MODELING ERROR IN RASTER-BASED SPATIAL DATA

Dr. Michael F. Goodchild
Visiting Professor
Department of Geography
University of California
Santa Barbara, California 93106
U. S. A.

Wang Min-hua
University of Western Ontario

Introduction

The problem of error in GIS products has attracted considerable attention in recent years (see for example Chrisman, 1987; Goodchild and Dubuc, 1987; Burrough, 1986; Walsh, Lightfoot and Butler, 1987). Spatial data handling systems process data with high precision, so it is perhaps surprising to find that the results of query or analysis frequently conflict with ground truth. Unfortunately the input to such systems is almost always an abstraction of reality, and processing frequently increases the degree of abstraction. For example the data input from a forest inventory often consist of classified stands of timber, defined by polygonal boundaries and homogeneous attributes. Since the forest stand itself is rarely if ever homogeneous, even on the simplest attributes, such input constitutes an abstraction of reality; furthermore the degree of abstraction is usually unknown. To compute an estimate of marketable timber, the attributes and area of the stand will be input to a set of yield tables, but the results will be subject to additional uncertainties because yields are only crudely predictable.

Research on error in GIS has two objectives, both of substantial practical significance: first, to minimize error in products, and second, to develop models of error which can be used to compute measures of uncertainty. At present we know of no system which routinely estimates and reports measures of the error inherent in any of its products.

Some GIS operations contribute directly to error, and such cases are relatively easy to characterize and model. Errors introduced during digitizing and processing are termed processing errors. Consider a homogeneous patch of land, represented as a polygon in a vector database. Conversion to raster representation introduces a distortion, since the patch must now be represented by a set of pixels of fixed size. If the area of the patch is estimated from the raster representation, it is relatively straightforward to model the effects of this distortion on the estimate of area, in the form of a measure of standard error (Goodchild, 1980). But while such measures can be used as useful guides in selecting pixel sizes, they are based on the assumption that the homogeneous patch with precise boundary is a correct

representation of reality. They fail, therefore, to deal with lack of homogeneity, or uncertainty in the location of the boundary, despite the obvious dominance of these sources of error in many resource management applications, as demonstrated in several experiments (McAlpine and Cook, 1971; MacDougall, 1975). Differences between reality and the representation input to the system are termed source errors.

Modeling source errors due to fuzzy boundaries or heterogeneous patches is much more difficult. In many GIS applications the database was derived from a cartographic representation with precise boundaries and homogeneous patches; information on fuzziness and heterogeneity was lost long before the data entered the system. It is clear that our ability to model error in a comprehensive fashion will depend on new approaches to data collection and representation which avoid some of the abstraction present in traditional methods.

Consider once more the example of a forest database. A digital forest inventory is created by first obtaining aerial photography or remotely sensed imagery. Manual interpretation then converts this to a cartographic representation of homogeneous stands; attributes are added by interpretation and direct ground observation. The cartographic representation is then digitized as a vector database. An alternative approach would input the raw imagery and ground observation directly to the digital system, as pixels and point or line features respectively. The system would now contain the necessary information both to compute measures of heterogeneity and boundary fuzziness, and hence uncertainty in products, and also to derive the cartographic view of homogeneous stands as and when required.

The interfacing of remote sensing systems and GIS presents similar issues. Once an image has been interpreted and classified in a remote sensing system it is common to generate information products by transferring the data to a GIS, which allows such operations as topological overlay, calculation of area of homogeneous patches, etc. Processing errors due to the use of pixels of fixed size to represent homogeneous spatial objects can be estimated relatively easily. However variation of class within pixels, heterogeneity of classes and problems of class definition, and uncertainties due to the physics of the sensor often make much larger contributions to error. Information on source errors is not available to the GIS from classified pixels.

The objective of this paper is to develop methods of characterizing source errors in a GIS. Several techniques of classifying remotely sensed images are capable of yielding probabilities of class membership for each pixel; we explore the notion that these can be passed to a GIS, rather than a single class for each pixel, and used to derive appropriate measures of product error. In effect, we propose to approach the problem of source error estimation by reducing the level of abstraction in GIS databases.

The next section of the paper discusses the nature of class membership probabilities, and associated stochastic processes. The following section explores the relationship between pixel probabilities and certain well-known concepts in cartographic error modeling. The final section discusses the implications of this approach for GIS error estimation.

Class Membership Probabilities

Consider the classification problem as that of assigning each pixel in an image to one of m classes. Several methods of classification, such as discriminant analysis, are capable of yielding membership probability vectors for each pixel, $\{p_{j1}, p_{j2}, \ldots, p_{jm}\}$ where p_{ji} denotes the probability that pixel j is a member of class i given its spectral response. Clearly $\sum_i p_{ji} = 1$ for all j.

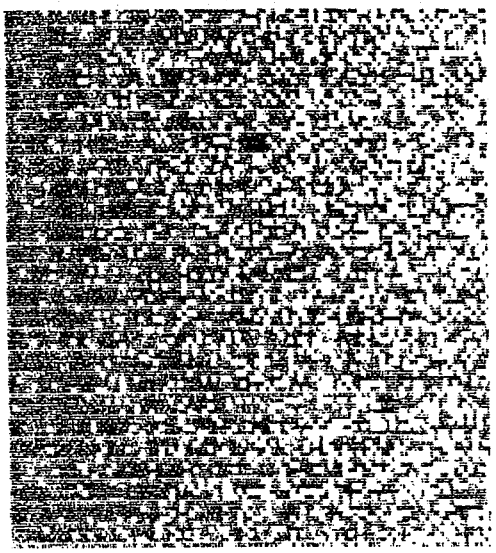
The probability vectors can be viewed as a multivariate surface, with the value in any pixel equal to the mean value of the surface over the domain of the pixel. Depending on the pixel size, we expect to find positive spatial autocorrelation in each surface, as the value of p_{ji} in any pixel is expected to be more like neighboring values than distant values: $E(p_{ji}-p_{ki})^2$ is an increasing function of the spatial separation of j and k.

Probability vectors can provide very simple estimates of the uncertainty inherent in responses to certain types of query. A query concerning the attributes of a point could be satisfied by returning the probability vector for the pixel containing the point. A more reliable response might be based on interpolation of continuous surfaces, subject to appropriate constraints (Tobler, 1979); the volume under each surface within each pixel would have to equal the observed probability for that class, and the surfaces at each point would have to sum to one.

The expected value of area of class i within some defined region S is given by $a \sum_{j \in S} p_{ji}$ where a is the area of a pixel. However the variation which is expected about this mean, in other words the uncertainty of an area estimate, depends on our conceptualization of the associated stochastic process. We use the term **realization** to denote a process by which pixel probability vectors are converted into a specific class at every point.

One simple realization is an independent Bernouilli trial in each pixel under the given probabilities. The result would be unrealistic because the variation between neighboring pixels would be inconsistent with the assumed homogeneity within cells, and therefore conflict with our expectations of the way variance should increase with distance. Figure 1 shows this realization applied to a 100 by 100 array with m=2; the probability that a cell is class black varies linearly from 1 on the left to 1/2 on

the right. The variance of an area estimate for class i in region S under this realization is given by $a^2 \sum_{j \in S} p_{ji}(1-p_{ji})$.



1. Realization by independent trials in each pixel, probabilities declining linearly from 1.0 on the left to 0.5 on the right.

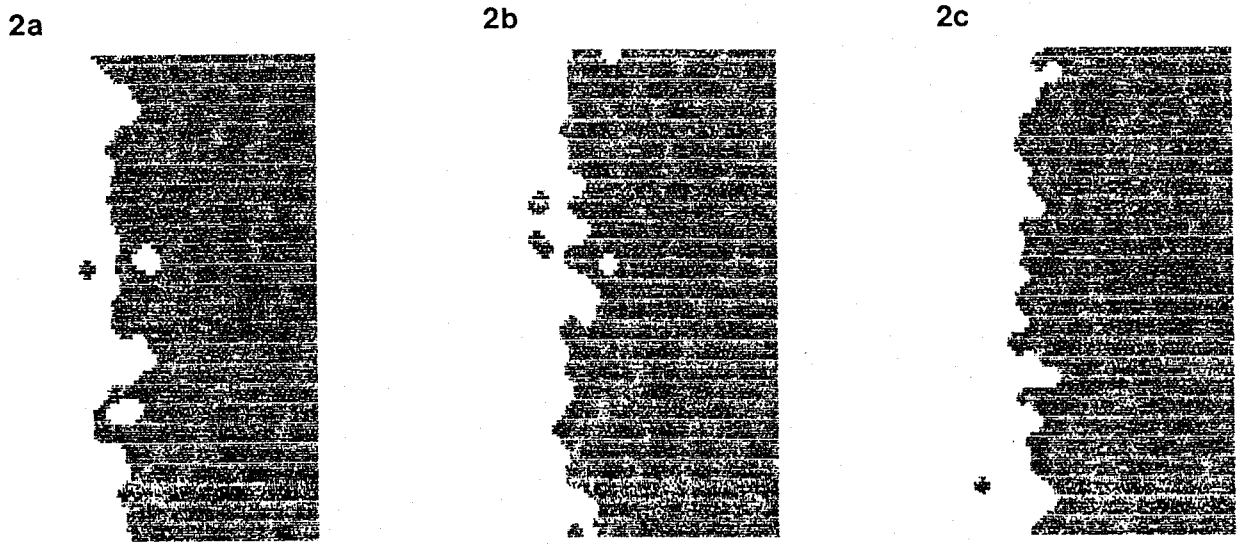
Another simple approach would be to assume the process to be spatially divisible, with independent trials within each subpixel. This would imply an assumption of homogeneity within subpixels and independence between, and would require that the size of subpixel be chosen appropriately. If each pixel is divided into n subpixels the variance in an area estimate is given by $a^2/n \sum_{j \in S} p_{ji}(1-p_{ji})$, which declines with n to reach 0 when the process is infinitely divisible.

Finally, consider the "realization" implied by maximum likelihood classification, in which each pixel is assigned the class of its greatest probability. The expected area is now given by $\sum_{j \in S} a\,\delta_{ji}$ where δ_{ji} is 1 if $p_{ji} > p_{jk}$ for all $i \neq k$, else 0. This estimate is clearly biassed with respect to the estimates from the previous realizations. In the case of Figure 1 it gives an estimate of 10000 (the entire array) rather than 7500, while the actual count of black pixels in Figure 1 is 7518.

Intuitive expectations about the relationship between variance and distance (the **variogram** function) are determined largely by context. A forester defining stand boundaries may apply subjective rules about the minimum size of a stand and the minimum curvature of stand boundaries. Suppose that the size of the pixel is defined by the context such that homogeneity within pixels is acceptable. It is still unacceptable to generate a

realization with independent trials because of the consequent excessive variance between neighbors.

A simple way to reduce interpixel variance is to apply a smoothing filter to the result of independent trials. Figure 2 presents three different realizations of the same set of probabilities; in each case these vary from 0 on the left to 1 on the right according to a logistic function. Maximum likelihood classification would produce a vertical boundary. In each case the initial map has been convolved with a 3 by 3 pixel filter by replacing the class of the central cell by the modal class until no further change occurs. Within any one realization the outcome for a particular pixel is no longer independent of neighboring outcomes. Across trials the posterior probability that a particular pixel has class i is determined by the prior probability and the nature of the filter. Clearly it is desirable that filters be chosen such that prior and posterior probabilities are equal for all pixels and classes, although for this particular filter the posterior probability is always closer to 0.5.



2. Three realizations of the same set of probabilities after applying a low-pass filter to independent trials.

The realization shown in Figure 2 provides a means of generating different representations of objects under the same set of input probabilities. Thus we now have a model for the distortions inherent in an object-oriented spatial database. Furthermore the model is based entirely on information readily available to a GIS provided a suitable method of image classification has been used, and provided some real basis can be given to the filter. In practice we suggest that the criteria for filter selection should be based on achieving equality in prior and posterior probabilities, and a desirable functional relationship between interpixel variance and distance. In the latter case there are clearly analogies to the choice of variogram function in Kriging.

Cartographic Distortion

In this section we consider the relationship between this model of pixel probabilities and more widely known concepts in cartographic distortion. The edge of a cartographic object is represented in a vector database as a line, usually by connecting points with straight line segments, despite the presence of both source and processing errors. Distortion in lines has been described through the concept of an epsilon band (Perkal, 1956; Chrisman, 1982; Blakemore, 1984); in its simplest version the true position of the line is believed to lie within a band of width epsilon about the observed line with probability 1.

The three realizations in Figure 2 can be described by a probabilistic version of the epsilon band. The true line will be taken to be the locus of p=0.5, which is a vertical line midway along the x axis. An estimate of an epsilon-like parameter can be obtained from the mean displacement of the observed line parallel to the x axis in each realization. Over 100 realizations, the mean displacement was 3.4 pixels with a standard deviation among realizations of 0.5. Mean displacement was calculated by taking all white cells to the right of the true line, and all black cells to the left, and averaging the horizontal distance to the true line. The islands shown in Figure 2, which occur frequently in this process, were therefore included in the calculation of the mean displacement.

Blakemore (1984) has described a modification of the point in polygon problem to allow for uncertainty in the position of the polygon boundary. The process described above provides a direct link between pixel probabilities and the containment of a point within a homogeneous patch inferred from a classification of pixels. However the results will certainly be different. Based on pixel probabilities a point will never be identified as certainly contained within a homogeneous patch; to do so requires a realization of prior probabilities based on some assumed level of spatial autocorrelation.

Part of the error in estimating the area of a patch derives from uncertainty about the position of the patch boundary. Consider the black areas of Figure 2 as patches whose area must be estimated. Under a maximum likelihood "realization" the estimate has zero standard error; under an independent series of Bernouilli trials the standard error can be obtained from the prior probabilities as noted above. In the realization process of Figure 2 the standard error is substantially reduced, since at several pixels distance from the boundary there is no longer uncertainty as to a pixel's class. Over 100 realizations the standard error was 164 about a true area of 5000 pixels.

Spurious or sliver polygons result when two independent distortions of the same line or polygon boundary are overlaid, and in large databases the number of spurious polygons so created can easily overwhelm the system (Goodchild, 1978). It is common in vector systems to include code to remove slivers after overlay, using rules based on area and perhaps shape. However

such algorithms have no sound basis in statistical models of error.

Figure 3 shows two independent realizations of a multinomial array, derived from a section of a Landsat scene. The scene was classified using cluster and discriminant analyses, to yield vectors of membership in each of four classes. These were then realized with independent trials followed by passes of a simple 3 by 3 filter. Note that the peripheral pixels reflect the difficulty of using the modal filter on the edge of the array.

Several authors (see for example Greenland and Socher, 1985) have compared two such versions of the same image using a bivariate table in which pixels are crosstabulated according to their classes on the two versions. The degree of agreement can be summarized by a chisquared or kappa statistic as a measure of distortion. The table below shows the mean counts for 100 comparisons of pairs of independent realizations of the same set of probabilities used to generate Figure 3:

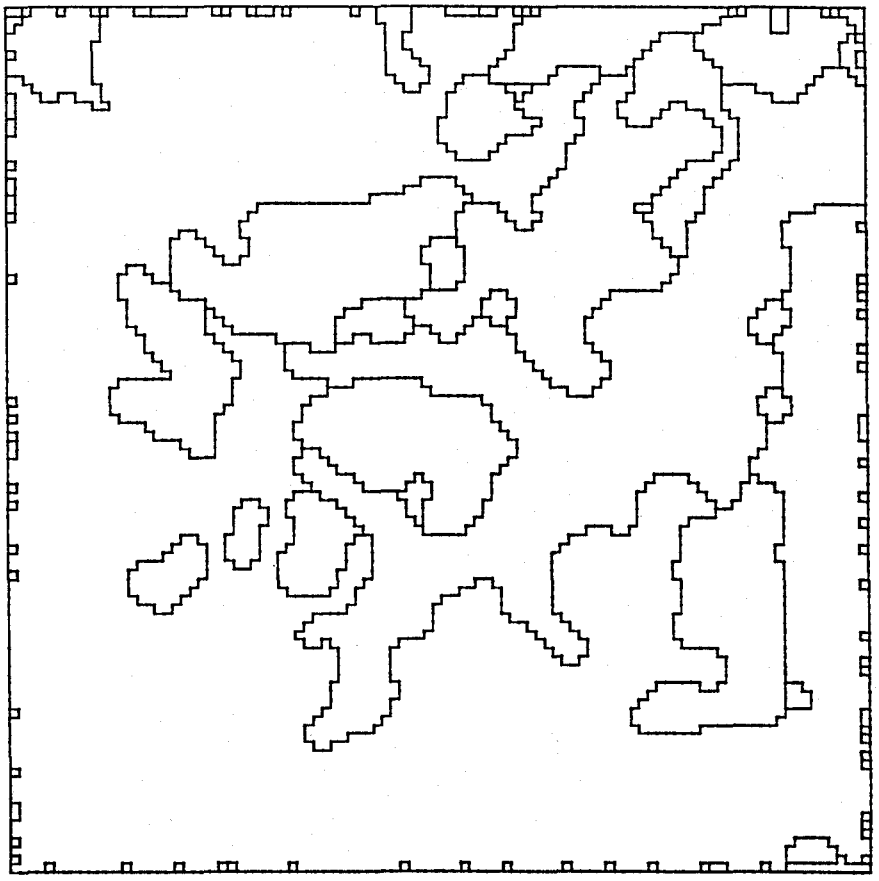
	Image A				
Image B	1	2	3	4	total
1	5427	176	49	120	5772
2	179	1635	3	36	1853
3	45	3	441	6	495
4	120	34	6	1720	1880
total	5771	1848	499	1882	10000

As expected most cells have the same class in both images of each pair, and thus cluster along the diagonal of the table.

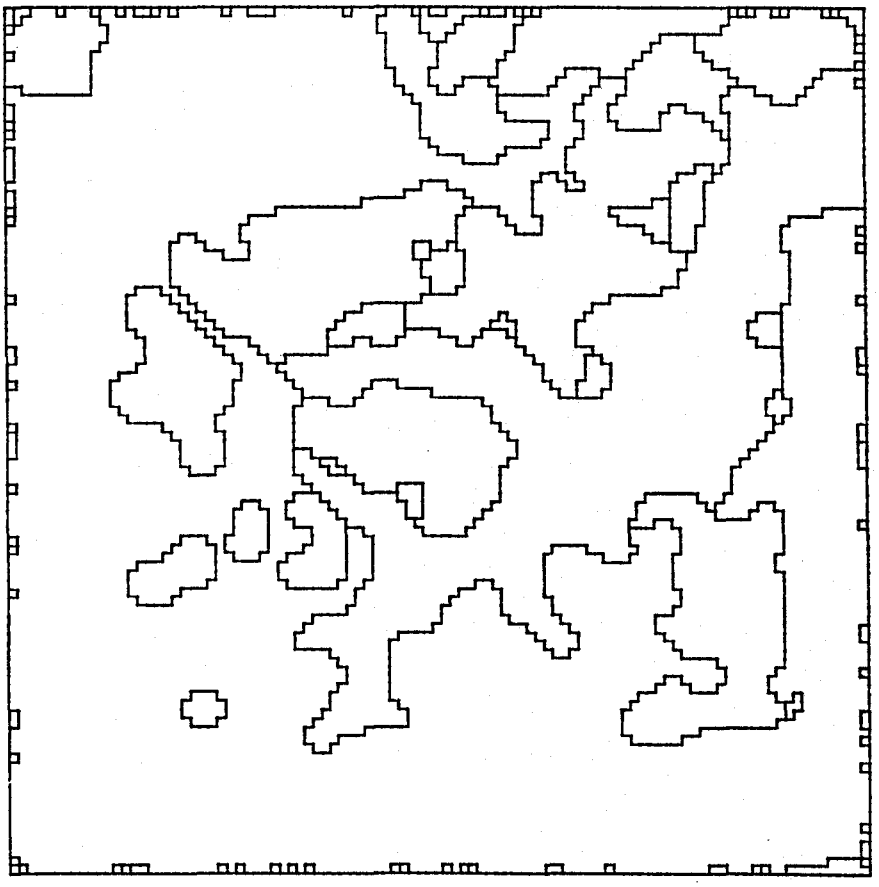
Both chisquared and kappa compare the cells of the table to the counts which would be expected if pixel classes were assigned randomly; for example any pixel in image A would have the same chance of being one of the 5772 class 1 pixels. However this is an entirely inappropriate null hypothesis which bears no relationship to any possible process of realization of pixel probabilities.

In the context of our model, disagreement between the two images results from the independent realization of a constant set of prior probabilities. Maximum likelihood and infinitely divisible realizations will produce no disagreement. High disagreement will occur if independent trials are used in each pixel, but this will be reduced when positive spatial autocorrelation is induced by passes of a low-pass filter. The values of chisquared and kappa are therefore artifacts of the prior probabilities and processes of realization, and not directly interpretable.

3a



3b



3. Two realizations derived from probabilities obtained by applying a discriminant classifier to a portion of a Landsat scene.

Implications

We have argued that errors in spatial databases, which lead to artifactual features such as sliver polygons and to uncertainties in GIS products, can be divided into processing errors introduced during digitizing and manipulation, and source errors in the form of discrepancies between input data and ground truth. Processing errors are likely to be more easily modeled, but unfortunately are likely less significant in many applications. The difficulty of modeling source errors is attributable largely to the fact that relevant information is often lost in the process of creating an input document, in the form of a cartographic view of reality.

We have suggested in this paper that class membership probabilities, which are obtained as a byproduct of several methods of classification for remotely sensed imagery, contain useful information on source error. If these are passed into a GIS, instead of a single class for each pixel, then they can be used to develop appropriate representations of uncertainty for GIS products. However, to do so requires a conceptualization of the manner in which pixel probabilities are realized in a stochastic process. Conducting trials in each pixel assumes that spatial variation is zero within pixels, and assigns independent classes to neighbors. Realization has been modeled in this paper by independent trials in each pixel followed by convolution with a low-pass filter.

Prior probabilities provide a satisfactory way of establishing confidence limits on GIS products in a raster database. In an object-oriented database the probabilities must be realized, requiring knowledge both of the prior probabilities and of a suitable filter. With this information it is possible to estimate the errors associated with areas of patches. However it is important to note that such errors are not homogeneous attributes of polygons. If polygon A on one image is overlaid with polygon B on another, the error in the area of intersection cannot be computed through a simple operation on the attributes of A and B. Our model does not support the notion of tracking error through the operations of an object-oriented database.

Unfortunately our model offers little support for estimation of source errors in conventional spatial databases containing classified pixels or homogeneous, precisely bounded objects. It seems that effective treatment of source errors will require the kinds of restructuring of data collection, interpretation and input outlined in the introduction.

References

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. Monographs on Soil and Resources Survey No. 12. Oxford University Press.

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

Chrisman, N.R., 1987. "The accuracy of map overlays: a reassessment". Landscape and Urban Planning 14 427-39.

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. "A fractal approach to the accuracy of geographical measures". Mathematical Geology 12 85-98.

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

Greenland, A. and R.M. Socher, 1985. "Statistical evaluation of accuracy for digital cartographic databases". Proceedings, AutoCarto 7. ASPRS/ACSM, Falls Church, Virginia, 212-21.

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

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

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

Tobler, W.R., 1979. "Smooth pycnophylactic interpolation for geographical regions". Journal of the American Statistical Association 74 519-30.

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 1423-1430.