1-1

MODELING ERROR IN SPATIAL DATABASES

Michael F. Goodchild
NCGIA
University of California
Santa Barbara, CA 93106

ABSTRACT

Geographic Information Systems process spatial data to generate information products through such operations as overlay, area measurement and buffer zone generation. The paper reviews the techniques available for modeling error in the various data models commonly used in GIS, and the relationships between them. Methods are discussed for incorporating error model parameters in the database, for tracking the propagation of error through computational processes and for reporting uncertainty in system products. The paper provides a general introduction to the research completed under the NCGIA's Initiative 1: Accuracy of Spatial Databases.

INTRODUCTION

Geographical variation is infinitely complex, and must be generalized, abstracted and simplified to fit into the finite storage space offered by a digital database. In addition it is impossible to measure position or the values of variables on the earth's surface with perfect accuracy, because of the limitations of instruments. All geographical data is therefore uncertain, and it is important to devise ways of describing uncertainty, displaying it and determining its effects on the products of analysis. NCGIA's Initiative 1 on Accuracy of Spatial Databases focussed on these issues, and the research necessary to deal with them effectively.

One of the problems facing any comprehensive approach to data uncertainty in GIS is the wide range of data models commonly employed in this technology. Peuquet (1984) defines a data model as "a general description of specific sets of entities and the relationships between these sets of entities". Date's definition (Date 1983) is more general: "The purpose...is...to provide a formal means of representing information...". In the case of geographic information, data modeling is the process of abstracting and generalizing, or discretizing real geographical variation, as raster cells, digitized contours, etc. The entities that populate many non-spatial databases exist in reality, e.g. individuals, families, or parts in a production process, but in many forms of geographical data they are imaginary abstractions.

We define the accuracy of a geographical database by comparing the database's response to a query, e.g. the soil type present at a specified point, with ground truth. This has several important implications. First, for many geographical variables, ground truth is itself uncertain. It may be impossible, for example, to visit a specified location and determine the local soil type with certainty, because of the subjectivity of soil classification and uncertainty in geographical position. In such cases it seems reasonable to include this uncertainty in the measure of database accuracy. Second, by this definition accuracy is a larger issue than simply the relationship between the contents of the database and the source document, probably a map: the errors inherent in the map itself are included. It is common to distinguish between the two types, by referring to errors in the map as source errors and those introduced by digitizing and analysis as processing errors. Third, the definition implies that a GIS is a means of accessing and analyzing real geographical distributions, rather than simply the contents of maps.

The first part of the paper describes the data models in common use in GIS. Each of these is then discussed from the perspective of error modeling and analysis. Some data models are found to be much more amenable to error modeling than others, and the paper ranks data models on this basis. The final section of the paper describes a prototype GIS with full error modeling and propagation capabilities.

GEOGRAPHIC DATA MODELS

The purpose of geographic data modeling is to discretize real, observable geographic variation, thus expressing it in a suitable form for representation in a discrete, digital store. Goodchild (1990) distinguishes between two radically different approaches, termed the layer and object views (for the purposes of this paper we will not cover modeling of phenomena distributed over networks, e.g. rail, road, or hydrologic, where variation occurs along a 1-dimensional manifold embedded in a 2-dimensional plane). The layer view sees geography as a set of variables, each observable at every point on the land surface, e.g. soil type or elevation. Each variable is a single-valued function of location. Each may or may not have continuity of value (there may be cliffs), and may or may not be differentiable (there may be breaks of slope at e.g. ridges or channels). The object view sees geography as an empty space occupied by objects. The objects have attributes, but the values of the attribute variables are defined only for objects. Any point in the space may be empty, or be occupied by one or more objects. An object is indivisible, so its attributes do not apply to its geographical parts. The geographical form of an object may be different at different scales. "The Bay Area" is an example of an object. It may appear as an area object at some scales and as a point at others. Points located within the extent of the area object do not have the properties of "The Bay Area", e.g. population - only the entire object. Points can be simultaneously within the objects "San Francisco", "The Bay Area" and "Northern California".

For the object view, the question of accuracy can be resolved into two issues: accuracy of location, and accuracy of attributes. This fundamental separability lies behind most attempts to define standards for cartographic data, where the objects are map features and are inherently well-defined (DCDSTF 1988). This paper concentrates on the layer view because of its greater importance to GIS.

Layer data models

Current GIS technology uses at least six different ways of representing the geographic variation of a single variable (Table 1). They can be grouped according to whether the variable is measured on a discrete (e.g. soil type) or continuous (e.g. elevation) scale (note that this refers to continuity of the variable, not of the geographical space, which is always continuous in the layer view). Two are important for discrete data: the raster model, where variation is described by determining its value within each rectangular cell; and the polygon model, where the plane is divided into irregular polygons. The raster model approximates by ignoring variation within cells, while the polygon model ignores variation within polygons and also often removes some of the geometrical complexity of boundary lines.

Six data models are common in representing continuous variables. Many DEMs represent elevation using a rectangular array of point samples. In remote sensing, variation in an image is represented as a mean value in each raster cell. Although the models are similar in form, the interpretation of the data is different in each case. A problem for GIS is that these models can be represented internally using virtually identical data structures, inviting the user to analyze them using the same processes and model error in similar ways despite the fundamental differences in meaning. Continuous variables can also be represented using irregularly spaced sets of sample points. Digitized contour lines are a common form of discretization of continuous variables in GIS. The polygon model is used frequently for social and economic variables: population density, for example, can be represented using irregular polygonal reporting zones together with the mean value of the surface within the polygon, or its integral over the polygon (volume under the surface). Finally, the TIN model is an increasingly common data model for continuous variables, particularly elevation.

Table 1: Summary of alternative GIS data models in common use

Model		Objects created
Discrete:		
Raster	Y	Pixels
Polygon	Y	Polygons
Continuous:		
Regularly spaced sample	N	Pixels
Mean value raster	Y	Pixels
Irregularly spaced sample	N	Points
Digitized contours	N	Lines
Polygon	Y	Polygons
TIN	Y	Polygons

With some data models, it is possible to query the value of the variable at every point on the plane. But with others, values can be obtained at arbitrarily located points only through the use of an additional process. For example, even though digitized contours model a continuous surface, the data model defines the value of the variable only at points on contours - elsewhere it is necessary to invoke a suitable spatial interpolation procedure. There is some standardization of such procedures in the case of regularly spaced samples (a plane fitted to a 3x3 window) but much variability in the case of irregularly spaced samples (e.g. Kriging, distance weighted interpolation, NURBS) or contours. Only the data models labeled Y in Table 1 are fully layer models in this sense. Moreover, because interpolators are not standardized, it is much more difficult to tackle the accuracy problem for the N models, as the error in an estimate of elevation from e.g. an irregularly spaced sample model depends on the interpolator used.

Table 1 also shows the type of objects created by the discretization inherent in each model. In many systems, the method of internal representation of data depends only on the type of objects, so all of the models producing polygons - TINs, discrete and continuous polygon models, and sets of area objects from the object view, will be treated similarly using the same data structure. This tendency to group different data models into common data structures is confusing, and can lead to inappropriate use of algorithms. For example, a set of network links and a set of contours might both be treated as collections of lines, yet there would be no meaning to a routing algorithm on a set of contours, or a spatial interpolation on a set of network links. Many of these problems arise because the set of data models available in a system is incomplete. If a data model is missing but the system is able to handle the type of object created by it, then it may be possible to process the data using another model with the same type of object: for example, contours might be represented as network links.

In many current systems, the sets of polygons created by the object and layer views are distinguished by the use of planar enforcement. In such cases, the data structure used for polygons in the layer view differs by requiring all points in the plane to lie in exactly one polygon. It is then efficient to store polygons as collections of arcs. Problems arise, of course, in using planar enforced data structures to represent polygons in the object view, where points may lie in any number of polygons. Planar enforcement can also be imposed in the object view in the form of integrated topology (e.g. in Integraph's TIGRIS). Here all area objects are represented as one layer: given a point, the database will return a list of all objects enclosing it. Line and point objects also carry pointers to enclosing area objects.

In the next section, we discuss the issue of accuracy in the context of the layer models. The N models of Table 1 are not discussed because of the problem noted earlier, that accuracy depends on a method of spatial interpolation.

As noted earlier, the accuracy of models such as digitized contours is regularly assessed using measures of cartographic data quality, but these object-oriented measures are not useful as estimates of the relationship between the value of a variable and ground truth at an arbitrarily located point. This leaves three models: raster, polygon and TIN, the first two for both discrete and continuous variables.

MODELS OF ERROR

The simplest kind of query from a discrete-variable database requests the class present at a location. Thus a straightforward approach to uncertainty is to measure the probability that the ground class at that point is not the database class. The misclassification matrix popular in remote sensing is an appropriate approach, and the Kappa index a suitable measure of uncertainty (Congalton and Mead 1983; Rosenfield and Fitzpatrick-Lins 1986). For continuous-variable data the standard error is an appropriate statistic, measuring the root mean square difference between the ground value and the value obtained from the database in units of the variable.

While these are useful statistics for summarizing the error of the map as a whole, they are of little use on their own in predicting the uncertainty of GIS products. For example, slope and aspect are often computed from DEMs by estimating surface derivatives. The polygon model would be useless because slopes are everywhere zero and aspect indeterminate. But slopes are readily available from the triangular facets of the TIN model, and can be obtained from the regularly spaced sample model by fitting planes to 3x3 windows (implying a method of spatial interpolation). But in all cases slope depends on more than one value, and so error in slope requires knowledge of the spatial correlation of errors, which is almost certainly non-zero. The same argument applies to estimating uncertainty in almost any derivative product of a DEM, including areas of intervisibility and inferred hydrology.

A similar case for knowledge of spatial dependence applies to the discrete-variable case. A misclassification matrix alone is not sufficient to estimate errors in such GIS products as overlay. Suppose 25% of a map is misclassified, and overlaid with another map that is similarly misclassified. To estimate the accuracy of the overlay, we need to know the probability that any point is correct on both maps, correct on one of the maps or incorrect on both. These probabilities are different depending on whether errors are spatially independent and uncorrelated between maps (probability .25x.25 that a point is wrong on both maps), or show some degree of either correlation between maps (probability .25 under perfect positive correlation, 0 under perfect negative correlation), or spatial dependence. An effective system for propagating error through GIS operations would have to be aware of the amount of spatial dependence present on each map. Moreover this may be a function of class: errors may occur in small patches on one class, and larger

patches on another, indicating different structures of spatial dependence.

In the pixel-creating data models, the geometric form of the objects is independent of the geographical distribution of the variable. Errors in the variable therefore do not affect the geometry of the objects. But in the polygon- and line- creating models, the objects' geometries are determined by the phenomenon, and are thus affected by uncertainty. Error in the determination of elevation can move a contour or a TIN triangle vertex or edge (e.g. many vertex selection algorithms for TINs will select spurious pits and peaks in areas of low relief), and error in determining soil type can move the boundary of a polygon. From a statistical rather than a data modeling perspective, Goodchild (1989) calls these fields and objects respectively, and argues that it is much more difficult to model error in objects because of the interaction between variable and geometry. Even though contours and soil class boundaries are line objects, it is not possible to separate locational and attribute uncertainty, since the line objects are abstract creations of the data modeling process. Error in the variable's value at a point, which is the central concern of accuracy in GIS, is only indirectly related to uncertainty in the positions of these line objects.

From the point of view of error modeling, therefore, it is possible to rank the alternative GIS data models in a clear order. The N models in Table 1 - regularly and irregularly spaced sample points and digitized contours - are lowest in order because they require a specified spatial interpolation procedure to be useful as GIS data models. The latter two are especially problematic because any spatial interpolation procedure will produce highly non-uniform spatial patterns of uncertainty, with lowest errors close to the sample points and contour lines. Of the Y models, polygon and digitized contours are less useful because of the non-separability of locational and attribute errors, the issue of locational error being peripheral to GIS accuracy.

GIS makes abundant use of discrete-variable data - soil type, land use/land cover, etc., and many popular GISs are incapable of handling any other type. Table 2 shows the summary ranking of data models from the perspective of simplicity of error modeling, and is in sharp contrast to the relative popularity of many of these models in current databases. The remainder of the paper focuses on the highest-ranked, the discrete-variable raster.

Table 2: Summary ranking of GIS data models from the perspective of simplicity of error modeling.

Data model	Rank
Discrete:	
Raster	1
Polygon	2
Continuous:	
Raster	1
TIN	2
Polygon	3
Regularly spaced	4
Irregularly spaced	5
Contour lines	6

ERROR MODELING IN A DISCRETE-VARIABLE RASTER

Goodchild and Dubuc (1987) and Goodchild (1989) describe two feasible models for spatially dependent errors in discrete-variable rasters. Of these the first model is generally infeasible due to the large number of parameters involved, although it may be a more realistic model of the actual map compilation process, but the second has only one parameter, of spatial dependence, and is briefly summarized here. Suppose it is known that a pixel has probabilities $\{p_1,p_2,...,p_i,...\}$ of being class 1,2,...,i,... on the ground. These probabilities might be the result of an image classifier, or might be known from the legend of a soil map, or might be subjective estimates of mixtures or uncertainties of interpretation (the Canada Land Inventory, the source of the CGIS database, has this type of attribute in its major layers). It is possible to generate a series of 'maps' from these probabilities by assigning classes to each pixel under two conditions: 1) that over many such 'maps', the probability that a given pixel is class i is p_i as input, and 2) that the class assigned to a pixel is correlated with the class assigned to its neighbors, to a degree given by a user-defined spatial dependence parameter. Suppose, for example, that a patch is class A with probability .8 and class B with probability .2. Each simulated 'map' will show an area of predominantly class A, but with inclusions of class B, the average size of inclusions depending on the value of the parameter. Although there is indirect control of the size of inclusions, there is no control over their locations within the patch, and the process does not allow us to simulate the effects of correlations of error between maps (note the conflict here with the notion of transition zones, which assigns greater uncertainty to the areas near the patch edge). On every simulation the inclusions will occupy about 20% of the patch area, and have the same average size, but there will be no correlation in their locations from one simulated 'map' to another.

IMPLEMENTATION

Implementation of the approach, which is limited to

discrete-variable rasters, requires the following:

1. A means of storing layers of probability vectors, e.g. as output from probabilistic classifiers in image processing systems.

2. Its vector equivalent, a means of storing planar-enforced area objects, with attributes of probability vectors, e.g. as obtained from a digitized soil map.

3. A suite of templates for selected values of the spatial dependence parameter (details of the generation of templates are given in Goodchild 1989). Each template consists of a layer of a spatially autoregressive random variable. The suite of templates can be used to create a series of simulated 'maps' for a given input layer and value of the parameter. A suite of 10 is normally adequate to obtain a meaningful estimate of product uncertainty.

4. A command to convert an input layer of probability vectors into a suite of simulated maps given a user-supplied value of the spatial dependence parameter.

5. Commands to perform such operations as overlay on entire suites of simulated maps rather than single layers.

6. Functions to compare products from a suite of simulated maps and compute the implied uncertainty in output measures.

For example, a query such as "compute the area that is class A on layer 1 and class F on layer 2" is normally handled in a raster system by cell-by-cell comparison. In this system, two suites of 'maps' are be simulated, and overlaid a pair at a time. Areas are computed for each overlaid pair, and the system compares the area estimates across the suite, expressing the result in the form of a mean and estimated standard error.

CONCLUSIONS

The issue of accuracy brings a radically different perspective to the issue of error modeling, and calls into question the usefulness of such data models as TINs and digitized contours for representing geographical variation in a GIS database. Despite their inherent simplicity, the pixel-based models appear to have distinct advantages in dealing with uncertainty. The prototype error-propagating system described here has inevitable disadvantages in storage requirements and processing time over more conventional approaches, but provides objective estimates of the uncertainty inherent in every GIS product. By recognizing the importance of spatial dependence explicitly, it goes much further than models based on propagation of spatially independent errors.

REFERENCES

Congalton, R.G. and Mead, R.A. 1983, A quantitative method to test for consistency and correctness in photointerpretation: *Photogrammetric Engineering and Remote Sensing*, Vol. 49 No. 1, pp. 69-74.

Date, C.J. 1983, *An Introduction to Database Systems*, Vol. 11, Addison-Wesley, Reading MA.

Digital Cartographic Data Standards Task Force (DCDSTF) 1988, The proposed standard for digital cartographic data: *The American Cartographer* Vol. 15 No. 1.

Goodchild, M.F. 1989, Modeling error in objects and fields: in M.F. Goodchild and S. Gopal, editors, *Accuracy of Spatial Databases*, Taylor and Francis, London, pp. 107-113.

Goodchild, M.F. 1990, Geographical data modeling: *Computers and Geosciences* (to appear).

Goodchild, M.F. and Dubuc, O. 1987, A model of error for choropleth maps, with applications to geographic information systems: in *Proceedings, AutoCarto 8*, ASPRS/ACSM, Falls Church, VA, pp. 165-174.

Peuquet, D.J. 1984, A conceptual framework and comparison of spatial data models: *Cartographica*, Vol. 21 No.4, pp. 66-113.

Rosenfield, G.H. and Fitzpatrick-Lins, K. 1986, A coefficient of agreement as a measure of thematic map accuracy: *Photogrammetric Engineering and Remote Sensing*, Vol. 52 No. 2, pp. 681-686.

GIS TRANSFORMATIONS

Waldo Tobler
National Center for Geographic Information and Analysis
University of California
Santa Barbara CA 93106

ABSTRACT

If one considers GIS information to be interpretable as point, line, area, or field phenomenon, then the sixteen common classes of transformation are:

point -> point	point -> line	point -> area	point -> field
line -> point	line -> line	line -> area	line -> field
area -> point	area -> line	area -> area	area -> field
field -> point	field -> line	field -> area	field -> field

Within these classes one can further distinguish between categorical (binary, n-ary) and numerical (scalar, vector, tensor) data, to obtain some eighty distinct possible classes of transformation. The common vector <-> raster conversions can also be added to this list. It is further useful to distinguish between locative and substantive transformations, but it makes sense to neglect the ubiquitous and necessary, but uninteresting, conversions between formats. In addition, many analysis techniques can be considered transformations; for example, computing terrain slope from topographic data by applying the derivative operator converts a scalar field to a vector field. An attempt is made to enumerate, via example, many of these transformations, and to comment on implementation algorithms.

The usual cartographic paradigm as related to a Geographic Information System considers geographic phenomena to be composed of points, lines, or areas. This is obviously an oversimplified view of reality, but has some merit when looked at from the point of view of the draughtsman. Flows and surfaces are often appended to the point-line-area classes as a crude recognition of the inadequacies of the three-fold categorization. As a slightly more advanced point of view one can consider phenomena to be located at a point, or along a line, or distributed over an area. An alternative is to view the phenomena from the point of view of the measurement scale. Thus one obtains the Nominal, Ordinal, Interval, or Ratio types of data. In the physical sciences data are often treated as fields. The term field is used in physics to describe any collection of numbers, each of which is attached to a position in space. More generally, a field in geography can be used to denote any phenomena which has an assigned property (which may be null or zero) at every terrestrial location. Thus we can have categorical fields, such as land use, or binary fields, such as urban or rural. Many fields, such as elevation above (or below) a datum, are usually given as numerical quantities. If there is one number at each location the field is usually known as a scalar field; two numbers at every place are often represented by a vector field. The common wind field is a well known example. Color is often treated as a three component vector. A two by two matrix at every geographic location, such as a strain or stress matrix, is a second order tensor field. Consider a wind-rose, a diagram showing the frequency by direction of winds at a place over some time interval. The number of directions around the complete circle is obviously infinite, and there are an infinity of