MODELING ERRORS FOR REMOTELY SENSED DATA INPUT TO GIS

Michael F. Goodchild
Natioal Center for Geographic Information and Analysis
University of California
Santa Barbara, CA 93106

Wang Min-hua
Department of Geography
University of Western Ontario
London, Ontario N6A 5C2, Canada

ABSTRACT

Different views of spatial resolution and accuracy present a major obstacle to the integration of remote sensing and GIS. Accuracy in remote sensing is modeled using probabilities of class membership in each pixel; in vector-based GIS it is modeled using concepts such as the epsilon band. The problem of linking the two views of accuracy reduces to one of realizing a stochastic process which must satisfy conditions of prior and posterior probabilities, and spatial dependence. We propose two suitable methods, one storage intensive and the other computationally intensive. The methods can be adapted to incorporate various forms of prior knowledge.

INTRODUCTION

Remotely sensed imagery provides a fast and efficient means of collecting large volumes of information about the earth's surface. Raw spectral responses can be registered, corrected and interpreted using sophisticated image processing systems, and a variety of methods of pixel classification have been developed to transform imagery into rudimentary maps for such themes as land use or vegetation cover. The response recorded for each pixel in a particular band is an integral over the area of the pixel of a continuous, spatially autocorrelated variable, and it is common to think of response data as a random sampling of a continuous surface or field. On the other hand a classified image can be conceptualized as an array of discrete values in which each pixel has been assigned to one of a number of classes.

A GIS can be defined as a system for input, storage, analysis and output of spatial information. As such, its main strengths lie in its ability to give the user access to an apparently scale-free, seamless electronic map, to analyze simultaneously different layers or coverages of the same area, to measure the lengths and areas of geographical objects, and to allow easy updating and editing. The capabilities of a GIS can greatly extend the usefulness of a classified, remotely sensed image by allowing access to other data either to improve the accuracy of classification, or to enhance the range of possible analyses. On the other hand remote sensing has much to offer GIS as a source of

easily updated and low cost input data. For these reasons numerous attempts to integrate remote sensing and GIS have been described in the literature.

Vector-based GISs model the world as populated by objects, specifically classes of points, lines or areas. Land cover is often modeled in such a system as a class of non-overlapping areas which exhaust the study space, each area being associated with one or more attributes which describe its land cover class. In practice the use of this model is largely independent of the means by which the data was acquired, whether by digitizing the lines on an existing map of land cover, scanning an existing map and vectorizing the resulting image, or using an image processing system to classify and vectorize a remotely sensed image. However the appearance of the data may reveal the source, as pixel edges will likely still be evident in a layer obtained from remote sensing, unless the lines have been subsequently smoothed.

Although there are undoubtedly significant technical problems in interfacing remote sensing and GIS, we wish to argue in this paper that the conceptual problems of interfacing systems which view the world respectively as fields and objects are in the long run more challenging, and will be a more substantial obstacle to the use of remotely sensed images in object-based systems. Our purpose in this paper is to explore the implications of such interfacing from the perspective of the interrelated issues of spatial resolution, error and accuracy. More specifically, the paper examines the extent to which concepts of error in imagery can be related to corresponding concepts of error in objects. The paper expands on work previously described by Goodchild and Wang (1988).

The next section reviews recent efforts to deal with the problem of uncertainty in object-based GIS. This is followed by a discussion of error in classified imagery, and by a review of techniques which can be used to interrelate these two views of the accuracy of spatial information.

ERROR IN OBJECT-BASED GIS

The use of high precision digital processing on data of undetermined accuracy has inevitably raised awareness of the problems of error in spatial data handling in recent years (see for example Walsh, Lightfoot and Butler 1987; Burrough 1986), besides leading to specific artifacts such as sliver polygons (Goodchild, 1979) and conflicts between geometry and topology (Franklin, 1984). Problems are made particularly acute by the ease with which a GIS can be used to change the scale of data without a corresponding change in its spatial resolution, and by the degree to which GIS processing of data from multiple sources distances the user from the data collection and interpretation process. As a result the users of GIS products are often unaware of the uncertainties and caveats which surround any spatial information.

Statistical models of the uncertainty in the locations of

point objects are well developed in surveying and geodesy. However their extensions to more complex line and area objects are not straightforward for several reasons. A model of the relationship between a true line on the ground and its representation as a series of digitized points and connecting straight lines in a spatial database must include not only the correlations which exist between errors at neighboring points (Keefer, Smith and Gregoire 1988), but also the process by which the points themselves were selected by the digitizer operator.

Despite these difficulties, simple approaches to describing accuracy of object representations can be found in the concepts of tolerance and error bands used in many digitizing and overlay systems. The Perkal epsilon band (Perkal 1956; Blakemore 1984) is a buffer of width epsilon on either side of a line or polygon boundary. In digitizing, two lines can be assumed to join and are consequently 'snapped' together if one lies within the other's epsilon band; similarly, in overlay, a line on one map which lies within the epsilon band of a line on the other map is assumed to represent the same line on the ground, and any associated sliver polygons are therefore removed. Unfortunately this deterministic view of the epsilon band can produce unwanted results in the following way. Line A can be found to lie within line B's band, indicating that A and B are the same; A can lie in C's band indicating that A and C are the same; but C can lie outside B's band. In this situation it is easy to generate inconsistencies, particularly if the positions of objects are adjusted in operations such as snapping. A probabilistic version of the concept could potentially resolve such problems.

The process of digitizing tends to result in errors and distortions which are substantially constant over a map, and depend only on the scale at which the map was digitized. On the other hand other, often more significant sources of error are unfortunately not as constant. It is common to distinguish between processing errors, which include those introduced during digitizing, and the source errors which exist between the source document and the reality which the document models. In the case of a land cover map these include the inaccuracies which result from modeling a complex pattern of spatial variation with a relatively small number of homogeneous areas separated by sharp discontinuities; in reality areas are not homogeneous and boundaries mark zones of transition rather than sharp breaks. Although it may be possible to model many forms of processing error (see for example Amrhein and Griffith 1987; Keefer, Smith and Gregoire 1988), it is virtually impossible to describe source errors without access to additional information such as ground surveys.

ERROR IN CLASSIFIED PIXELS

Many methods of image classification estimate the probability that a pixel belongs to each of a set of possible classes: commonly, the class to which the pixel is

finally assigned is that having the largest probability. However the complete set of probabilities for each pixel constitutes a useful source of information on the uncertainty of classification. Let the subscript i denote one of the n pixels, j denote one of the m classes, and let the vector $\{p_{i1},p_{i2},\ldots,p_{im}\}$ denote the set of probabilities for pixel i. Let M_i, $M_i = k \mid p_{ik}>p_{ij} \; \forall \; j\neq k$, be the most likely class.

A maximum likelihood classification based on M_i allows easy restructuring of the pixels to objects using a raster/vector conversion algorithm, but it implicitly deletes all potentially useful information on uncertainty, thus creating the kind of situation we have already described as common for object-based models of such themes as land cover. We propose instead that the entire vector be passed to the GIS, allowing GIS analysis to incorporate uncertainty into its processes and products. In most applications it is likely that only a small proportion of the m probabilities for each pixel will be significantly large, so we need not necessarily assume that this strategy will result in an m-fold increase in the storage requirements of this particular layer.

In order to obtain objects from the vectors of probabilities we must first create a realization, or a specific outcome of the stochastic process which the probabilities define. Let X_i denote the class to which pixel i is assigned in a particular realization: the maximum likelihood classification generates an outcome of the stochastic process by simply assigning $X_i = M_i$. The same set of probabilities can be used to produce multiple realizations or outcomes, corresponding to the tossing of a dice, and the differences between outcomes represent uncertainty.

The simplest realization would be a multinomial process in which the outcome in each pixel is determined independently, based on the known probabilities. A simple approach would be to generate a random number x_i, $0\leq x_i\leq 1$, and assign class k if:

$$\sum_{j=1}^{k-1} p_{ij} < x_i \leq \sum_{j=1}^{k} p_{ij} \qquad (1)$$

However the result would appear unreasonably fragmented because of the independence of the outcome in neighboring pixels, and it is very unlikely that large, homogeneous patches of similarly classified pixels would develop except where one probability is close to 1 and classification is therefore almost certain. This process would fail therefore to model the common situation in remote sensing where a large patch of many pixels returns a homogeneous spectral response, but nevertheless has a very uncertain classification. A further objection is that by ensuring homogeneity within pixels but independence between them, we create a result which is very dependent on pixel size.

These objections can be removed if the outcomes in neighboring pixels are allowed to be correlated. In essence, we require a process of realization in which two

properties are satisfied: a) the proportion of realizations in which pixel i is assigned class j tends to p_{ij} as the number of realizations tends to infinity (posterior and prior probabilities are equal), and b) outcomes within one realization display a prescribed level of spatial dependence.

METHODS OF REALIZATION

Goodchild and Wang (1988) described a process in which each pixel was first independently assigned to a class. This initial image was then repeatedly convolved with a low-pass filter, in order to impose spatial dependence (see also the ICM technique of Besag 1986). The paper illustrated the use of a 3 by 3 filter with the rule that in each pass the central pixel was assigned the modal class of the 9 pixels within the filter window. This process was demonstrated to generate spatially dependent realizations, allowing uncertainty in pixel classifications to be converted to uncertainty in the location of objects and to concepts such as the epsilon band. However it is easy to show that the low-pass filter generates posterior probabilities which are not equal to the prior probabilities, violating the first requirement above, except in special cases.

Cross and Jain (1983) have described a process of modeling spatially dependent images in which an initial set of outcomes, such as that produced by our simple multinomial process above, is modified by selectively swapping the contents of randomly selected pairs of pixels (see also Goodchild 1980). Again, while the result is a pattern which has strong spatial dependence, in general the prior and posterior probabilities in each pixel are not equal.

Two approaches appear to offer a way of satisfying both requirements simultaneously, one computationally intensive and the other storage intensive. The latter is conceptually simpler and will be described first. Let q denote a number of realizations, say 100, and suppose that initial classes are assigned to each pixel in each of q realizations by the multinomial, spatially independent process. The proportion of realizations in which a given pixel is assigned to a given class will be approximately equal to the prior probability. Now suppose that some means exists to measure the level of spatial dependence present in any one realization, and that a target value for this measure has been defined. Suitable measures can be found in the literature on indices of spatial autocorrelation (Cliff and Ord 1981; Goodchild 1988). The technique then executes the following steps until the target is reached, or no further improvement can be obtained:

```
select a pixel at random;
for that pixel, select a pair of realizations at random;
if the pixels are currently assigned to different
classes, then;
     swap the contents of the pixels if by so doing the
     recomputed measure is closer to the target;
end if;
```

repeat.

Because the technique cannot change the numbers of each class assigned to any one pixel across the set of realizations, we ensure that the posterior and prior probabilities are equal.

The second method implements a spatially autoregressive process in which the value in any cell is correlated with the value in nearby cells (Haining, Griffith and Bennett 1983). A spatially autoregressive process on a lattice can be defined as:

$$\mathbf{z} = \rho \mathbf{W} \mathbf{z} + \boldsymbol{\epsilon} \quad (2)$$

where: z_i is the value assigned to pixel i by the process;
ρ is a spatial autocorrelation parameter;
$\mathbf{W}$ is an n by n array of interactions between pixels;
ϵ_i is an independent, normally distributed error term with zero mean and variance σ^2.

We assume that W_{ij} is 1 if pixels i and j are 4-adjacent, else 0. The solution for $\mathbf{z}$ is given by:

$$\mathbf{z} = (\mathbf{I} - \rho \mathbf{W})^{-1} \boldsymbol{\epsilon} \quad (3)$$

The $\mathbf{z}$ are known to be multivariate normal with zero mean and with variance-covariance matrix given by (Haining, Griffith and Bennett 1983):

$$\sigma^2[(\mathbf{I} - \rho \mathbf{W})^T (\mathbf{I} - \rho \mathbf{W})]^{-1} \quad (4)$$

We can obtain a class X_i for pixel i from the following rule:

$$X_i = k \quad \text{iff} \quad F(z_i) < p_i \quad (5)$$

where F(z) is the probability that an independent, normally distributed random number with mean 0 and variance given by the diagonal terms of (4) exceeds z.

Unfortunately the technique requires the inversion of an n by n matrix, and special methods are necessary to generate realizations in arrays of more than about 8 by 8 pixels.

Both techniques have the advantage that it is easy to include prior information about such objects as field boundaries, roads or water. The spatial dependence between pixels across a significant boundary can be deleted by setting the appropriate terms in $\mathbf{W}$ to zero instead of 1, which will cause the boundary to emerge in each realization. In the swap technique the same effects can be achieved by setting appropriate terms in the evaluation function, which will in most cases include the equivalent of the $\mathbf{W}$ matrix. Similarly the presence of known classes such as water can be dealt with by setting the associated probability to 1 and all others to 0 in affected pixels.

DISCUSSION

The techniques described can be used to simulate the effects of uncertainty in both field and object views of spatially varying phenomena. Goodchild and Wang (1988) illustrate the generation of a cross-classification matrix, which is the approach often used in image processing to assess accuracy, and the sliver polygons and epsilon bands of the object-based approach to accuracy. Although these methods emphasize the equivalence between the measures used in both views, we stress that it is the field-based probabilities which are externally generated, while the object-based measures must be derived from them. This serves to emphasize the earlier point that an object-based view of spatial data rarely carries information on which an objective model of accuracy can be based.

The techniques provide a framework within which it is possible to discuss a number of conceptual models of uncertainty in spatial data. We have argued that the independent pixel is almost always inappropriate: because of spatial dependence these techniques produce patches whose size and shape are controlled by user-defined parameters and largely independent of pixel size. By setting appropriate levels of spatial dependence it is possible to produce a range of outcomes from highly fragmented and scattered patches when spatial dependence is low and local, through large patches which result from the aggregation of numbers of spatially dependent choices. With high levels of spatial dependence and with appropriately set terms in the **W** matrix, it is possible to have predefined patches in which the outcome is essentially the result of a single trial, thus simulating the example of the multi-pixel field whose entire class is uncertain.

We have thus far assumed that spatial dependence is a stationary property of the entire array. In reality some classes display patches which are more fragmented than others, and spatial dependence also varies from one region to another. In the future we hope to develop methods which will successfully simulate these conditions as well.

REFERENCES

Amrhein, C. and D.A. Griffith, 1987, GIS, Spatial Statistics and Statistical Quality Control: Proceedings, IGIS '87, ASPRS/ACSM, Falls Church, VA.

Besag, J., 1986, On the Statistical Analysis of Dirty Pictures: Journal of the Royal Statistical Society B48:259-302.

Blakemore, M., 1984, Generalization and Error in Spatial Databases: Cartographica 21:131-9.

Burrough, P.A., 1986, Principles of Geographic Information Systems for Land Resources Assessment, Oxford.

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

Cross, C.R. and A.K. Jain, 1983, Markov Random Field Texture Models: IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-5:25-39.

Franklin, W.R., 1984, Cartographic Errors Symptomatic of Underlying Algebra Problems: Proceedings, International Symposium on Spatial Data Handling, Zurich, University of Zurich, 190-208.

Goodchild, M.F., 1979, Effects of Generalization in Geographical Data Encoding: in H. Freeman and G.G. Pieroni, editors, Map Data Processing, Academic Press, New York, 191-206.

Goodchild, M.F., 1980, Simulation of Autocorrelation for Aggregate Data: Environment and Planning A 12:1073-81.

Goodchild, M.F. and M.-H. Wang, 1988, Modeling Error in Raster-Based Spatial Data: Proceedings, Third International Symposium on Spatial Data Handling, Sydney, IGU Commission on Geographical Data Sensing and Processing, Columbus, Ohio.

Goodchild, M.F., 1988, Spatial Autocorrelation, Concepts and Techniques in Modern Geography No. 48, GeoBooks, Norwich.

Haining, R.P., D.A. Griffith and R.J. Bennett, 1983, Simulating Two-Dimensional Autocorrelated Surfaces: Geographical Analysis 15:247-53.

Keefer, B.J., J.L. Smith and T.G. Gregoire, 1988, Simulating Manual Digitizing Error with Statistical Models: Proceedings, GIS/LIS '88, ASPRS/ACSM, Falls Church, VA, 475-83.

Prekal, J., 1956, On Epsilon Length: Bulletin de l'Academie Polonaise des Sciences 4:399-403.

Walsh, S.J., D.R. Lightfoot and D.R. Butler, 1987, Recognition and Assessment of Error in Geographic Information Systems: Photogrammetric Engineering and Remote Sensing 53:1423-30.