(168)

INT. J. GEOGRAPHICAL INFORMATION SYSTEMS, 1992, VOL. 6, NO. 2, 87–104

Development and test of an error model for categorical data

MICHAEL F. GOODCHILD

National Center for Geographic Information and Analysis,
University of California, Santa Barbara, CA 93106, U.S.A.

SUN GUOQING

Ressler Associates Inc., 14 440 Cherry Lane,
Court Suite 212, Laurel, MD 20707, U.S.A.

and YANG SHIREN

National Center for Geographic Information and Analysis,
University of California, Santa Barbara, CA 93106, U.S.A.

Abstract. An error model for spatial databases is defined here as a stochastic process capable of generating a population of distorted versions of the same pattern of geographical variation. The differences between members of the population represent the uncertainties present in raw or interpreted data, or introduced during processing. Defined in this way, an error model can provide estimates of the uncertainty associated with the products of processing in geographical information systems. A new error model is defined in this paper for categorical data. Its application to soil and land cover maps is discussed in two examples: the measurement of area and the measurement of overlay. Specific details of implementation and use are reviewed. The model provides a powerful basis for visualizing error in area class maps, and for measuring the effects of its propagation through processes of geographical information systems.

1. Introduction

The error problem in geographical information systems can be briefly summarized as follows. The contents of a geographical information system's database are almost without exception approximations to real geographical variation (exceptions include the geographical location of the North Pole and other mathematically defined objects, as their locations can be expressed with unlimited accuracy). Products of a geographical information system are thus similarly approximations. To estimate the uncertainty associated with a given product (e.g. its confidence limits), it is necessary to (*a*) model the uncertainty existing in the database and (*b*) model the propagation of uncertainty through the operations performed on the data by the system. Neither of these requirements is trivial.

For the purposes of this paper, an error model is defined as a stochastic process capable of generating a population of distorted versions of the same reality. Each version is a sample from the same population. The best known and perhaps the simplest such error model describes the variation in samples of a single, scalar measurement, such as air temperature, based on the Gaussian distribution. The mean of the distribution is an estimate of the true value and the standard deviation is a measure of uncertainty. In a spatial database context, the variation from sample to sample might represent the differences between the classifications of the same raw land cover data by different workers, or between different digitizations of the same map. Simple models of error, such as the Gaussian model, attempt to provide a general framework for

0269–3798/92 $3·00 © 1992 Taylor & Francis Ltd.

describing and simulating error, rather than an accurate model of the actual process by which error accumulates in a given instance. The theory behind the Gaussian model merely asserts that the error process combines a large number of random, independent and additive distortions of unknown origin. If more information is available, such as detailed calibrations of particular error sources, then the Gaussian model may be replaced by a more specific model. Similarly, our model is presented as a generic framework appropriate to the example where nothing is known about the actual processes contributing to the error. Where these are known and are suitable for modelling, as might be true when positional errors are introduced in line objects during digitizing, then it may be appropriate to design and calibrate a more informed error model. Thus we do not claim that our model is an accurate representation of the actual processes by which error is introduced into spatial data.

Many commonly used descriptions of map error fail to meet the requirements of an error model, as they fall short of complete specification of a stochastic process. Such descriptions include the width of an epsilon band, the circular map accuracy standard used as the basis for many statements of positional accuracy, the statistics of the misclassification matrix used in remote sensing and the reliability diagram found on many topographic maps. None of these provides sufficient detail to support simulation of the distortions present in complex spatial objects.

One source of difficulty in dealing with error is that the objects that populate a geographical information system database are frequently artefacts of the process of modelling continuous geographical variation (fields), and are thus subject to uncertainty not only in their locations and attributes, but also in their number and topological properties. For example, suppose two observers map land cover for the same area, using the same raw data, and record their results as classified polygons. The two sets of results would probably differ not only in the classes given to polygons and the precise locations of polygon boundaries, but also in the numbers of polygons, edges and nodes in the boundary networks. A previous paper (Goodchild 1989) discussed the difficulty of modelling error in such databases, and argued that error models based on fields are fundamentally easier to construct than models based on objects.

Consider, for example, the problem of modelling uncertainty in topographic elevation. Although elevation is geographically continuous, we suppose that the database contains a sample of digitized contours, that is, discrete objects. Uncertainty in this instance is represented by distortion in the positions of contour lines, and is often described by a statement such as 'the true location of the contour lies 90% of the time within a band of width x drawn around its recorded location'. A comparison of two versions of the same set of contours might be expected to yield large numbers of slivers, a commonly observed artefact of uncertainty in object positions. However, although error bands provide a description of uncertainty, they fall short of modelling it, as the published work on epsilon bands (e.g. Chrisman 1982) seems to provide no means of generating a population of distorted versions that satisfy all of the requirements of contours.

To model error in this situation, we need to step back from the contour representation to the data from which the contours were derived. Suppose these were a set of spot heights, used to interpolate a continuous field. Without going into details, it would be relatively straightforward to distort the elevation values at the spot heights, reinterpolate the field, and then recompute the contours. Of course, the new set of contours would be distorted relative to the first, but would satisfy all of the requirements of contours. Moreover, the amount of distortion of contour positions would not be constant, but would vary depending on local slope and the proximity of spot heights. It is perhaps worth noting that 'kriging' (Burrough 1986) can provide direct estimates of the uncertainty in contour positions that is due to the interpolation process itself.

The distinction between field and object errors models can be illustrated by considering a simple type of query: the determination of the value of a variable (e.g. soil type or elevation) at a specified point. In such examples the fact that objects are used to represent geographical variation is not immediately relevant, and accuracy in the positions of contours or polygon boundaries would be a very indirect way of expressing uncertainty in the value of a variable at a point.

As geographical information systems make use of both field and object representations, it is essential that any comprehensive approach to error modelling is meaningful in both domains. The purpose of this paper is to present such a model for categorical data, to illustrate its implementation and to present some results from the generation of a simple product of a geographical information system.

2. Categorical data

The focus of this paper is the so-called 'area class map' (Mark and Csillag 1989) or 'categorical coverage' (Chrisman 1982). This is a single variable taking a finite number of discrete, nominal categories, and whose value is known everywhere in the mapped region. Two data models are commonly used to present such variables. The polygon (often described as the vector) model identifies areas of homogeneous value, and is typified by the lines drawn on soil maps, or their digitized equivalent. The raster model identifies the class that is dominant in every cell of a regular (normally rectangular) array or tesselation, and is typified by a classified remotely sensed scene. Polygon and raster models are commonly used in spatial databases to represent such categorical data.

One source of uncertainty in polygon model databases derives from the digitizing process. The source map may have been distorted by folding, a change of humidity, or its placement on the table; registration points may have been located incorrectly; the cursor may not be positioned accurately; and there will be significant variation between individual operators in the selection of points to be digitized. Several error models of the digitizing process have been proposed (Keefer *et al.* 1988, Dunn *et al.* 1990, Bolstad *et al.* 1990).

However, the errors introduced by digitizing categorical data using the polygon model are generally small compared with the uncertainties present in the source document and passed intact into the spatial database. First, the boundaries digitized often represent generalizations of transition zones of various widths rather than real, sharp changes of category. The class assigned to a polygon is not, in fact, a homogeneous property of the entire polygon—many legends include such terms as 'mostly', 'predominantly', 'a mixture', 'substantial inclusions'. Holes and islands, i.e. inclusions of another class, are often deleted during the map-making process, particularly if they fall below some established threshold or 'minimum mapping unit'. Finally, the boundaries of polygons are often deliberately smoothed to create a more pleasing cartographic effect. In short, such maps, and particularly their digitized versions, often create a false sense of accuracy by being stripped of any indication of uncertainty.

Figure 1 shows a typical polygon model categorical dataset. Its polygons have homogeneous attributes, its boundaries are infinitely thin lines, it has an imposed

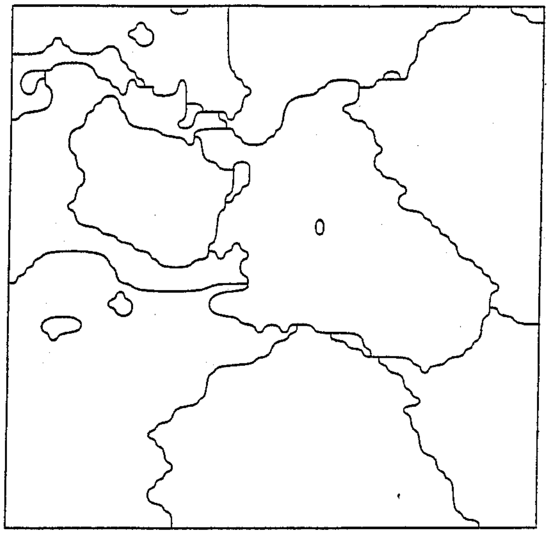


Figure 1. Typical categorical coverage using the polygon model.

threshold of minimum polygon area and its boundaries have been smoothed. In the context of this paper it represents a single sample from a population of distorted versions of the same truth, just as the Gaussian distribution is used to represent different observations of the same scalar measurement. Samples can be expected to vary not only in the positions of boundaries, but also in the number of polygons and the positions and number of nodes.

This is the object based ('cartographic') view of uncertainty that underlies the data quality sections of the proposed US Spatial Data Transfer Standard (DCDSTF 1988). In the contrasting field based view, polygons and their geometry and topology are themselves artefacts of the data modelling process and are thus not central to the issue of error. The real geographical variation is complex and continuous, and uncertainty is better described by devices such as the misclassification matrix (van Genderen and Lock 1977, Mead and Szajgin 1982), which provides empirical estimates of the probabilities of observing class i at a point when in reality class j is present, for all i and j. One obvious shortcoming of the matrix is its implicit assumption that these probabilities are constant over the mapped area.

3. The model: conceptual discussion

The mapped area is assumed to be discretized as an array of N cells, with rows and columns indexed by i and j, respectively. The raster model is chosen over the polygon model because of its greater suitability for capturing inherently continuous variation. The variable that is mapped over the area has n discrete classes, indexed by k. Finally, the probability that cell i, j falls into category k is given by p^k_{ij}; associated with each cell is a vector of probabilities $\{p^1_{ij}, p^2_{ij}, \ldots, p^n_{ij}\}$.

There are two common sources of this type of data, and together they cover many applications of geographical information systems. One is remote sensing, where it is assumed that the uncertainty associated with a classified scene is represented by the probability vectors. Many classifiers are capable of yielding this type of data, but in practice such probability vectors are commonly replaced by the most likely class, $s|p^s_{ij} > p^k_{ij}$ for all $k \neq s$. Probabilities might be interpreted as indicative of uncertain classification, of mixed pixels, of heterogeneous classes, or might be the product of a fuzzy classification. The other common source is interpretive mapping, as it is common for legends on land cover, soil, vegetation, or land use maps to include percentages, as in '90% class A with 10% inclusions of class B'. The well known Canada Geographic Information System (Tomlinson *et al.* 1976) provides three such classes and percentages for every polygon, although many analyses ignore all but the most common class. In this second example probabilities can be expressed as attributes of polygons, but are assumed to be expressed in raster form for the purposes of this error model.

The probability vector approach can also be used to characterize transition zones, as in the example now given. For example, the transition between a class A polygon and a class B polygon might be represented by cells with probability vectors varying gradually from {1·0, 0·0} through {0·5, 0·5} to {0·0, 1·0} (Mark and Csillag 1989). The degree of fuzziness of the boundary (an object concept) can thus be represented in the steepness of the probability gradient (a field concept). However, continuous variation is not limited to polygon boundaries. In some instances lower levels of certainty may occur in the interiors of polygons, rather than at the edges. For example, the class 'wetland' may be typified more by the sedge marshes around the edge of a wetland polygon than by the small area of open water in the middle.

In the error model described in this paper, the probability vectors associated with each pixel are assumed to be given by, and to summarize what is known about, detailed geographical variation within the mapped area. In contrast, the earlier model of Goodchild and Dubuc (1987) assumed that nothing was known about detailed geographical variation and instead generated samples of all possible maps. This earlier model operationalized the idea of a 'random map', but, unlike the new model, it cannot generate distortions of the same truth. In the new model, the absence of uncertainty is represented by having exactly one non-zero probability in every vector, i.e. by being certain about the class present in every cell. In remote sensing, the 'mixed pixel' effect ensures that uncertainty can never be completely absent from a scene.

Each sample map is a realization of the stochastic process defined by the probabilities, in which each cell is assigned to one and only one category. If one cell of the raster is compared across a large number of realizations, the proportion of realizations in which the cell was classified into a given category will approximate the given probability. One obvious way of obtaining a realization is to assign classes independently in each cell. Let c_{ij}^k represent the sum of class probabilities up to and including class k; c_{ij}^k is therefore the sum of p_{ij}^s from $s=1$ to k. Mechanically, we could achieve such a realization by generating a uniformly distributed random number in the range 0–1 in each cell, denoted by z_{ij}. Then cell i, j would be assigned to class k if $c_{ij}^{k-1} < z_{ij} < c_{ij}^k$.

Such an independent assignment of classes in each cell would fail to address the all important issue of texture. Consider a polygon with a homogeneous probability vector {0·8, 0·2}. At one extreme, the 20 per cent inclusions of class 2 might be fully mixed with the dominant class, so that any subarea within the polygon, however small, would contain exactly 20 per cent class 2. Every cell in a raster representation would be classified as class 1. At the opposite extreme, the probabilities might be interpreted as referring to the polygon as a whole. For example, the polygon might correspond to the boundaries of an agricultural field, and the uncertainty might relate to the type of crop growing in the field. In between lies a range of geographical scales of mixing between the two types, and corresponding size distributions of inclusions of class 2.

The texture can be controlled in an appropriate manner by inducing correlation between outcomes in neighbouring cells. When the correlation is zero, the pattern of inclusions is dominated by the cell size, but as the correlation increases, larger and

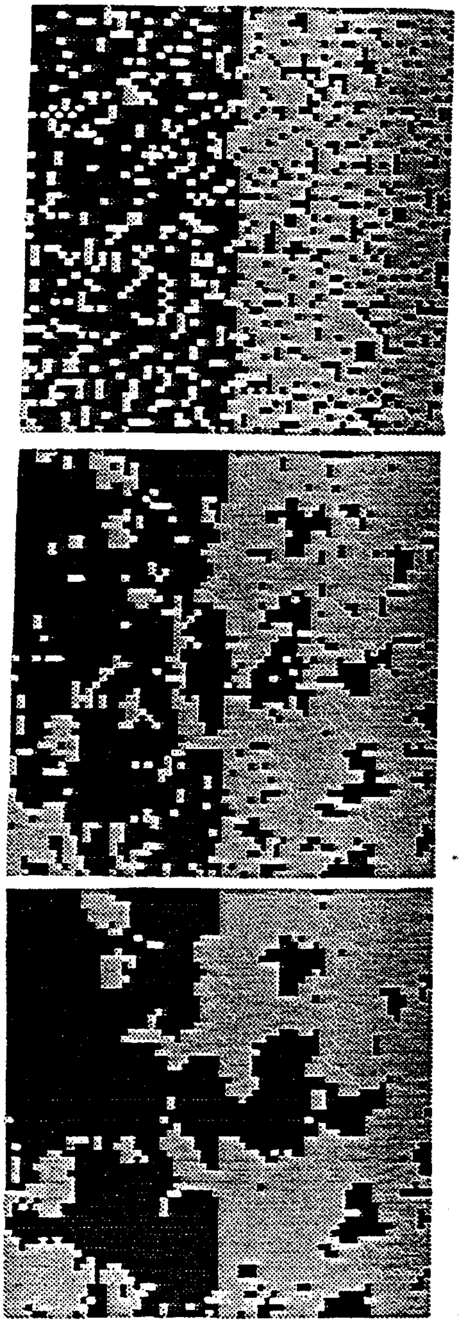


Figure 2. Simulation of inclusions within two polygons, with increasing levels of spatial dependence.

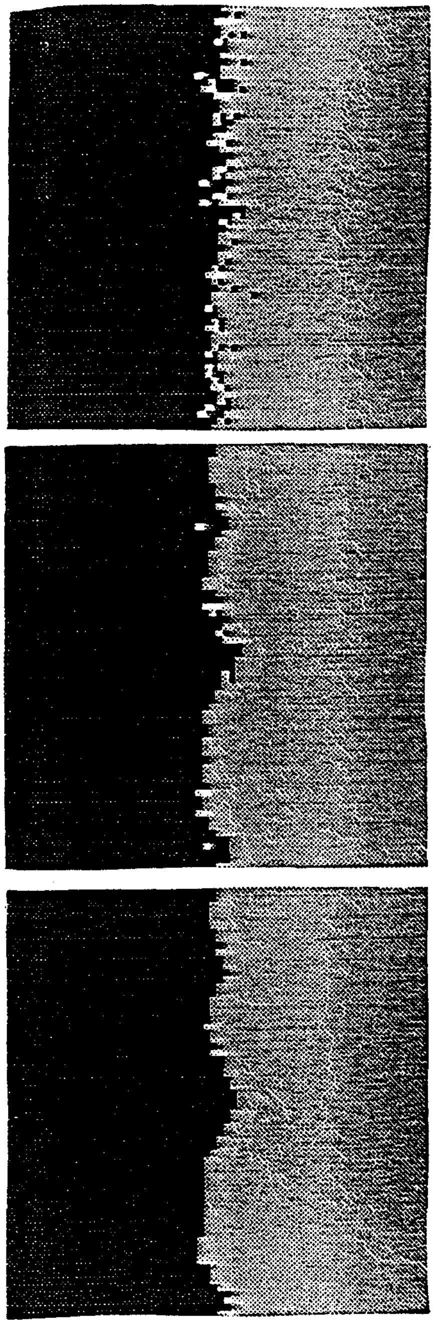


Figure 3. Simulation of a transition zone between two homogeneous areas, with increasing levels of spatial dependence.

larger aggregates begin to appear. The precise details are given in the following section. In summary, the requirements of the error model are (*a*) a vector of given probabilities in each cell determines the proportion of realizations in which the cell is assigned to each given class and (*b*) a correlation parameter determines the level of dependence between outcomes in neighbouring cells in any one realization. One realization of the process might correspond to the map drawn by one interpreter; the difference between realizations in this instance would represent the uncertainty introduced by variations between individual interpretations.

Figure 2 shows a simple illustration of the error model for two polygons, and demonstrates its ability to handle the problem of heterogeneity. Cells in the left polygon of each pair have a 75 per cent probability of being black, 25 per cent white, i.e. a probability vector of $\{0{\cdot}75, 0{\cdot}25\}$; probabilities are reversed in the right-hand polygon. The top panel is a realization with very low spatial dependence and shows the dominant influence of the cell size. The level of spatial dependence is increased in the other panels, leading to increasingly large inclusions. At the limit the entire left polygon would be either black or white in any one realization, with probabilities of 0·75 and 0·25, respectively. Note that the boundary between the two polygons remains clear in all illustrations because of its sharp change of probabilities.

Figure 3 shows the error model in the context of boundary transition, rather than polygon heterogeneity. The probability vectors in all three panels range from $\{1{\cdot}0, 0{\cdot}0\}$ on the left to $\{0{\cdot}0, 1{\cdot}0\}$ on the right, and follow a logistic (S-shaped) curve in the area of the boundary; the gradient of probability is steepest half-way across the diagram. Again, the upper panel shows the example where spatial dependence is almost absent. As dependence increases, the pattern is increasingly interpretable as the boundary between two polygon objects. As the position of the inferred boundary is different in every realization, the process provides a model of distortion in the position of a boundary between two polygons, and could be interpreted as a simulation of the uncertainty introduced by two digitizer operators. Superimposing a larger number of realizations with high spatial dependence would give an interpretation of the uncertainty band concept.

The two figures show a clear linkage between field and object concepts, first in the appearance of inclusion objects as a result of heterogeneity, and second in the modelling of positional uncertainty in the boundary between two polygon objects. In both instances the model provides a means of generating samples of distorted versions of the same basic truth (embedded in the probability vector fields) under different levels of spatial dependence.

The next section provides a precise statement of the model, and of the method used to generate these examples.

4. The model: technical discussion

Goodchild and Wang (1988) described an early version of the model in which spatially dependent realizations were computed by an averaging process. Each cell was first independently assigned a class, using the method described earlier. A 3 × 3 moving window was then passed repeatedly over the map, centring it on every cell in turn. At each position the central cell was assigned the modal class of the nine cells in the window. Unfortunately, it is easy to show that in this process the proportion of realizations P_{ij}^k in which cell i, j is given class k will not be equal to p_{ij}^k—in fact:

$$P_{ij}^k > p_{ij}^k \quad \text{if} \quad p_{ij}^k > 0{\cdot}5; \qquad P_{ij}^k < p_{ij}^k \quad \text{for} \quad p_{ij}^k < 0{\cdot}5. \tag{1}$$

Thus, although the process induces spatial dependence, it fails to satisfy the first requirement of the model, that the proportion of realizations in which a given cell is assigned to a given class is equal to the prescribed probability.

The two requirements can be satisfied if spatial dependence can be induced in the continuous variable z_{ij} (recall that z_{ij} is a uniform random deviate in the range 0–1). Thus, we seek a process for generating a random field in which (*a*) the marginal distribution of the variable in each cell is known across realizations and (*b*) spatial dependence is controlled in some straightforward fashion.

Consider the spatially autoregressive process:

$$X = \rho WX + \varepsilon, \qquad (2)$$

where X is a vector of length N (rows × columns), ρ is a parameter, W is a matrix of dimensions N by N, and ε is a vector of length N (our previous row and column indexing of cells has been replaced by a single index). Elements w_{uv} of matrix W are unit if cells u and v are Rook's case neighbours (share a common edge) and 0 otherwise. The elements of ε are independent Gaussian random deviates with mean 0 and a standard deviation of unity. Conceptually, equation (2) states that the value of the variable X in a cell can be predicted from a weighted combination of the values of the variable in neighbouring cells, plus a random value unique to the cell itself. When ρ is zero, the variable has no spatial dependence. The maximum value of ρ is 0·25 to ensure stationarity (Cliff and Ord 1981).

Equation (2) provides a system of simultaneous linear equations that can be solved to find X. In principle, this requires the inversion of an N by N matrix (rows × columns by rows × columns, e.g. at 64 × 64 map requires inversion of a 4096 × 4096 matrix):

$$X = (I - \rho W)^{-1}\varepsilon, \qquad (3)$$

where I is an N by N matrix with ones on the diagonal and zeroes elsewhere. To reduce the importance of edge effects, the array can be mapped on to a torus. In this way cells on the left edge of the array become neighbours of those on the right, and similarly for the top and bottom of the array, ensuring that all cells have the full complement of four Rook's case neighbours. An alternative method for dealing with edge effects is to exclude the outer cells from the calculation [see Griffith (1983) for a comprehensive discussion of edge effects].

The variable X has a level of spatial dependence determined by ρ, and a marginal distribution which is known to be Gaussian with a mean of zero. The variance–covariance matrix of X is given by Haining *et al.* (1983):

$$[(I - \rho W)^T (I - \rho W)]^{-1}. \qquad (4)$$

A simple transformation is therefore sufficient to obtain the desired values z_{ij}, with controlled spatial dependence and a uniform distribution in the range 0–1.

As a result of the need to invert an N by N matrix, this method of generating a stationary random field presents difficult computational problems, despite its conceptual simplicity. King and Smith (1988) used a spectral approach to calculate the inverse (see also Larimore 1977). Heuvelink (in press) has proposed a computationally simple iterative method which can be implemented for any matrix W, whereas the approach used in this paper is effectively restricted to correlation with Rook's case neighbours. For the examples in this paper, it was possible to solve equation (2) directly by matrix inversion, by exploiting the inherent simplicity of W and reducing inversion to the use of a set of simple rules (an algorithm coded in C is available from the authors). In this way it has been possible to simulate arrays as large as 512 × 512 (W dimensioned to $512^2 \times 512^2$). However, Heuvelink's method provides a much more practical and robust approach and it is recommended that it is used in practice.

This process was used to obtain figures 2 and 3, based on arrays of 64 rows and 64 columns. The upper panels in these figures were obtained by setting ρ to 0·050; $\rho = 0·230$ in the centre panels and $\rho = 0·247$ in the lower panels. Inclusions grow very rapidly as ρ approaches 0·250, and edge effects and numerical errors in the inversion become important near the limit.

5. Application I

The data quality statements that often accompany digital spatial databases provide useful information on data accuracy. For example, the accuracy of a digital elevation model is often characterized by a root mean square error (RMSE) in its elevation measurements. An RMSE of 7 m indicates that the square root of the average squared difference between the true and recorded elevations at randomly chosen points in the mapped area is 7 m. However, such measures are not sufficient to estimate the uncertainty associated with products of geographical information systems. In the elevation case, the RMSE is not sufficient to place confidence limits on slopes or aspects calculated from the digital elevation matrix, or on more complex products such as the areas of viewsheds (Felleman and Griffin 1990). In these instances it is necessary to have information not only on the RMSE, but also on the spatial dependence of errors (Fisher 1990). The uncertainty in slope estimates is clearly much higher if errors are locally spatially independent than if they are positively correlated.

For similar reasons, although it is possible to describe the accuracy of a categorical variable with a simple misclassification matrix, such measures are insufficient to provide measures of uncertainty on computed products, e.g. the total area of a certain class. The band concept described earlier is a useful description of positional accuracy in a polygon boundary, but not sufficient to place confidence limits on the polygon's area.

However, a stochastic error model, as distinct from a simple descriptive statistic such as the RMSE, is an adequate basis for determining levels of uncertainty in products. Chrisman and Yandell (1988) and Keefer *et al.* (1988) have shown how an error model can be used to characterize the uncertainty associated with digitizing and to estimate the standard error of polygon areas. In this section it is shown how the error model presented earlier can be used to establish confidence limits on area estimates using a simple example.

5.1. *Data*

Fisher (1991) provides an interesting example of the effects of uncertainties in area estimates. In certain areas of the United States the assessed value of agricultural land is based in part on soil class. Soil mapping involves substantial professional judgment and subjectivity and uncertainty is often expressed explicitly in the finished product. A given unit, mapped as one or more polygons, might be described as '80% A with 20% inclusions of B'. As the map gives no information on the locations of inclusions of B, there is substantial uncertainty about the amount of land on a given farm that is actually A, and thus in the assessed value and taxes. The example provides a direct link between geographical uncertainty and economic value.

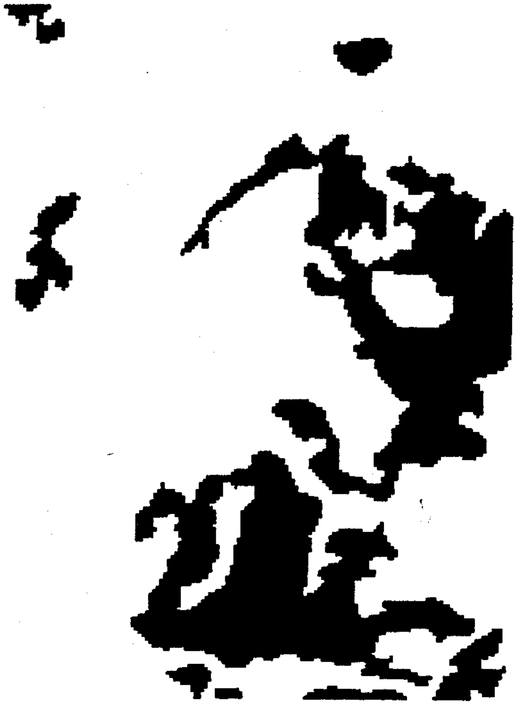


Figure 4. Data used in the example applications, digitized by Fisher (1991) from the Medina, Ohio soil map and simplified by merging three classes.

To illustrate the application of the error model, rectangular study area on the Medina, Ohio, soil map previously coded by Fisher was taken. Some 39 soil classes occur in the 150 × 200 cell area, and three of these classes were arbitrarily merged to form an area of 7513 cells, or approximately 25 per cent of the mapped area. The merged classes were given a value of unity, and all other classes a value of zero. Figure 4 shows the smooth boundaries typical of this series of soil maps, although these are distorted by the coarse pixels.

For the purposes of illustration, it is proposed that class 1 is described as '80% A with 20% inclusions of B'. Thus an area equivalent to 6010 cells would be expected to be truly A. In the following section the use of the error model to estimate the actual area of A and the associated uncertainty is investigated. Uncertainty is measured in two ways: relative to the expected area of 6010 (denoted by s_1), and relative to the mean estimated area (denoted by s_2). In both instances the standard error, or RMSE, is used, which is crudely interpreted as the 'typical' difference between the estimated area and the true value.

5.2. *Subcell mixing*

The example is first considered where A and B are fully mixed at the subcell level. If mixing is spatially homogeneous, then the dominant class in every cell will be A. Thus, the estimated area of A will be heavily biased at 7513 cells and no cells of class B will be observed. As this estimate is constant, s_2 is zero. The uncertainty relative to the expected area, s_1, is much larger because of the bias, at $(7513-6010)=1503$.

5.3. *Independence* ($\rho=0$)

When outcomes are independent in each cell, the binomial distribution can be used to calculate the expected area A and its uncertainty [Fisher (1991) used Monte

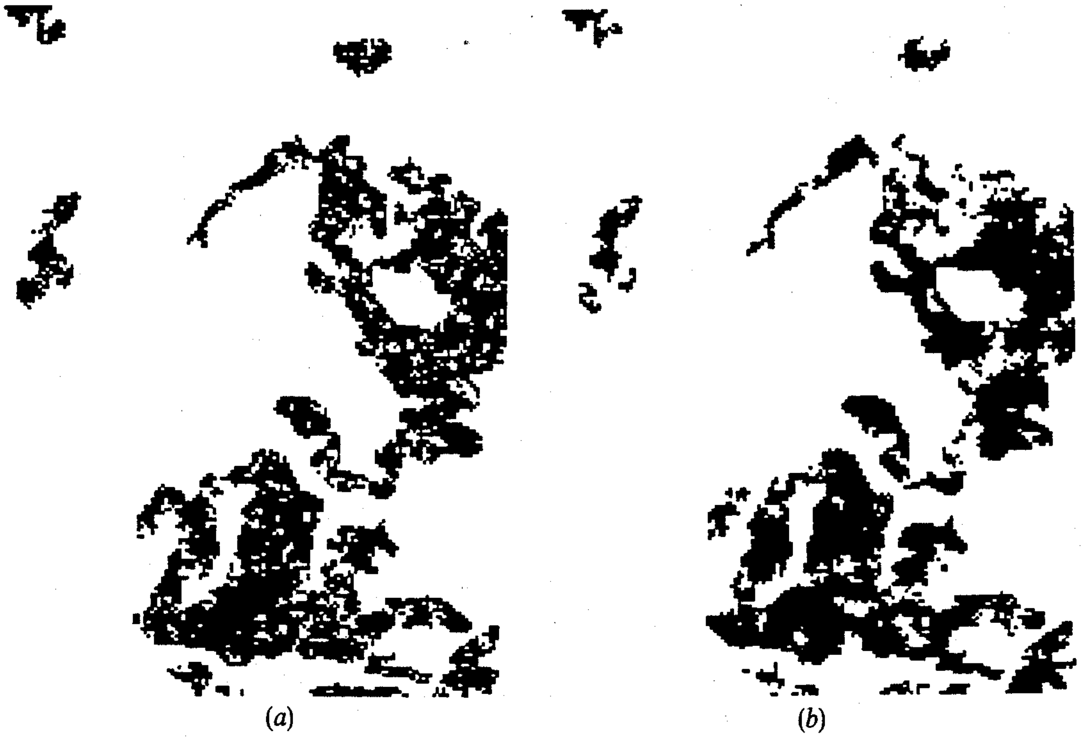
Carlo simulation]. On average, 6010 cells will be classified as A and 1503 as B. From the standard deviation of the binomial, it is known that $s_1 = s_2 = [7513 \times 0{\cdot}80 \times (1 - 0{\cdot}80)]^{0{\cdot}5} = 34{\cdot}7$. Thus, uncertainty about the locations of inclusions might lead to estimates of the area of A as high as 6050, or as low as 5970. More precisely, and approximating the binomial distribution by the Gaussian distribution for large sample sizes, the 95% confidence limits on the area of A are $6010 \pm (1{\cdot}96 \times 34{\cdot}7)$. The observed area can therefore be expected to lie between 5942 and 6078 in 95% of trials. For independence, the size of inclusions is roughly equal to the cell size.

5.4. *Spatial dependence* ($0{\cdot}25 > \rho > 0$)

Spatial dependence is now considered. Using the probability vector $\{0{\cdot}8, 0{\cdot}2\}$ for all cells within class 1, the patterns of inclusions of B for various values of the spatial dependence parameter ρ were simulated. The error model was realized ten times for each value of ρ. Figure 5 shows example realizations for $\rho = 0{\cdot}200$ and $\rho = 0{\cdot}240$. Owing to the high standard error, ten realizations were not sufficient to obtain accurate estimates of the mean, so it is not clear whether the observed deviations from 6010 indicate substantial bias, although it is observed that $s_1 > s_2$. Table 1 shows the results obtained for various values of ρ. Although 0·250 is the theoretical upper limit of the range of ρ, it has nevertheless been possible to obtain useful simulations at this value, and simulated inclusions are large but still less than the size of the mapped area.

5.5. *Limit* ($\rho = 0{\cdot}25$)

In the limit, the entire area of class 1 (7513 cells) is either A with probability 0·80, or B with probability 0·20. In this instance the uncertainty of an estimate of the area of A can be calculated directly, and $s_1 = s_2$.



(*a*) (*b*)

Figure 5. Example realizations of the error model for two levels of spatial dependence and 20% inclusions.

Table 1. Estimates of area and associated uncertainties for various levels of spatial dependence.

ρ	Sample size	Mean estimate	s_1	s_2	DF[a]
0·000	(Calculated)	6010	35	35	7513
0·200	10	6139	156	88	1166
0·240	10	6153	210	150	401
0·245	10	6158	268	223	182
0·249	10	6254	395	311	93
0·250	10	6514	753	559	29
Limit	(Calculated)	6010	3005	3005	

[a] Number of trials needed to produce the observed value of s_2 in a binomial distribution (see text).

5.6. *Summary*

These simulations indicate, as might be expected, that the uncertainty associated with an estimate of the area of A rises rapidly with the level of spatial dependence. As the inclusions of B increase in size, the ability to estimate the true area of A deteriorates, although its theoretical value remains constant at 6010 pixels, or 80 per cent of the polygon area. In Fisher's example, the uncertainty associated with taxes based on measured area is much higher for $\rho > 0$ than under spatial independence, and may be as much as an order of magnitude higher. Moreover, subcell mixing also produces a large discrepancy between the expected area of 6010 and actual estimates.

Suppose, for example, that a farmer pays annual taxes of $6010 based on ownership of 7513 pixels of this soil type, identified as '80% A with 20% inclusions of B'. If the inclusions of B are small ($\rho = 0$), an accurate assessment based on the true area of A mapped pixel by pixel might correct the assessment to as much as $6078 or as little as $5942, or an uncertainty of ±1·1 per cent. However at $\rho = 0{\cdot}249$, with much larger inclusions, the estimated 95% confidence limits become $6884 and $5236, or ±12·9 per cent.

One way to think of spatial dependence is as a loss of effective degrees of freedom (Cliff and Ord 1981). In the limit at $\rho = 0{\cdot}25$, the class of all 7513 cells is assigned by a single trial, whereas at $\rho = 0$ each cell's class is the outcome of a separate, independent trial. In effect, the number of degrees of freedom has changed from 1 to 7513. We can infer the effective number of degrees of freedom for a given ρ by asking what number of independent trials would be required to yield the observed standard error in area estimates, s_2, using the standard deviation of the binomial distribution. The results are shown in table 1, and might be thought of as crudely indicative of the number of inclusions present at each level of ρ.

6. Application II

One of the common applications of geographical information systems is in estimating the area with some combination of characteristics on more than one layer using an overlay process. Several studies have been published (MacDougall 1975, Chrisman 1987, Veregin 1989, Newcomer and Szagin 1984) on the accuracy of overlays and the propagation of uncertainty in the input layers. However, these have been concerned only with estimating the probability that the characteristics ascribed to the overlay are correct. Suppose, for example, that map 1 depicts two classes, 1A and 1B, and assigns a probability vector $\{0{\cdot}8, 0{\cdot}2\}$ to a given point. In a similar fashion map 2 depicts classes 2A and 2B and assigns a probability vector $\{0{\cdot}9, 0{\cdot}1\}$ to the same point. After overlay the point may have any one of four classes—1A2A, 1A2B, 1B2A, or 1B2B—and the associated probability vector will have four components. If the errors are independent on the two maps, the overlay probability vector will be found by simple multiplication to be $\{0{\cdot}72, 0{\cdot}08, 0{\cdot}18, 0{\cdot}02\}$. However, conditional probabilities must be used if correlations exist between the errors on the two maps. This type of correlation is referred to as thematic dependence to distinguish it from the spatial dependence that is likely to exist in each layer.

To estimate the uncertainty of products derived from the overlay, the same problem is faced as before, namely that probability vectors alone are not sufficient to model error, as they contain no information on the spatial dependence of errors within each layer, or in the overlay.

In this section the effects of spatial dependence on the results of overlay are considered, assuming as before that the product of interest is a measure of area. For convenience, overlays of the same Medina data are used, and the task of estimating the area of soil type A is considered as before. In the overlay, the area that is classified as AA is of interest, that is, class A on both input layers. The layers may have the same or different values of ρ, and no attempt will be made to simulate thematic dependence.

Consider two of the realizations used in the application, with the same value of ρ. With 10 realizations, 45 overlays can be generated (100 where ρ is different on the two maps), and that number of estimates of the area that is A on both maps can be made. For $\rho=0$ the binomial distribution can again be used to predict the mean and standard deviation, as the probability that a cell is A on both maps is $0{\cdot}8\times0{\cdot}8=0{\cdot}64$. The expected area is 4808 with a standard error $s_1=s_2=33$. However, standard errors are expected to increase with increasing ρ.

Table 2 shows the results of overlaying all combinations of realizations and values of ρ. Only s_2 is shown, i.e. the uncertainty with respect to the mean estimate rather than the expected estimate. As expected, standard errors are highest when two layers with high spatial dependence are overlaid.

The diagonal terms in table 2 show the uncertainty in area estimates when two layers with the same levels of spatial dependence, but zero thematic dependence, are overlaid. In proportion to the mean estimate, uncertainty in areas estimated from two overlaid layers is roughly 32 per cent higher than in areas estimated from one layer, if spatial dependence is held constant. Table 2 shows the uncertainty to be generally

Table 2. Estimates of area and associated uncertainties (s_2) for overlaid combinations.

ρ	$\rho=0{\cdot}000$	$\rho=0{\cdot}200$	$\rho=0{\cdot}240$	$\rho=0{\cdot}245$	$\rho=0{\cdot}249$	$\rho=0{\cdot}250$
0·000	4808 (33)					
0·200	4911[a]	5015 (94)				
0·240	4922[a]	5109 (243)	5034 (162)			
0·245	4926[a]	5089 (202)	5131 (327)	5037 (243)		
0·249	5003[a]	5141 (134)	5176 (238)	5190 (338)	5193 (341)	
0·250	5211[a]	5338 (97)	5363 (167)	5373 (229)	5478 (323)	5660 (643)

Sample size 45 for diagonal elements, 100 for off-diagonal elements.
[a] Standard errors not determined for overlays with $\rho=0$.

highest where the averaged spatial dependence of the two layers is highest. It is concluded that the effect of a given level of spatial dependence on the ability to estimate area from uncertain layers worsens substantially as layers are overlaid.

7. Implementation

The error model requires that uncertainty is represented in the database in the form of probability vectors, either on a cell by cell basis or for entire polygons, although in the latter instance the model can deal only with heterogeneity (figure 2) and not with transition effects (figure 3). It has already been discussed how such data might be obtained from image classification or from the legends of thematic maps. Probability vectors would clearly take more storage than simple classes, leading to a very substantial increase in the volume of the database. However, it seems likely that in most instances these probability vectors would contain only a few non-zero terms, and assorted compression techniques might be used to cope with the problem of volume.

In this section two further issues in implementing the model are discussed: the determination of ρ and the design of an uncertainty-reporting geographical information system.

7.1. *Determination of ρ*

To characterize spatial dependence, it is necessary to provide the model with a value of ρ. Note that the effects of ρ increase sharply as the maximum value of 1/4 is approached. In this paper ρ has been presented as a property of the map as a whole, and this may be appropriate in examples where the map has been created by the interpretation of imagery at a constant scale. However, in many applications ρ might be allowed to vary regionally. For example, on a land use map it is likely that inclusions in urban areas would require a different value of ρ from inclusions in forested areas, although this clearly presents an interesting paradox at any urban–forest boundary. Techniques to relate image texture to ρ might be helpful, and it might also be possible to generalize the error model to incorporate the more detailed descriptions of spatial dependence that are available from such indicators as the variogram.

In many applications, ρ is seen as being supplied by the user, based on intuitive expectations about the sizes of inclusions. Unfortunately, it is not possible to provide a simple relationship, e.g. between mean inclusion size or minimum mapping unit and ρ, owing to the confounding effects of the probability vectors. For example, table 3 shows mean inclusion size under various combinations of ρ and probability (where probability refers to the class forming the inclusions; the pixels in an inclusion are

Table 3. Mean area of simulated inclusions for various levels of spatial dependence and probability of the included class.

	Probability				
ρ	0·5	0·4	0·3	0·2	0·1
0·200	19	10	6	3	2
0·240	46	27	17	10	6
0·245	62	39	26	17	10
0·249	131	97	73	48	29
0·250	408	304	274	219	116

Rook's case neighbours; means were estimated over 10 realizations and with a total area of 7513 pixels). However, it is possible to simulate the visual appearance of various values of ρ, and to make a subjective choice on that basis.

Ideally, the probabilities and ρ are seen as being captured during the process of interpretation and classification by the worker creating the database from field or remotely sensed data. A soil scientist might be asked to attach percentage estimates to each polygon, to estimate ρ from a set of visual representations of various inclusion patterns and to attach a measure of fuzziness or width to each polygon boundary. In current practice, although much of this information is available, at least as crude estimates to the discipline specialist creating a map, it is lost during the processes of map-making and digitizing.

7.2. *Implications for the design of geographical information systems*

The nature of the error model and its dependence on input probability vectors ensure that no general analytical results can be obtained, and that the model must therefore be applied using simulation. In a conventional raster geographical information system, the area of class A on a layer is measured by counting cells and no measure of uncertainty is given. In a vector system, area is obtained from geometrical calculations on every class A polygon, and again no measure of uncertainty is available.

The error model proposed here must be implemented in the raster domain to accommodate simulation. Thus it is assumed that the database is structured as a set of layers, each consisting of an array of cells and associated probability vectors. A value of ρ is also available for every cell, either as a constant over the mapped area or as a region based variable.

This uncertainty-reporting geographical information system would store permanently a set of realizations of the autoregressive model, in the form of arrays of z_{ij} values. It is expected that ten realizations would be adequate for each of five selected values of ρ (e.g. 0·200, 0·240, 0·246, 0·249, 0·250). The arrays would have to be as large (in rows and columns) as the largest anticipated application.

To measure the area of class A on a given layer, instead of counting cells, the system would compute ten classified maps, using the ten stored realizations of z_{ij} values for the chosen value(s) of ρ. Cells would be counted on each of the ten maps and used to compute a mean estimate and standard error, which would then be reported to the user. The process of simulating error would not be shown to the user, but the system would have the ability to display realizations of classified maps on request for visual assessment. The authors are in the process of implementing this model at the University of California at Santa Barbara using the GRASS system.

Clearly this process has a substantial computational overhead, although the operation would be essentially transparent to the user. If speed of processing is a problem, it might be possible to develop a series of general rules of thumb rather than to invoke simulation on every process. However, in general it would be possible to use this approach to provide uncertainty measures on every product of the geographical information system, not only area estimates.

8. Conclusions

Categorical spatial data are commonly produced by disciplines concerned with the interpretation of complex geographical variation and represent little more than the primitive scientific process of classification applied in a spatial context. As such, they are the basis for many applications of geographical information systems, particularly in resource analysis and management. Seen in this context, the results obtained in this paper are disturbing, although they could be anticipated if the effects of spatial dependence are considered carefully. In reality, both uncertainty (in the form of class probabilities of less than 1) and spatial dependence are almost always present in categorical spatial data, and area estimates are one of the most common products of a geographical information system. Given the magnitudes of the uncertainties shown by this example, it seems essential that greater efforts are made to deal explicitly with the error problem in spatial databases.

In this paper overlay has been considered as a process of combining attributes from two input layers and the influence of thematic dependence between layers has been ignored. In most applications, overlay is followed by some form of reclassification, and the rules that determine the output class as a function of two or more input classes are often complex. The only rule considered here is the 'AND' condition, although Veregin (1989) and others have also considered the 'OR' condition. Given the nature of this model, it seems unlikely that any analytical results could be obtained for these different conditions that go beyond those already known, and that simulation is therefore the only effective way to proceed.

The applications discussed in this paper have been limited to measurement of area. However, the error model could be applied to any other operation of a geographical information system on categorical data. Tasks such as the estimation of the length of the common boundary between two polygons, or of the sizes and shapes of slivers, would require the classified cell realizations presented here to be converted to vector representations of objects, as discussed for figure 3.

The probability vectors and spatial autocorrelation parameter that together characterize the model provide an adequate description of generic problems of within polygon heterogeneity and transition across polygon boundaries. More complex and specific error models would be required to handle examples where spatial dependence is anisotropic, or where detailed information is available on the form of spatial dependence present. In some instances spatial dependence might be expected to be specific to class. However, the model provides a robust, simple and straightforward approach in those instances where very little is known about the forms of error present or the processes contributing to them. For reasons already discussed, this lack of knowledge seems broadly characteristic of the use of map or remote sensing data in geographical information systems.

Acknowledgments

The National Center for Geographic Information and Analysis is supported by the National Science Foundation, grant SES 88-10917. This research constitutes part of the Center's Research Initiative 1 on Accuracy of Spatial Databases. Financial support of the US Geological Survey is also acknowledged, as is a gift of computing time at IBM Corporation's Palo Alto facility for the large matrix inversions.

References

Bolstad, P. V., Gessler, P., and Lillesand, T. M., 1990, Positional uncertainty in manually digitized map data. *International Journal of Geographical Information Systems*, **4**, (4), 399–412.

Burrough, P. A., 1986, *Principles of Geographical Information Systems for Land Resources Assessment* (Oxford: Clarendon Press).

CHRISMAN, N. R., 1982, Methods of spatial analysis based on error in categorical maps. *Unpublished PhD Dissertation*, University of Bristol, Bristol, UK.

CHRISMAN, N. R., 1987, The accuracy of map overlays: a reassessment. *Landscape and Urban Planning*, **14**, 427–439.

CHRISMAN, N. R., and YANDELL, B. S., 1988, Effects of point error on area calculations: a statistical model. *Surveying and Mapping*, **48**, 241–249.

CLIFF, A. D., and ORD, J. K., 1981, *Spatial Processes: Models and Applications* (London: Pion).

DIGITAL CARTOGRAPHIC DATA STANDARDS TASK FORCES (DCDSTF), 1988, The proposed standard for digital cartographic data. *The American Cartographer*, **15**, 1–140.

DUNN, R., HARRISON, A. R., and WHITE, J. C., 1990, Positional accuracy and measurement error in digital databases of land use: an empirical study. *International Journal of Geographical Information Systems*, **4**, 385–398.

FELLEMAN, J., and GRIFFIN, C., 1990, The role of error in GIS-based viewshed determination—a problem analysis. Report No. IEPP-90-2, Institute for Environmental Policy and Planning, College of Environmental Science and Forestry, State University of New York, Syracuse, NY.

FISHER, P. F., 1990, Simulation of error in digital elevation models. *Papers and Proceedings of Applied Geography Conferences*, **13**, 37–43.

FISHER, P. F., 1991, Modelling soil map—unit inclusions by Monte Carlo simulation. *International Journal of Geographical Information Systems*, **5**, 193–208.

GENDEREN VAN, J. L., and LOCK, B. F., 1977, Testing land-use map accuracy. *Photogrammetric Engineering and Remote Sensing*, **43**, 1135–1137.

GOODCHILD, M. F., 1989, Modelling error in objects and fields. *Accuracy of Spatial Databases*, edited by M. F. Goodchild and S. Gopal (London: Taylor & Francis), pp. 107–114.

GOODCHILD, M. F., and DUBUC, O., 1987, A model of error for choropleth maps with applications to geographic information systems. *Proceedings, Auto Carto 8* (Falls Church, VA: ASPRS/ACSM), pp. 165–174.

GOODCHILD, M. F., and WANG, M.-H., 1988, Modelling error in raster-based spatial data. *Proceedings, Third International Symposium on Spatial Data Handling* (Columbus, OH: IGU Commission on Geographical Data Sensing and Processing), pp. 97–106.

GRIFFITH, D. A., 1983, The boundary value problem in spatial statistical analysis. *Journal of Regional Science*, **23**, 377–387.

HAINING, R., GRIFFITH, D. A., and BENNETT, R., 1983, Simulating two-dimensional autocorrelated surfaces. *Geographical Analysis*, **15**, 247–255.

HEUVELINK, G. B. M., An iterative method for multi-dimensional simulation with nearest neighbour models. *Proceedings of the Second CODATA Conference on Geomathematics and Geostatistics, 10–14 September 1990, Leeds*, edited by P. A. Dowd (Sciences de la Terre), in press.

KEEFER, B. J., SMITH, J. L., and GREGOIRE, T. G., 1988, Simulating manual digitizing error with statistical models. *Proceedings, GIS/LIS '88*, Vol. 2 (Falls Church, VA: ASPRS/ACSM), pp. 475–483.

KING, P. R., and SMITH, P. J., 1988, Generation of correlated properties in heterogeneous porous media. *Mathematical Geology*, **20**, 863–877.

LARIMORE, W. E., 1977, Statistical inference on stationary random fields. *Proceedings of the IEEE*, **65**, 961–970.

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

MARK, D. M., and CSILLAG, F., 1989, The nature of boundaries on 'area-class' maps. *Cartographica*, **26**, 65–78.

MEAD, R. A., and SZAJGIN, J., 1982, Landsat classification accuracy assessment procedures. *Photogrammetric Engineering and Remote Sensing*, **48**, 139–141.

NEWCOMER, J. A., and SZAJGIN, J., 1984, Accumulation of thematic map error in digital overlay analysis. *The American Cartographer*, **11**, 58–62.

TOMLINSON, R. F., CALKINS, H. W., and MARBLE, D. F., 1976, *Computer Handling of Geographical Data* (Paris: UNESCO).

VEREGIN, H., 1989, Error modelling for the map overlay operation. *Accuracy of Spatial Databases*, edited by M. F. Goodchild and S. Gopal (London: Taylor & Francis), pp. 3–18.