(116)

Chapter 10

Modeling error in objects and fields

Michael F. Goodchild

Introduction

The current interest in spatial databases stems largely from their role in supporting geographic information systems, and the rapidly growing GIS industry. GISs are powerful systems for handling spatial data, and in recent years they have found application in fields as different as transportation, forestry and archaeology. Yet the power of a GIS to input, store, analyze and output geographic information of all kinds is at the same time a major liability. To the database, the structure used to store a polygon or pixel has almost no connection to the real meaning of the polygon, as a parcel of land, object on a topographic map or stand of timber. The analyst making use of a polygon overlay operation has similarly little pressure to be sensitive to the interpretation of the data layers being overlaid. In reality a GIS may encourage poor analysis by separating the data collection, compilation and analysis functions, and failing to make the user aware of the possible dangers of indiscriminate use of such functions as scale change, reclassification and overlay.

One of the more obvious issues from this perspective is the existence of two traditions of GIS analysis. The distinction between raster and vector is often seen as a problem of system design, but actually presents a major issue of data interpretation. To emphasize this difference, and to stress the context of data interpretation rather than system design, we will use the terms field and object in this paper, although they are to some extent synonymous with raster and vector respectively. Some spatial databases represent the world as if it were populated by objects - points, lines and areas - with associated attributes, continuing a tradition developed in cartography. Others represent the world as fields, or arrays of pixels, again with associated attributes. The choice between the two representations has variously been seen as depending on the method of data collection (satellites generate fields,cartographers generate objects), the degree of spatial resolution required (objects appear to imply higher levels of spatial resolution, whereas pixels imply a level which is fixed by the pixel size), and the efficiency of algorithms (for example, the widely held perception that overlay is faster in raster). However we will argue in this paper that fields and objects represent fundamentally different forms of abstraction of geographical reality.

This paper examines the relationship between fields and objects from the perspective of database error. It is clearly possible to represent a given set of data in either form, and to derive one from the other by a simple GIS operation such as raster/vector conversion. But conversion must be sensitive to the nature of the data and its uncertainty if subsequent analysis is to be successful.

Errors in objects

Consider the common process of creating a spatial database of objects by digitizing a topographic map. Attributes will likely be entered from a keyboard, and provided they are keyed correctly, we can reasonably expect them to be perfectly accurate. The locations of objects will be obtained by digitizing or scanning, and will be subject to assorted errors. A number of factors may contribute to distortion of the source document, including folding and stretching, changes in humidity, copying processes etc., and the process of map registration will introduce additional error. If digitizing is used, positional accuracy will be affected by the operator's precision in positioning the cursor, and by the rules used to select points to be digitized from line or polygon objects, whether in point or stream mode. Finally the positional accuracy of a scanned line will be affected by the resolution of the scanner.

Of all of these errors, only cursor positioning has been subject to successful analysis. Keefer, Smith and Gregoire (1988) have described a model in which each digitized point on a line or area object is distorted from its true position by a bivariate distribution. If each point were distorted independently we would reach the unreasonable conclusion that the expected error at digitized points is greater than between digitized points, but it is likely that errors are positively correlated between adjacent points along the digitized line. Chrisman and Yandell (1988) and Griffith (this volume) have obtained useful estimates of the expected error in polygon area measures based on this type of model of digitizing error. However it is much more difficult to devise a reasonable model of the process of point selection,which varies substantially between digitizer operators and types of lines, and likely contributes at least as much to error in area estimates. Unfortunately any such model would have to be sensitive to the type of line being digitized, as meandering rivers clearly present very different problems from topographic contours or highways, whereas this seems less important for modeling cursor positioning error.

In the case of a county boundary or the outline of a building, the database object corresponds directly to a clearly defined object in the real, geographical world. But in many cases the object in the database is an abstract model of real, continuous and complex geographic variation. Although counties and buildings are frequently found in spatial databases, many GIS applications have been developed for abstracted objects. In the forest industry, an inventory of forest resources is commonly maintained in digital form by dividing the forest into area objects or "stands" with descriptions which are attributed homogeneously to the entire stand. In reality the boundaries of stands are transition zones, and the attributes are heterogeneous. Despite the popularity of vector databases, the object model is often a poor representation of geographic variation.

Greater difficulty may arise if the attributes assigned to an object are themselves abstractions. For example the term ""old growth"" may be attached to a forest stand, but ambiguities in the definition of old growth may make it impossible to resolve whether a particular point within the stand is or is not covered by old growth. Similarly it may be necessary to observe a significant area in order to assign a classification such as"aspen parkland", which is not strictly an attribute of a point,but of an extended area.

Various terms have been used to distinguish between the two types of error implied by this argument. In this paper we use the term processing error to describe the uncertainty introduced by digitizing and any subsequent form of digital processing, such as vector/raster conversion. The term source error is used to describe the differences which may exist between the object model and the geographical truth which it represents. While source errors may be absent in representations of counties or buildings,we conjecture that they will exceed processing errors in representations of such geographical variates as vegetation, soil or land use. Since the results of a GIS analysis will be

interpreted by reference to geographic truth rather than to the source documents used to create the database, the existence of large source errors is a problem of major significance for the field. Several early papers (see for example McAlpine and Cook, 1971; MacDougall, 1975) drew attention to the importance of source errors.

The distinction between processing and source error is critical in estimating the accuracy of measures derived from a database. The error in the estimated area of a forest stand can be obtained by analyzing the errors introduced in processing its area object, and is independent of source errors implicit in the object. The area of a given species can be estimated by summing the areas of the objects classified as containing the species, but will be subject to source errors if the species attribute does not apply homogeneously to the area within each stand. Since an object database commonly contains no information about heterogeneity, it may be difficult or impossible to estimate source errors.

Maps of forest stands or soil types are compiled by a complex process which combines two forms of data. Information is first obtained on the ground from a series of point samples or transects, and then extended spatially using an aerial photograph or remotely sensed scene, or similar image. The area objects are compiled by interpreting a field, augmented by point attributes, so in these examples the object representation is clearly a derivative of the field representation. In compiling the map from the image, the cartographer imposes his or her own expectations on the data. The boundaries between area objects will be smooth generalizations of complex transition zones whose width will likely vary depending on the classes on either side. The holes and islands which one would expect in and around the transition zone will be commonly deleted. Finally, in the case of forest stands the cartographer may impose some concept of minimum size, since it is difficult to administer stands of less than, say, 10 hectares. In essence the cartographer acts as a low-pass filter, removing much of the detailed geographical variation in converting from field to objects. Unfortunately this removes much of the information on which an error model might be based, since error is selectively deleted by a low-pass filter, and suggests that we might focus on modeling errors in fields rather than derivative objects.

While some progress has been made in developing error models for digitized lines and areas, it has proven much more difficult to construct comprehensive models of processing and source errors for complex geographical objects. Chrisman (1982), Blakemore(1984) and others have described the uncertainty in the position of a line using a band of width ε (Perkal, 1956), which is a useful model of certain kinds of processing and source errors but does not deal with the problem of heterogeneity of attributes. Moreover while the epsilon band can be used to give a probabilistic interpretation to the point in polygon operation (Blakemore, 1984) it is not adequate as the basis for estimating error in area measures, or for simulating error in object databases to benchmark GIS error propagation. At this point we have no fully satisfactory means of modeling error in complex spatial objects.

Contours provide another example of the difficulty of modeling error in objects derived from fields. A model of object distortion might take each contour line and displace it using a random, autocorrelated error process, but it would be easy to produce topological inconsistencies such as crossing contours or loops. However contours are derived from an elevation field: no matter how elevations are distorted, the contours derived from them must always be topologically consistent. This suggests a general proposition - that the solution to modeling error in complex spatial objects may lie in modeling error in the fields from which many of them are obtained. In the next section we explore this possibility in more detail.

Field models of error in area objects

The previous section concluded with the proposition that satisfactory models of error in complex spatial objects could be formulated as models of error in fields. In this section we consider two such models.

Goodchild and Dubuc (1987) proposed a model based on an analogy to the effect of mean annual temperature and precipitation on life zones. We first generate two random fields, using one of a number of available methods, such as the fractional Brownian process, turning bands or Fourier transforms (Mandelbrot, 1982). The autocovariance structure of the fields can be varied to create a range of surfaces from locally smooth to locally rugged. Imagine that one field represents mean annual temperature and the other, annual precipitation.

Holdridge *et al.* (1971) have proposed a simple two-dimensional classifier which, given temperature and precipitation, yields the corresponding ecological zone. By applying such a classifier to the two fields, we obtain a map in which each pixel has been classified into one of the available zones. If the pixels are now vectorized into homogeneous areas, the map satisfies the requirements of many types of area class maps: the space is exhausted by non-overlapping, irregularly shaped area objects, and edges meet in predominantly three-valent vertices. On the other hand if a single field had been used with a one-dimensional classifier, the result would have the unmistakable features of a contour or isopleth map. To simulate the influence of the cartographer, which we have previously compared to the action of a low-pass filter, Goodchild and Dubuc (1987) applied a spline function to the vectorized edges, and selectively removed small islands.

The model has interesting properties which it shares with many real datasets of this class. Because the underlying fields are smooth, classes can be adjacent in the simulated map only if they are adjacent in the classifier, so certain adjacencies are much more common than others. The response of an edge to a change in one of the underlying fields depends on the geometry of the classifier: it is maximum if the corresponding edge in the classifier space is perpendicular to the appropriate axis, and minimum (zero) if the edge is parallel to the axis. So distortion or error can be simulated by adding distortion to the underlying fields.

The model successfully simulates the appearance of area class maps, and is useful in creating datasets under controlled conditions for use in benchmarking spatial databases and GIS processes, and for tracking the propagation of error. On the other hand the large number of parameters in the model make it difficult or impossible to calibrate against real datasets. Parameter values would have to be established for the underlying fields, the distorting field(s), the classifier, and the spline function used to smooth vectorized edges.

Goodchild and Wang (1988) describe an alternative model based on an analogy to remote sensing. Suppose that the process of image classification has produced an array of pixels, and that associated with each pixel is a vector of probabilities of membership in each of the known classes. For example the vector{0.3,0.3,0.4} would indicate probabilities of 0.3, 0.3 and 0.4 of membership in classes A, B and C respectively. In practice we would expect the proportion of non-zero probabilities to be small, allowing the vectors to be stored efficiently.

We now require a process of realizing pixel classes such that (1)the probabilities of each class across realizations are as specified, and (2) spatial dependence between pixels in any one realization. The degree of spatial dependence will determine the size of homogeneous patches which develop. With high spatial dependence, a given pixel will belong to large patches in each realization: in 30% of realizations the example pixel will be part of a patch of class A, 30% B and 40% C.

Goodchild and Wang (1988) described a simple process which satisfies only one of the requirements. Initial classes were assigned to each pixel by independent trials, and a 3x3 modal filter was then passed over the array, replacing the value of the central pixel by the modal value of the 3x3 window. While this ensures spatial dependence, the posterior probabilities are not equal to the priors except in special cases.

More recently, we have experimented with two methods which satisfy both requirements. In the first, we first generate a large number of realizations using independent trials. On each simulated map, we count the number of 4-adjacencies between unlike classes. We next execute a number of cycles to induce spatial dependence without at the same time changing the posterior probabilities. The two realizations with the lowest levels of spatial dependence (highest number of 4-adjacencies between unlike classes) are selected in each cycle. A random pixel is selected, and the contents of the pixel are swapped between the two selected realizations if the result would yield a higher level of spatial dependence. After examining a large number of pixels, another pair of realizations is selected. Because the method conserves the numbers of each class of pixel across realizations, while increasing spatial dependence, it clearly satisfies both of our requirements. Finally the simulated maps are vectorized and smoothed to create area objects.

In the second method, we make use of a simple spatially autoregressive process (Haining, Griffith and Bennett, 1983), to generate a random field of known distribution, and classify the result by comparing the value in each pixel to the prescribed probabilities. Since only two classes can be simulated, the method must be repeated n-1 times to develop a map of n classes.

Let $\mathbf{x}$ denote the random field. The spatially autoregressive process is defined as:

$$\mathbf{x} = \rho \mathbf{W} \mathbf{x} + \varepsilon \qquad (1)$$

where $\rho<0.25$ is a parameter of spatial dependence, $W_{ij}=1$ if i and j are 4-adjacent, else 0, and ε_i is a normal deviate of zero mean. We find $\mathbf{x}$ by inverting the matrix $(\mathbf{I} - \rho \mathbf{W})$. Unfortunately an array of n by n pixels requires the inversion of an n^2 by n^2 matrix, but it is possible to do this for arrays as large as 64 by 64 by taking advantage of the block structure of the W matrix. Given the known distribution of x_i across realizations, we can compute $P(X<x)$ and compare it to the pixel's specified probability p_i.

Certain types of prior information can be introduced into both of these models in order to broaden their applications. For example, suppose there exists a predefined "parcel" of known boundaries, and it is suspected that the class or classes within the parcel are independent of those outside. The parcel boundary is adjusted to pixel edges: $W_{ij}=0$ if i is inside the parcel and j is outside, even though i and j may be 4-adjacent. This has the effect of removing spatial dependence between the parcel and its surroundings. The homogeneity of the parcel depends on the magnitude of ρ; if ρ is sufficiently large, the parcel will act as an object whose attribute is determined by a single trial. Geman and Geman (1984) have described an edge process with similar objectives.

The vectors of probabilities required by this process are readily obtainable from many remote sensing classifiers. Conventionally, pixels are classified by maximum likelihood even though the classifier yields a complete vector. This results in the loss of valuable information on uncertainty, and leads to severe bias in derived estimates of area, particularly for large patches. The model has only one parameter, ρ. Its value might be established by calibration against ground truth, or might be set to reflect expectations about patch size and map complexity. Inverses of $(\mathbf{I}- \rho \mathbf{W})$ might be precomputed for various values of ρ, and multiplied by various ε to obtain multiple realizations. In this way it would be possible to simulate error rapidly for any classified image. Although the first method above based on swapping between realizations is conceptually simpler, it does not lend itself as readily to precomputing, and so would be more difficult to implement in practice.

Discussion

The process by which a spatial database is created from a source map is complex, and error of various types is introduced at each step. Some of the error components can be modeled, and progress has been made in analyzing the errors due to cursor positioning, but others such as point selection are more difficult, and the goal of a comprehensive model of spatial data processing errors is still elusive. Yet despite this, for most types of spatial data the errors inherent in the source document are clearly more significant than those introduced by processing. This is particularly true when the source document contains objects which are approximate abstractions of complex and continuous spatial variation.

We have argued in this paper that many of the more abstract types of spatial objects have been obtained or compiled from raw data in the form of fields, and that the process of compilation often removes much of the information on which a useful model of uncertainty or error might be based. The role of the cartographer in compiling vegetation or soil maps was compared to the action of a low-pass filter in selectively removing the high spatial frequencies which contain diagnostic information on uncertainty, such as wiggly lines and small islands.

From a spatial statistical point of view, the central proposition of this paper is that uncertainty in the objects on a map results from different outcomes of a stochastic process defined for a field. This is substantially different from the approach which has underlain much work on stochastic processes for images. The problem of image restoration has the objective of finding the"true" value for each pixel given some distorted value. However in the geographical case there is usually no comparable notion of true value, because spatial variation extends to all scales, and because class definitions are frequently ambiguous. Similarly, although spatial statistics contains extensive literature on image segmentation, this problem is seen as one of a range of possible models for the cartographic process of forming objects from fields.

From the perspective of spatial databases and GIS, there are several possible roles for a model of error in spatial data. On the one hand it would be useful to have a calibrated model which could be used to describe error in a particular dataset, to track the error through GIS processes, and to report uncertainty in the results of processing. However such models may be unobtainable in many cases, due to the lack of adequate information for calibration, and to the complexity of the error process itself. There seems to be no equivalent with the generality of the Gaussian distribution for complex spatial objects. On the other hand models of error are useful for generating simulated datasets under known conditions, for benchmarking GIS processes and storage methods.

Abstract point, line and area models developed in cartography because of the need to represent complex spatial variation on paper using images which could be created with a simple pen (Goodchild, 1988). DEMs, TINs and quadtrees are a few of the new data models which have been developed for spatial databases to take advantage of the removal of cartographic constraints. The contour was devised as an efficient way to display spatial variation of elevation on a topographic map with a pen capable of drawing lines of fixed width, but DEMs and TINs have distinct advantages over digitized contours in terms of sampling efficiency and the execution of various analytic operations. We have seen in this paper that the selection of the cartographic model also has the effect of removing information on uncertainty, since it is easier to model error in fields than in derivative objects. A database populated by fields is more useful for modeling and tracking uncertainty than one populated by abstracted objects. From an accuracy perspective, then, it would be preferable if databases representing spatial variation of parameters such as soil type, vegetation or land use contained the raw point samples and images from which such maps are usually compiled. If the cartographic view of the database were required, it could be generated by interpreting areal objects, either interactively or automatically. But this strategy would avoid the common practice of imposing the cartographic view as a filter between the database and the reality which it represents.

References

Blakemore, M., 1984, Generalization and error in spatial databases. *Cartographica* **21**, 131-9.

Chrisman, N. R., 1982, Methods of spatial analysis based on errors in categorical maps. Unpublished PhD thesis, University of Bristol.

Chrisman, N. R., and Yandell, B., 1988, A model for the variance in area. *Surveying and Mapping* **48**, 241-6.

Geman, S., and Geman, D.,1984, Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-**6**, 721-41.

Goodchild, M. F., 1988, Stepping over the line: technological constraints and the new cartography. *American Cartographer* **15**, 311-20.

Goodchild, M. F., and Dubuc, O., 1987, A model of error for choropleth maps, with applications to geographic information systems. Proceedings, *AutoCarto 8*. ASPRS/ACSM, Falls Church, VA, 165-74.

Goodchild, M. F., and Wang, M.-H., 1988, Modeling error in raster-based spatial data. *Proceedings, Third International Symposium on Spatial Data Handling*. IGU Commission on Geographical Data Sensing and Processing, Columbus, Ohio, 97-106.

Haining, R. P., Griffith, D.A., and Bennett, R.J., 1983. Simulating two-dimensional autocorrelated surfaces. *Geographical Analysis* **15**, 247-55.

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

Keefer, B. J., Smith, J. L., and Gregoire, T. G., 1988, Simulating manual digitizing error with statistical models. Proceedings, *GIS/LIS '88*. ASPRS/ACSM, Falls Church, VA, 475-83.

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

McAlpine, J. R. and Cook, B.G., 1971, Data reliability from map overlay. In *Proceedings, Australian and New Zealand Association for the Advancement of Science*, 43rd Congress, Brisbane, Section 21, Geographical Sciences.

Mandelbrot, B. B., 1982, *The Fractal Geometry of Nature*. (San Francisco: Freeman)

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