, 1964, 1970). All of them satisfy the necessary topological and geometrical constraints on boundaries. However none are sufficiently irregular in appearance to be acceptable as simulations of real choropleth boundaries.

If a suitable method for generating boundaries could be found, the second stage of the simulation process would be to distribute attributes over the polygons in some reasonable fashion. A random allocation is unacceptable on two grounds; it fails to reproduce the spatial autocorrelation of attributes observed on almost all maps, and allows adjacent zones to receive the same attribute. Goodchild (1980) and Haining, Griffith and Bennett (1982) have discussed the simulation of autocorrelation.

The focus of this paper is on the other type of choropleth data, which we refer to as natural resource data. In this case boundaries are intimately related to the variable being mapped, and are in most cases unique to it. For example the boundaries on a soil map occur along lines of change in soil type, and are unlikely to coincide with boundaries on any other coverage. Boundaries are inherently uncertain, and the level of uncertainty is related to the change in soil class which occurs at the boundary; it is easy to believe that a transition from class A to class B might be more readily determined on the ground than a transition from A to C, for example. Under such circumstances it seems clear that an error model which separates attributes from locations must be inadequate.

The next section of the paper describes the proposed model. We then discuss the implications of the model for the analysis and description of natural resource data, and its potential applications.

THE MODEL

Consider a number m of continuous variables $z_1, z_2, \ldots, z_m$ distributed over the (x,y) plane. The variables will probably show spatial autocorrelation, and may or may not be correlated. Now consider an m-dimensional space defined by these variables; we will refer to this

ts phase space by analogy to phase diagrams in thermodynamics. The space is divided into a number of domains, each of which is associated with one of a set of n classes:

$$C(z_1,z_2,\ldots,z_m) \in S \qquad (1)$$

where C is the class assigned to a point in phase space and S is the set of all possible classes. The domains provide a mapping from a set of m continuous variables to one of a set of classes. There may be more than one domain associated with a particular class, and some classes may not appear in the phase space. Finally, since the input variables are by definition continuous, it follows that if two zones share a common boundary on the choropleth map, then their corresponding classes must have been obtained from adjacent domains in phase space.

A simple model of world life zones by Holdridge et al. (1971) provides an illustration. Suppose that vegetation is largely controlled by temperature and precipitation variables, which have been mapped over the surface. Holdridge's diagram relating temperature and precipitation to vegetation class, reproduced in Figure 1, is a simple example of domains in phase space.

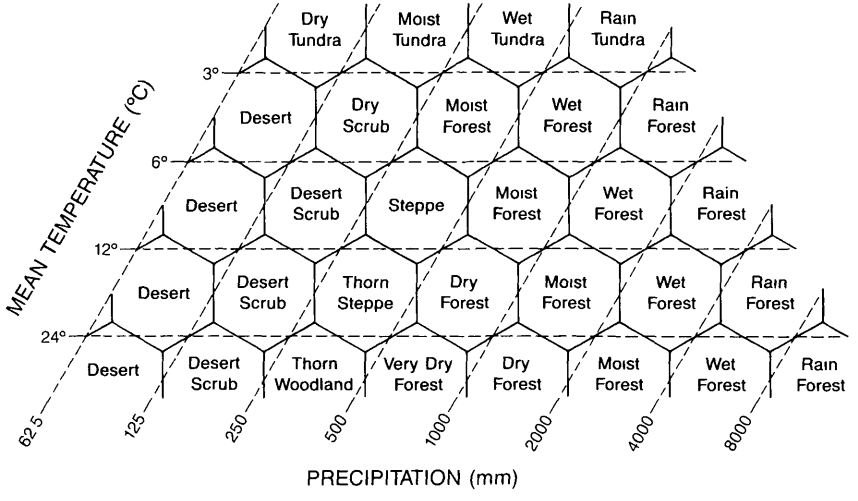

Figure 1. Example phase space for world life zone classification, from Holdridge et al. (1971).

If applied to the two input variables, it would map every combination of temperature and precipitation to a vegetation class, and thus convert two isopleth maps into one choropleth map. Errors in the choropleth map could then be ascribed to two sources: errors in the values of the continuous variables, and uncertainty in the delimitation of domains.

The visual appearance of the simulated choropleth map will clearly depend on the input surfaces. Highly irregular surfaces will produce highly fragmented choropleth zones, while smooth surfaces will produce large zones with relatively smooth boundaries, suggesting a

direct relationship between the degree of spatial autocorrelation of the input surfaces and the nature of the resulting map. For this reason we propose to use fractional Brownian surfaces as input variables, because they allow control over the level of spatial autocorrelation: a single parameter H can be varied to generate a continuum from very smooth (H=1) to very rugged (H=0) surfaces. A value of 0.7 has often been identified as giving the closest visual appearance to real terrain (Mandelbrot, 1977, 1982).

To illustrate the model, two surfaces were generated, at H=0.7 and H=0.6, and sampled with a 64 by 64 array. Each cell's values of z_1 and z_2 were mapped into the five-class phase space shown in Figure 2: the 4096 points are shown as dots. The resulting classified raster was vectorized to give the polygons shown in Figure 3.

It is likely that the boundaries produced by this simulation process are too irregular to be acceptable: they also show many isolated islands, which are rare on real maps. We suggest that these differences are the result of cartographic smoothings which take place during the drawing of choropleth boundaries. To allow for this, and also to remove the visual effects of pixel boundaries, we have added

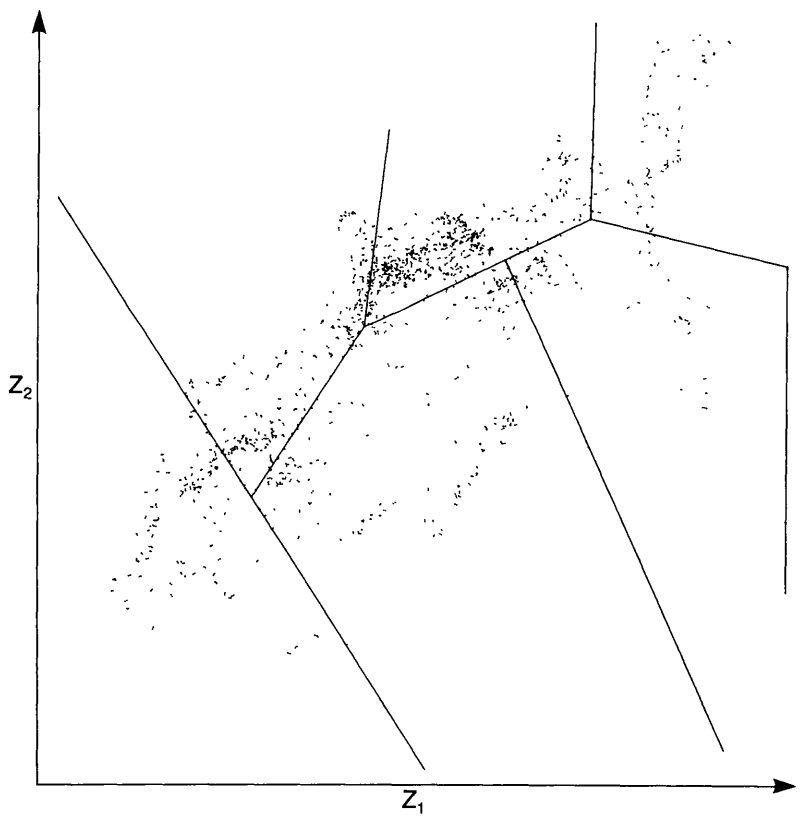

Figure 2. Phase space used in example simulation, with points from two 64 by 64 rasters.

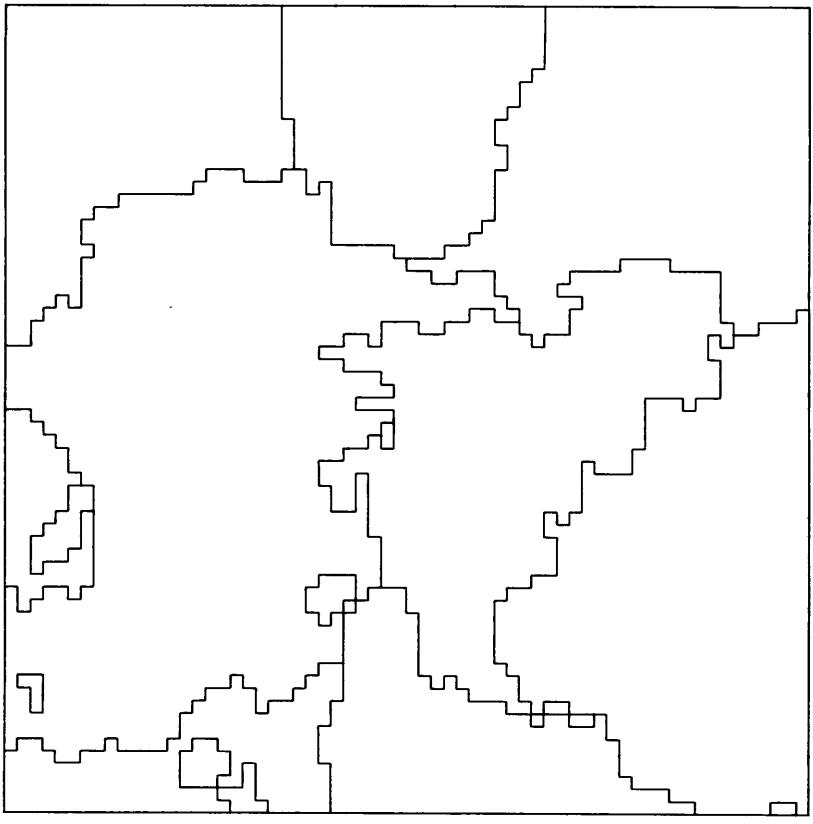

Figure 3. Classified 64 by 64 raster simulation.

two stages to the simulation process. First, the vectorization algorithm has been biassed against small islands. The normal criterion for contiguity is rook's case: a cell is not part of a larger choropleth zone unless at least one of its four rook's case neighbours is also part of the zone. However we allow an additional case: a pixel can be part of a larger zone if at least one of its bishop's case (diagonal) neighbours is also part of the zone, provided all of its four rook's case neighbours are part of some other, second zone. Second, we smooth the vectorized boundary between topological vertices by using a simple spline. This has the effect both of removing the pixel outlines, and also of reducing the irregularity of the line to emulate the cartographer's implicit generalization.

IMPLICATIONS OF THE MODEL

A contour map can be seen as a choropleth map in which the zones between every pair of adjacent contours are given a unique colour or class. In terms of our model, this choropleth map would be generated from a single variable, m=1, using a phase space of one dimension in

which the domains appear as divisions along the axis of that variable. Classes can be adjacent on the choropleth map, and have non-zero common boundary length, only if their corresponding domains are adjacent in phase space. It follows that there is a unique ordering of the classes such that when adjacencies are counted in a table in which the classes have been placed in the correct order in both rows and columns, the only non-empty cells will be those immediately adjacent to the diagonal.

The same property holds for two input variables if the domains are bounded by parallel lines, and similarly for more than two variables. If domain boundaries are parallel, it follows that some linear combination of the two input variables can be found, perpendicular to the domain boundaries, which would produce the same choropleth zones.

In terms of the model, the relative frequencies of adjacencies on a choropleth map are therefore an indication of the complexity of the phase space and the number of input variables, independently of the error or distortion of the data. For example, error can never produce an adjacency between two classes which are not adjacent in the underlying phase space. It can, however, produce an adjacency which was not previously present on the choropleth map but which is nevertheless present in phase space.

While the model replicates the observed crude characteristics of much natural resource choropleth data, we do not wish to imply that all such data is generated by processes of this type. The model seems reasonable as a mechanism for determining vegetation zones in relation to continuous, climatological variables, but no comparable continuous variables control bedrock geology or soil class. Some characteristics of choropleth data are clearly not replicated, such as the long, contorted polygons which follow rivers on maps of floodplains and related phenomena.

APPLICATIONS

The model provides a method for simulating choropleth boundary networks and associated attributes under a variety of conditions from small, fragmented zones to large ones and from highly irregular boundaries to smooth ones. We plan to use it to investigate a number of questions related to error and accuracy in choropleth maps, the answers to which are significant in the design and operation of geographic information systems.

First, the model will allow us to investigate the relationships between the accuracy of a spatial data base and the accuracy of measures derived from it, under a full range of conditions. For example, there is need for empirical work to examine further the effects of pixel size in raster data bases, and of digitizing errors and line generalization in vector data bases. The use of simulated rather than real data allows greater control over the characteristics of the data, and a wider range of experimental conditions.

Second, the model may provide a better understanding of the sources of error in choropleth maps. Error can occur at several stages in the simulation; in the measurement of the continuous variables, in the spatial sampling design (the density and position of the raster),

in the delimitation of domains in phase space, and in the vectorization and smoothing of polygon boundaries. Each of the various sources of uncertainty in a soil map boundary can be related to one or more of these sources. For example, uncertainty may be due to a low density of sampling of soils near the boundary, to inaccurate measurement of parameters such as soil colour, to subjective smoothing of the boundary by a cartographer, or to imprecision in the definition of soil classes. It is possible to simulate each of these separately, and to observe their effects. For example, error due to smoothing will produce uniform uncertainty for all lines, whereas error due to inaccurate measurement of one underlying continuous variable will produce degrees of error in boundary lines which are a function of the classes separated by the line, and depend directly on the slope of the relevant domain boundary in phase space.

Third, we can observe the effects of each source of error on GIS operations such as polygon overlay and sliver removal. Algorithms designed to remove slivers can be tested under a variety of conditions and forms of error.

One of the more desirable objectives of a study of error in spatial data bases would be the development of hypothesis tests to resolve such questions as whether a particular sliver polygon is real or spurious, based on its area or shape, or whether a point lies inside or outside a polygon. To do so would require a simple error model characterized by a very small number of parameters. The model proposed here is clearly not suitable; its parameters include the number and level of spatial autocorrelation of the underlying continuous variables, the spatial sampling design, the geometry of the phase space and the nature of the splining process. Although various simplifying assumptions might be made (for example that all boundaries in phase space are straight), there seems little prospect of calibrating a model of this complexity.

Greenland and Socher (1985) have proposed a simple measure of the degree of agreement between two versions of the same choropleth map. The proportion of area which has been assigned the same class on both maps, p_o, is compared to an expected proportion p_e in an index denoted by kappa. The basis for the calculation of p_e is the assumption that class is randomly allocated, in other words that the proportion of area allocated to class A on one map and to class B on the other is simply the product of the proportion which is A on the first map and the proportion which is B on the second.

If the maps show highly fragmented polygons, it is relatively easy for errors in boundary positions to produce agreements no higher than the expected proportion, and thus low values of kappa. But if the polygons are large, the same degree of boundary distortion will reduce kappa only slightly, and it will be almost impossible to find distortions which yield low kappa values. In other words, kappa is highly sensitive to the degree of spatial autocorrelation in attributes, and cannot be compared usefully across different types of data. To do so requires a more appropriate model of error. However given the variety of possible sources and forms of error in the model proposed in this paper, it is unlikely that a simple measure of data base distortion could be devised which would be valid across a range of data types.

REFERENCES

Blakemore, M. 1984, Generalization and error in spatial databases: Cartographica 21, (2,3), 131-139.

Boots, B.N. 1973, Some models of the random subdivision of space: Geografiska Annaler 55B, 34-48.

Chrisman, N. 1982, Methods of spatial analysis based on errors in categorical maps: Unpublished Ph.D. thesis, University of Bristol.

Getis, A. and Boots, B. 1978. Models of Spatial Processes, Cambridge University Press, London.

Goodchild, M.F. 1978, Statistical aspects of the polygon overlay problem: Harvard Papers on Geographic Information Systems 6, Addison-Wesley, Reading, Mass.

Goodchild, M.F. 1980, Simulation of autocorrelation for aggregte data: Environment and Planning A 12, 1073-1081.

Goodchild, M.F. 1982, The fractional Brownian process as a terrain simulation model: Modelling and Simulation 13, Proceedings of the 13th Annual Pittsburgh Conference, 1133-1137.

Goodchild, M.F., Klinkenberg, B., Glieca, M. and Hasan, M. 1985, Statistics of hydrologic networks on fractional Brownian surfaces: Modelling and Simulation 16, Proceedings of the 16th Annual Pittsburgh Conference, 317-323.

Greenland, A. and Socher, R.M. 1985, Statistical evaluation of accuracy for digital cartographic bases: Proceedings, AutoCarto 7, 212-221.

Haining, R.P., Griffith, D.A. and Bennett, R.J. 1983, Simulating two-dimensional autocorrelated surfaces: Geographical Analysis 15, 247-253.

Holdridge, L.R., Grenke, W.C., Hathaway, W.H., Liang, T. and Tosi, J.A. Jr. 1971. Forest Environments in Tropical Life Zones: A Pilot Study, Pergamon, Oxford.

Honeycutt, D.M. 1986. Epsilon, generalization and probability in spatial data bases, Unpublished manuscript.

MacDougall, E.B. 1975, The accuracy of map overlays: Landscape Planning 2, 23-30.

Mandelbrot, B.B. 1975, Stochastic models of the Earth's relief, the shape and the fractal dimension of the coastlines, and the number-area rule for islands: Proceedings of the National Academcy of Sciences 72, 3825-3828.

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

Mandelbrot, B.B. 1982. The Fractal Geometry of Nature, Freeman, San Francisco.

Mark, D.M. and Aronson, P.B. 1984, Scale-dependent fractal dimensions of topographic surfaces: Mathematical Geology 16, 671-684.

Miles, R.E. 1964, Random polygons determined by random lines in a plane: Proceedings of the National Academy of Sciences 52, 901-907, 1157-1160.

Miles, R.E. 1970, On the homogeneous planar Poisson point process: Mathematical Biosciences 6, 85-127.

Perkal, J. 1956, On epsilon length: Bulletin de l'Academie Polonaise des Sciences 4, 399-403.

Perkal, J. 1966, On the length of empirical curves: Discussion Paper 10, Michigan Inter-University Community of Mathematical Geographers, Ann Arbor.

Poiker, T.K. 1982, Looking at computer cartography: GeoJournal 6, 241-249.