CHAPTER FIVE[1]

ISSUES OF QUALITY AND UNCERTAINTY

Michael F. Goodchild

National Center for Geographic Information and Analysis

University of California

Santa Barbara, CA 93106, USA

INTRODUCTION

Digital processing of cartographic data brings immense benefits in the form of rapid, precise and sophisticated analysis, but at the same time it reveals weaknesses which may not otherwise be apparent. Computers are very precise machines, and errors and uncertainties in data can lead to serious problems, not only in the form of inaccurate results but in the consequences of decisions made on the basis of poor data, and increasingly in legal actions brought by affected parties.

In this chapter we first review the dimensions of the accuracy problem, and discuss the premises and assumptions on which the remainder of the chapter is based. The second section reviews the models which are available for analyzing and understanding errors in spatial data. The chapter concludes with a summary of the current state of the art and an agenda for future research and development.

The terms accuracy, precision, resolution and scale are used almost interchangeably in reference to spatial data, but we will need to establish more exact definitions for the purposes of this discussion. **Accuracy** refers to the relationship between a measurement and the reality which it purports to represent. **Precision** refers to the degree of detail in the reporting of a measurement or in the manipulation of a measurement in arithmetic calculations. Finally, the **resolution** of a data set defines the smallest object or feature which is included or discernable in the data. Scale and resolution are intimately related because there is a lower limit to the size of an object which can be usefully shown on a paper map. This limit is often assumed to be 0.5mm as a rule of thumb, so the effective resolution of a 1:1,000 map is about 1,000 times 0.5mm, or 50cm,

---

[1] Draft of a chapter to appear in *Cartographic Research Agenda in the 1990s*, edited by J.-C. Muller, to be published by Elsevier in 1991

0.5mm as a rule of thumb, so the effective resolution of a 1:1,000 map is about 1,000 times 0.5mm, or 50cm, although the standards of most mapping agencies are substantially better. The effective resolutions of some common map scales, assuming 0.5mm resolution, are shown in Table 5.1.

Table 5.1 about here

Because digital spatial data sets are not maps, and have no similar physical existence, they do not have an obvious scale, and it is common to think of such data as essentially scale-free. But this is to oversimplify. If the data were obtained by digitizing, their resolution is that of the digitized map. If they were obtained from imagery, their resolution is that of the pixel size of the imagery. In spite of appearances to the contrary, then, no spatial data is ever resolution free. The term 'scale' is often used loosely as a surrogate for database resolution.

DIMENSIONS OF THE PROBLEM

Precision

One of the widely accepted design principles of spatial data handling systems is that they should be able to process data without significant distortion. It is important when calculations are carried out on coordinates, for example, that the results should not be affected by lack of precision in the machine. We have seen that a figure of 0.5mm can be assumed to be typical for the resolution of an input map document. 0.5mm is also roughly the accuracy with which an average digitizer operator can position the crosshairs of a cursor over a point or line feature. Most digitizing tables have a precision of 0.25mm or better, thus ensuring that the table itself does not introduce additional distortion as points are captured.

If we take the size of the source map sheet on the digitizing table to be 100cm by 100cm, then a precision of 0.5mm represents an uncertainty of 0.05% relative to the size of the sheet, or an uncertainty in the fourth significant digit. So in order not to introduce distortion, internal storage and processing of coordinates must be correct to at least four digits. However seven or eight decimal digits are commonplace in arithmetic calculations in spatial data handling systems, and many GISs operate at much higher precision, in effect allowing GIS users to ignore the possibility of arithmetic distortion and to assume that internal processing is carried out with infinitely high precision.

Despite this, arithmetic precision, or lack of it, is sometimes a problem in spatial data processing. One common situation occurs when point locations are specified in a global coordinate system. For example, the UTM system assigns coordinates with respect to an origin on the equator, so locations in the mid-Northern latitudes will have one of their coordinates in the millions of metres. Processing of data taken from large scale maps, perhaps 1:1,000, will therefore require precision in the sixth or seventh significant digit. This can be particularly problematic in performing operations such as calculation of the intersection point of two lines, or the centroid of a parcel of land, where differences in the products of six or seven digit numbers become important.

Another artifact of high precision is well known in connection with polygon overlay operations. It is common to find that two coverages of different themes for the same area share certain lines. For example the shoreline of a lake will appear both on a map of soils and a map of vegetation cover. When soils and vegetation cover are overlaid the two versions of the shoreline should match perfectly. In practice this is never the case. Even if the two lines matched perfectly on the input documents, errors in registration and differences in the ways in which the digitizer operators captured the lines will ensure different digital representations in the database. Moreover since the overlay operation is carried out to high precision, the differences are treated as real and become small sliver polygons. Table 5.2, from Goodchild (1979), shows the results of overlaying five layers from the Canada Geographic Information System, and the enormous numbers of very small polygons produced.

Table 5.2 about here

Several approaches have been taken to deal with the spurious polygon problem in polygon overlay, which has the annoying property that greater accuracy in digitizing merely leads to greater numbers of slivers. Some systems attempt to distinguish between spurious and real polygons after overlay, using simple rules, and to delete those which are determined to be spurious (see Figure 5.1). A suitable set of rules to define spurious polygons might be:

- small in area;
- long and thin in shape (high ratio of perimeter to square root of area);
- composed of two arcs only;

- both nodes have exactly four incident arcs.

In addition, the attributes of spurious polygons are characteristic. Suppose our soil map line marks the boundary between soils A and B, and our vegetation map shows the boundary between classes 1 and 2. Ideally, when overlaid the result should be a line with soil A and vegetation 1 on one side, and soil B and vegetation 2 on the other. Sliver polygons will be either soil B and vegetation 1, or soil A and vegetation 2. Moreover the attributes will alternate from one sliver to the next, as the soil and vegetation boundaries cross and recross (see Figure 5.1). So the detection of one spurious polygon provides additional evidence that nearby polygons are also spurious.

Figure 5.1 about here

Another approach to dealing with the sliver problem is in effect to reduce the precision of the overlay operation, by requiring the user to establish a tolerance distance. Any lines lying within the tolerance distance of eachother are assumed to be the same. A similar approach is often taken in digitizing, where two lines are assumed to join or 'snap' if the distance between them is less than some user-established tolerance distance (see Figure 5.2a).

Figure 5.2 about here

Because a tolerance distance is in effect a reduced level of spatial resolution, all objects or features smaller than the tolerance become ambiguous or indistinguishable. Figure 5.2c shows an example of a complex polygon with a narrow strait or isthmus. If the width of the strait is less than the tolerance distance, it is impossible to determine from the geometry of the feature alone whether it is indeed a single polygon with a narrow strait, or two polygons. In some cases it may be possible to resolve the ambiguity if the user has already identified labels or attributes for the polygon(s); if two labels have been input, it should be clear that the feature consists of two polygons. In such instances the **topology** (the existence of two polygons) is allowed to resolve ambiguity in the **geometry**.

The potential conflict between topology and geometry is a consistent theme in spatial data handling. Franklin (1984; Franklin and Wu 1987) has written about specific instances of conflict, and about ways of resolving them. For example, a point may be given an attribute indicating that it lies inside a given polygon,

which might be a county, but geometrically, because of digitizing error or other problems, the point may lie outside the polygon (Blakemore 1984). The conflict may be resolved by moving the point inside the polygon, or by moving the polygon boundary to include the point, or by changing the point's attributes. The first two are examples of allowing topology to correct geometry, and the third of allowing geometry to correct topology. In all cases, and particularly in the second, the implications of a change on other objects and relationships must be considered; if the polygon's boundary is moved, do any other points now change properties? Unfortunately the propagation of changes of this type can have serious effects on the overall integrity of the database.

In summary, the average user will approach a GIS or spatial data handling system with the assumption that all internal operations are carried out with infinite precision. High precision leads to unwanted artifacts in the form of spurious polygons. Moreover in reducing effective precision in such operations as digitizing and polygon closure it is common for unwanted effects to arise in the form of conflicts between the topological and geometrical properties of spatial data. The ways in which system designers choose to resolve these conflicts are important to the effective and data-sensitive application of GIS technology.

Accuracy

Thus far we have used examples of the distortions and errors introduced into spatial data by digitizing and processing. However if we are to take a comprehensive view of spatial data accuracy it is important to remember that accuracy is defined by the relationship between the measurement and the reality which it purports to represent. In most cases reality is not the source document but the ground truth which the source document models. The map or image is itself a distorted and abstracted view of the real world, interposed as a source document between the real world and the digital database.

We will use two terms to distinguish between errors in the source document and in the digitizing and processing steps. **Source** errors are those which exist in the source document, and define its accuracy with respect to ground truth. **Processing** errors are those introduced between the source document and the GIS product, by digitizing and processing. We will find that in general processing errors are relatively easier to measure, and smaller than source errors.

All spatial data without exception are of limited accuracy, and yet it is uncommon to find statements of data quality attached to such data, whether in the form of source maps, images or digital databases. The

accuracy of much locational or attribute data is limited by problems of measurement, as for example in the determination of latitude and longitude, or elevation above mean sea level. In other cases accuracy is limited by problems of definition, as when a soil type is defined using terms such as 'generally', 'mostly', 'typically' and thus lacks precise, objective criteria. Locations defined by positions on map projections are limited by the accuracy of the datum, and may change if the datum itself is changed.

More subtle inaccuracies result from the way in which a source document models or abstracts reality. For example an extended object such as a city may be shown at some scales as a point, or a census reporting zone may be located by its centroid or some other representative point (Bracken and Martin 1989). The transition between two soil types, which occurs in reality over some extended zone, may be represented on the source document and in the database by a simple line or sharp discontinuity. The attributes assigned to a polygon may not in fact apply homogeneously to all parts of the polygon, but may be more valid in the middle and less valid toward the edges (Mark and Csillag 1989). All of these errors result from representing spatial variation using objects; the objects are either of the wrong type for accurate representation, in the case of the city represented by a point, or used to model what is in fact continuous variation. Models such as these allow complex spatial variation to be expressed in the form of a comparatively small number of simple objects, but at a corresponding cost in loss of accuracy. Unfortunately the process of modeling has usually occurred in the definition of the source document, long before digitizing or processing in a GIS, and rarely is any useful information on levels of accuracy available.

Separability

Most digital spatial data models distinguish clearly between locational and attribute information. From an accuracy perspective, errors in location and attributes are often determined by quite different processes, and in these cases we term the two types of error **separable**. For example, a reporting zone such as a census tract may be formed from segments which follow street center lines, rivers, railroads and political boundaries. The processes leading to error in the digitizing of a river-based boundary are quite different to those typical of street center lines. Errors in imagery are similarly separable, attribute errors being determined by problems of spectral response and classification, while locational errors are determined by problems of instrument registration. The same is true to a lesser extent for topographic data, since errors in determining elevation at a point are probably only weakly dependent on errors in locating the point horizontally.

However in maps of vegetation cover, soils or geology, where polygon objects are commonly used to model continuous variation, the two types of error are not separable, as object boundaries are derived from the same information as the attributes they separate (Goodchild 1989). The transition zone between types A and B may be much less well defined than the transition between types A and C. The process by which such maps are created gives useful clues to the likely forms of error which they contain. A forest cover map, for example, will be derived from a combination of aerial imagery and ground truth. Polygons are first outlined on the imagery at obvious discontinuities. Ground surveys are then conducted to determine the attributes of each zone or polygon. The process may be iterative, in the sense that the ground survey may lead to relocation of the interpreted boundaries, or the merger or splitting of polygons. Figure 5.3 summarizes these processes of source error generation, and the subsequent steps which lead to processing errors.

Figure 5.3 about here

The forest example illustrates the relationship between different error processes particularly well. In forest management the homogeneous zones identified in the mapping process are termed **stands**, and become the basic unit of management. If a stand is cut and reseeded or allowed to regenerate, the processes determining its attributes may be no longer related to the processes determining its boundaries, and thus attribute and locational errors will become separable. Similarly, even in the initial map obtained from imagery and ground survey there will be boundary segments which follow roads, rivers or shorelines and again are subject to error processes which are independent of those affecting attributes.

The ideal

The arguments in the preceding sections lead to a clear idea of how an ideal GIS might be structured. First, each object in the database would carry information describing its accuracy. Depending on the type of data and source of error, accuracy information might attach to each primitive object, or to entire classes of objects. Accuracy might be coded explicitly in the form of additional attributes, or implicitly through the precision of numerical information.

Every operation or process within the GIS would track error, ascribing measures of accuracy to every new attribute or object created by the operation. Uncertainty in the position of a point, for example, would be used to determine the corresponding level of uncertainty in the distance calculated between two points, or

in the boundary of a circle of specified radius drawn around the point.

Finally, accuracy would be a feature of every product generated by the GIS. Again it might be expressed either explicitly or implicitly in connection with every numerical or tabular result, and means would be devised to display uncertainty in the locations and attributes of objects shown in map or image form.

Of course the current state of GIS technology falls short of this ideal in all three areas. We currently lack comprehensive methods of describing error, modeling its effects as it propagates through GIS operations, and reporting it in connection with the results of GIS analysis. The remaining sections of this chapter describe the current state of knowledge and such techniques as currently exist for achieving a partial resolution of the error problem.

MODELS OF ERROR

Background

With the basic introduction to accuracy and error in the previous section, we can now consider the specific issue of quality in cartographic data. This is not a simple and straightforward extension; as we will see, spatial data requires a somewhat different and more elaborate approach.

The objective is to find an appropriate model of the errors which occur in cartographic data. A model is taken here to mean a statistical process whose outcome emulates the pattern of errors observed in real digital spatial data. The first subsection considers models of error in the positioning of simple points; the second section looks at the ways this can be extended to more complex line or area features.

Points

The closest analogy between the classical theory of measurement and the problem of error in cartographic data concerns the location of a single point. Suppose, for example, that we wished accurately to determine the location, in coordinates, of a single street intersection. It is possible to regard the coordinates as two separate problems in measurement, subject to errors from multiple sources. If the coordinates were in fact determined by the use of a transparent roamer on a topographic sheet, then this is a fairly accurate model of reality. Our conclusion might be that both coordinates had been determined to an accuracy of 100m, or 2mm on a 1:50,000 sheet. This is illustrated in Figure 5.4.

If the errors in both coordinates are represented by classic bell curves, then we can represent accuracy

in the form of a set of ellipses centred on the point. If the accuracies are the same in each coordinate and if the errors are independent, then the ellipses become circles, again centred on the point. This gives us the circular normal model of positional error, as the two-dimensional extension of the classic error model. The error in position of a point is seen in terms of a bell-shaped surface centred over the point (Figure 5.4). The probability that any point is the true location is measured by the height of the surface of the bell over the point, and is clearly greatest at the measured location, decreasing symmetrically in all directions. Using the model we can compute the probability that the true location lies within any given distance of the measured location, and express the average distortion in the form of a standard deviation.

The **Circular Standard Error** is equal to the standard errors in the two coordinates, and can be portrayed by drawing a circle which passes through points representing one standard error to the left and right of the point in the x direction, and similarly above and below the point in the y direction (Figure 5.4). When the standard errors in x and y are not equal, the Circular Standard Error is approximated as the mean of the two. This approximation is valid only for small differences in the two standard errors, but it is difficult to imagine circumstances in which this would not be true. The probability that a point's true location lies somewhere within the circle of radius equal to the Circular Standard Error is 39.35%. The **Circular Map Accuracy Standard** (CMAS) requires that no more than 10% of the errors on a map will exceed a given value. For example, with a CMAS of 1mm, in order to meet the standard no more than 10% of points should be more than 1mm from their true positions. Under the circular normal model a radius of 2.146 times the circular standard error will contain 90% of the distribution, with 10% lying outside. The relationship between the CMAS and the circular standard error is therefore a ratio of 2.146.

These indices have been incorporated into many of the widely followed standards for map accuracy. For example, the NATO standards rate maps of scales from 1:25,000 to 1:5,000,000 as A, B, C or D as follows:

A: CMAS = 0.5mm (e.g. 12.5m at 1:25,000)

B: CMAS = 1.0mm (e.g. 25m at 1:25,000)

C: CMAS determined but greater than 1.0mm

D: CMAS not determined.

Many other mapping programs use CMAS as the basis for standards of horizontal or planimetric accuracy,

and use similar values.

Lines and areas: the Perkal band

In vector databases lines are represented as sequences of digitized points connected by straight segments. One solution to the problem of line error would therefore be to model each point's accuracy, and to assume that the errors in the line derived entirely from errors in the points. This is illustrated in Figure 5.5. Unfortunately this would be inadequate for several reasons. First, in digitizing a line a digitizer operator tends to choose points to be captured fairly carefully, selecting those which capture the form of the line with the greatest economy. It would therefore be incorrect to regard the points as randomly sampled from the line, or to regard the errors present in each point's location as somehow typical of the errors which exist between the true line and the digitized representation of it.

Figure 5.5 about here

Secondly, the errors between the true and digitized line are not independent, but instead tend to be highly correlated (Keefer, Smith and Gregoire 1988; Amrhein and Griffith 1987). If the true line is to the east of the digitized line at some location along the line, then it is highly likely that its deviation immediately on either side of this location is also to the east by similar amounts. Much of the error in digitized lines results from misregistration, which creates a uniform shift in the location of every point on the map. The relationship between true and digitized lines cannot therefore be modelled as a series of independent errors in point positions.

One commonly discussed method of dealing with this problem is through the concept of an error band. Figure 5.6 shows the observed line surrounded by a band of width epsilon, known as the Perkal epsilon band (Perkal, 1956, 1966; Blakemore, 1984; Chrisman, 1982). The model has been used in both deterministic and probabilistic forms. In the deterministic form, it is proposed that the true line lies within the band with probability 1.0, and thus never deviates outside it. In the probabilistic form, on the other hand, the band is compared to a standard deviation, or some average deviation from the true line. One might assume that a randomly chosen point on the observed line had a probability of 68% of lying within the band, by analogy to the percentage of the normal curve found within one standard deviation of the mean.

Figure 5.6 about here

Some clarification is necessary in dealing with line errors. In principle, we are concerned with the differences between some observed line, represented as a sequence of points with intervening straight line segments, and a true line. The gross misfit between the two versions can be measured readily from the area contained between them, in other words the sum of the areas of the spurious or sliver polygons. To determine the mismatch for a single point is not as simple, however, since there is no obvious basis for selecting a point on the true line as representing the distorted version of some specific point on the observed line. Most researchers in this field have made a suitable but essentially arbitrary decision, for example that the corresponding point on the true line can be found by drawing a line from the observed point which is perpendicular to the observed line (Keefer, Smith and Gregoire, 1988). Using this rule, we can measure the linear displacement error of any selected point, or compute the average displacement along the line.

Although the Perkal band is a useful concept in describing errors in the representation of complex objects and in adapting GIS processes to uncertain data, it falls short as a stochastic process model of error. An acceptable model would allow the simulation of distorted versions of a line or polygon, which the Perkal band does not, and would provide the basis for a comprehensive analysis of error.

Alternatives to the Perkal band

Goodchild and Dubuc (1987) have described a stochastic process which might be used to simulate the effects of error in soil, vegetation or forest cover maps. Consider the surface shown in Figure 5.7, which has the following properties. It is continuous, the elevation at every grid cell being a small displacement up or down from its neighbors, and therefore strongly spatially autocorrelated. It is also random, the surface having been generated by one of a number of stochastic processes known to produce random fields of this nature. In fact the surface was generated by the fractional Brownian process (Mandelbrot, 1982), although other methods such as moving bands, spatially autoregressive and Markov processes have been described (for review see Haining, Griffith and Bennett, 1983). The fractional Brownian process has the property of self-affinity, that is, in principle, that any part of the surface if suitably expanded would be indistinguishable in its statistical properties from the surface as a whole.

Figure 5.7 about here

Suppose now that the surface is classified by assigning each pixel or grid cell a color or class based on its elevation, using a set of elevation classes or ranges. The result would be a standard contour map, in which the image is divided into a number of bands, each corresponding to a range of elevation, with the rule that classes may not be adjacent in the image unless the corresponding ranges of elevation are adjacent. The boundaries of the polygons so formed may not intersect. Finally, boundaries can be distorted to simulate the effects of digitizing errors by adding a suitably autocorrelated surface of small relief to the existing one and recontouring.

To simulate the appearance of a soil or vegetation map, Goodchild and Dubuc (1987) took two independently generated surfaces and passed each pixel through a simple classifier as in the example shown in Figure 5.8. This process is analogous to the control of ecological zones by temperature and precipitation in the model of Holdridge et al. (1971). The resulting map has the required properties; boundaries intersect in nodes, and have a degree of fragmentation and boundary wiggliness which depends on the ruggedness of the underlying surfaces. Distorted versions of the map can be produced by adding autocorrelated surfaces to one or both of the underlying surfaces, as in the two versions of the same map crossclassified in Table 5.3. Moreover the degree of distortion generated in each line will be a function of the classes which the line separates. In the simulated map shown in Figure 5.9 the boundaries have been smoothed by using a spline function, simulating the tendency of interpreters to smooth out lines of discontinuity in images.

Figure 5.8 about here

Figure 5.9 about here

The model is useful as a simulation of the distortion process which can be used to study the effects of uncertainty on GIS operations. However the number of parameters in the model, which include the number and properties of the underlying surfaces, the pixel size, the spline function and the classifier, argues against its calibration against real data and therefore against its use in characterizing the error in real datasets. In practice, GIS source maps tend to be highly abstracted, and to contain very little information on which a

practical model of error might be based. It is only by returning to the original sources of the map's information that we can learn more about the nature of map error. In fact the map itself may be a barrier to better error modeling, by interposing a high level of abstraction between the raw data and the spatial database. Error modeling might be much easier if the raw imagery and ground surveys could be used to create a spatial database, and the abstract map obtained from the database by some objective process, rather than the reverse.

When the source of information is a remotely sensed image, a similar situation arises when pixels are classified and the result passed to a GIS for analysis, as this process similarly discards information which might be useful in modeling error. Many classification methods generate probabilities of class membership for each pixel, although usually only the most likely class is used. If the complete set of probabilities were passed to the GIS, then it would be possible to use these as the basis for uncertainty in GIS products. Goodchild and Wang (1988) describe a stochastic process of object distortion based on this concept. In effect, both of these methods provide links between uncertainty in the continuous variation of a field, and uncertainty in the objects used to model the field in a spatial database for phenomena such as soils and vegetation.

Models for dot and pixel counting

One of the traditional methods of measuring area from a map involves placing an array of grid cells or dots over the area to be measured, and counting. In principle this process is similar to that of obtaining the area of a patch from an image by counting the pixels which have been classified as belonging to or forming the patch. The accuracy of the area estimate clearly depends directly on the density of dots or the pixel size, but so does the cost of the operation, so it would be useful to know the precise relationship in order to make an informed judgment about the optimum density or size.

The literature on this topic has been reviewed by Goodchild (1980). In essence two extreme cases have been analyzed, although intermediates clearly exist. In the first the area consists of a small number of bounded patches, often a single patch; in the second, it is highly fragmented so that the average number of pixels or dots per fragment of area is on the order of one or two. We refer to these as cases A and B respectively. Case A was first analyzed by Frolov and Maling (1969), and later by Goodchild (1980). The limits within which the true area is expected to lie 95% of the time, expressed as a percentage of the area

being estimated, are given by:

$$\pm\ 1.03\ (kn)^{1/2}\ (SD)^{-3/4} \qquad (5.1)$$

where n is the number of patches forming the area;

k is a constant measuring the contortedness of the patch boundaries, as the ratio of the perimeter to 3.54 times the square root of area (k=1 for a circle);

S is the estimated area; and

D is the density of pixels per unit area.

The results for Case B follow directly from the standard deviation of the binomial distribution, since this case allows us to assume that each pixel is independently assigned. The limits within which the true area is expected to lie 95% of the time, again as a percentage of the area being estimated, are given by:

$$\pm\ 1.96\ [1 - \dot{S}/S_0]^{1/2}\ (S/D)^{1/2} \qquad (5.2)$$

where S_0 is the total area containing the scattered patches.

Error rises with the 3/4 power of cell size in case A, but with the 1/2 power in case B. Thus a halving of cell size will produce a more rapid improvement in error in case A, other things being equal, because only cells which intersect the boundary of the patch generate error in case A, whereas all cells are potentially sources of error in case B, irrespective of cell size.

Error in images

Images such as those derived from remote sensing are composed of pixels (picture elements) of uniform size and shape, with associated measures of reflected or emitted radiation. After classification, each pixel is assigned to one of a number of classes, often of land use or land cover. Unlike the cartographic case, there are no objects or features to be located, and accuracy is simply a function of the errors in the assignment of classes to each pixel. The standard method of measuring accuracy in a classified image is to compare the classes assigned to a sample of pixels to the true classes on the ground ('ground truth') and express the result in the form of a table, the rows representing the assigned class and the columns the ground truth. Pixels which fall on the diagonal of the table are correctly classified; pixels which fall off the diagonal in a column are termed 'errors of omission' since the true value is omitted from the assigned class; and pixels which fall off the diagonal in a row are termed 'errors of comission' since they appear as false occurrences of a given class in the data. The contents of the table can be summarized in statistics such as the percentage of cells

correctly classified, or using Cohen's Kappa statistic, which allows for correct classification by chance (see Congalton, Oderwald and Mead 1983; van Genderen and Lock 1977; van Genderen, Lock and Vass 1978; Greenland, Socher and Thompson 1985; Mead and Szajgin 1982; Rosenfield 1986; Rosenfield and Fitzpatrick-Lins 1986).

PROJECTIVE SUMMARY

As we have seen, techniques for dealing with uncertainty in spatial data are much better developed in some areas than others. Uncertainties in point locations are comparatively easy to deal with, and a significant proportion of geodetic science is devoted to dealing with the issue of accuracy in control networks. Substantial progress has been made in modeling the error introduced by digitizing, and in predicting its effects on the measures of perimeter and area of objects (Griffith 1989; Chrisman and Yandell 1988). But the more general problem of modeling the accuracy of spatial data with respect to ground truth remains. In part this is because of the complexity of the processes involved, and in part because many ground truth definitions are themselves uncertain. But in addition, we have seen how the representation of spatial variation as objects acts in many cases to make error modeling more difficult, simply because of the nature of the object data model. Both of the models of error in area-class maps presented above are defined first in terms of continuous variation, using the raster data model, and then used as the basis for deriving uncertainty in objects. If the accuracy issue leads us to question the use of traditional data models then any progress will certainly be slow.

The accuracy of spatial databases is the topic of the first research initiative of the National Center for Geographic Information and Analysis (NCGIA), which began with a meeting of specialists in Montecito, CA in December 1988. The collection of papers from that meeting has been published (Goodchild and Gopal 1989) and provides a broad overview of the accuracy issue in spatial data handling generally. The meeting identified an eleven-point research agenda:

- data structures and models;
- models of error and distortion;
- error propagation;
- product uncertainty and sensitivity;
- risk analysis;

- accuracy concerns among users and agencies;
- experimentation and measurement;
- error reduction methods;
- interpolation and surface modeling;
- aggregation, disaggregation and modifiable areal units;
- regularity and stability.

If accuracy is a problem in spatial data handling because of the high precision of the handling system, then one rational approach would be to reduce precision to match accuracy. In effect this is what happens in a raster system. Dutton (1984 1989; see also Goodchild and Yang 1989) makes this point, and proposes a tesseral scheme for global data that would replace common coordinate referencing with a finite-resolution hierarchical key.

There has been substantial progress in recent years in understanding the propagation of uncertainty that occurs in spatial data handling. Newcomer and Szagin (1984) and Veregin (1989c) have discussed error propagation during simple Boolean overlay; Lodwick (1989; Lodwick and Monson 1988) and Heuvelink et al. (1989) have analyzed the sensitivity of weighted overlay results to uncertainty in the input layers and weights; and Arbia and Haining (1990) have analyzed a general model of uncertainty in raster data. However overlay is perhaps the most simple of the primitive spatial data handling operations. Much more research is needed on the effects of buffering (dilating) uncertain objects, and of changing from one data model to another, e.g. raster/vector conversion.

If traditional map data models tend to give a misleading impression of the reliability of data, it is not surprising that many cartographic products fail to make the user aware of uncertainties. The move to standards of data quality (see Chapter 12) is a significant step toward greater awareness of the reliability of spatial data, and includes concepts of lineage, consistency and completeness as well as the more statistical issues discussed in this chapter. It is particularly important as GIS users continue to apply spatial data to purposes for which they may not have been designed.

This chapter has covered a range of material broadly connected with the issue of accuracy in spatial databases. As we noted at the outset, there are good reasons for believing that GIS technology will enhance our sensitivity to the nature of spatial data, and will increase the need to develop comprehensive and precise

models of its errors and distortions. This enhanced understanding is likely to come not from study of the abstract objects which populate our maps and databases, but from the processes which created spatial variation in the first place and by which that variation was interpreted and modeled. In this sense GIS technology will help to fill many of the gaps which currently exist in our understanding of spatial differentiation and the processes which produce it.

It seems unlikely that the ideal accuracy-sensitive GIS which was outlined earlier in this chapter will ever become a reality. Its analog in simple, aspatial measurement, the statistical analysis system which tracks uncertainty as an attribute of each of its data elements and through each of its procedures, is also readily justified but equally far from implementation at this time. On the other hand standard errors of estimate and confidence intervals are a common product of statistical systems, and it is not difficult to imagine similar products from GIS procedures.

One of the results of the information age is an erosion of the role of the central data collection agencies, and society has not yet begun to come to grips with a growing need to reorganize the entire system of public data provision. It is now possible to determine position at low cost and at any location using satellite receivers, and to do so with an accuracy which exceeds that of the USGS 1:24,000 map series or other readily available and authoritative sources. Population counts from local agencies are increasingly being used as alternatives to those of the Bureau of the Census. Accuracy is now increasingly seen as an attribute of spatial information, and as a component of its value to decisionmakers. In some areas its associated costs can be dramatic, when disputes based on spatial information involve litigation. The complex relationships between accuracy and cost in the GIS area will be a challenging area for research for many years to come.

PRACTITIONER'S SUMMARY

The measurement of data quality in cartographic products is dealt with in numerous publications and standards, many of them referenced in this chapter. Perhaps the most comprehensive view of cartographic data standards is that developed in the US by the Digital Cartographic Data Standards Task Force (DCDSTF 1988), which provides a five-fold model for describing data quality, although it sets no precise numerical thresholds in any of the five categories and defines no standard measurement procedures.

This chapter has taken the view that measurement of data quality for the digital databases now being

developed to support GIS is substantially different from measurement of the accuracy of cartographic features. In fact the two questions "Is this feature correct?" and "What is at this place?" require entirely different strategies. Some of these already exist, particularly in the feature-oriented domain, because they have been developed in support of traditional map-making. Others, particularly in the GIS-oriented domain, are still very limited and inadequate. Despite recent advances, the GISs of 1990 still regard the contents of the database as perfectly accurate, and make no attempt to estimate the uncertainties which are inevitably present in their products.

To summarize the state-of-the-art, we have good techniques for describing and measuring:

- the accuracy of location of a single point;
- the accuracy of a single measured attribute;
- the probability that a point at a randomly chosen location anywhere on a map has been misclassified;
- the effects of digitizing error on measures of length and area;
- the propagation of errors in raster-based area class maps through GIS operations such as overlay; and
- the uncertainty in measures of area derived from dot counting.

However many of these are still in the domain of research literature, and we are in a very early stage of implementing them in standard practice.

From a practical point of view, the coming decade will see a greater awareness of the problems caused by uncertainty in the manipulation and analysis of geographical data, driven largely by the increasing availability of GIS. There will be increasing awareness also of the distinction between accuracy of cartographic features, and accuracy of GIS databases and products. We will see the emergence of systems equipped with tools for tracking uncertainties and the lineage of data through complex series of operations, allowing the user to visualize and interact directly with such information through graphic interfaces. And finally, we will see increasing pressure on the providers of data to supply information on data quality.

REFERENCES

American Society of Photogrammetry (Committee for Specifications and Standards, Professional Practice Division) (1985) Accuracy specification for large-scale line maps. Photogrammetric Engineering and Remote Sensing **51(2)**: 195-199.

Amrhein C and Griffith D A (1987) GIS, spatial statistics and statistical quality control. Paper presented at IGIS '87, Washington DC.

Arbia G and Haining R P (1990) Error propagation through map operations. Unpublished MS.

Aronoff S (1982) Classification accuracy: a user approach. Photogrammetric Engineering and Remote Sensing **48(8)**: 1299-1307.

Aronoff S (1985) The minimum accuracy value as an index of classification accuracy. Photogrammetric Engineering and Remote Sensing **51(1)**: 99-111.

Barrett J P and Philbrook J S (1970) Dot grid area estimates: precision by repeated trials. Journal of Forestry **68(3)**: 149-151.

Bauer M E, Hixson M M, Davis B J and Etheridge J B (1978) Area estimation of crops by digital analysis of Landsat data. Photogrammetric Engineering and Remote Sensing **44(8)**: 1033-1043.

Baugh I D H and Boreham J R (1976) Measuring the coastline from maps: a study of the Scottish mainland. The Cartographic Journal **13(2)**: 167-171.

Beard M K (1989) Use error: the neglected error component. In: Proceedings, AutoCarto 9. Falls Church, VA, ASPRS/ACSM: 808-817.

Bedard Y (1987) Uncertainties in land information systems databases. In: Proceedings, AutoCarto 8. Falls Church, VA, ASPRS/ACSM: 175-184.

Blakemore M (1984) Generalization and error in spatial databases. Cartographica **21**: 131-139.

Blakemore M (1985) High or low resolution? Conflicts of accuracy, cost, quality and application in computer mapping. Computers and Geosciences **11(3)**: 345-348.

Blakney W G G (1968) Accuracy standards for topographic mapping. Photogrammetric Engineering and Remote Sensing **34(10)**: 1040-1042.

Bracken I and Martin D (1989) The generation of spatial population distributions from census centroid data.

Environment and Planning A 21(8): 537-544.

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

Campbell J B (1977) Variation of selected properties across a soil boundary. Soil Science Society of America Journal 41(3): 578-582.

Campbell J B and Edmonds W J (1984) The missing geographic dimension to soil taxonomy. Annals of the Association of American Geographers 74(1): 83-97.

Card D H (1982) Using known map category marginal frequencies to improve estimates of thematic map accuracy. Photogrammetric Engineering and Remote Sensing 48(3): 431-439.

Carter J R (1989) Relative errors identified in USGS gridded DEMs. In: Proceedings, AutoCarto 9. Falls Church, VA, ASPRS/ACSM: 255-265.

Chrisman N R (1982) Methods of Spatial Analysis Based on Errors on Categorical Maps. Unpublished Ph.D. Dissertation, Department of Geography, University of Bristol.

Chrisman N R (1983) The role of quality information in the long-term functioning of a geographic information system. Cartographica 21(2/3): 79-87.

Chrisman N R (1987) The accuracy of map overlays: a reassessment. Landscape and Urban Planning 14: 427-439.

Chrisman N R (1989) Error in categorical maps: testing versus simulation. In: Proceedings, AutoCarto 9. Falls Church, VA, ASPRS/ACSM: 521-529.

Chrisman N R and Yandell B (1988) A model for the variance in area. Surveying and Mapping 48: 241-246.

Congalton R G (1988a) Using spatial autocorrelation analysis to explore the errors in maps generated from remotely sensed data. Photogrammetric Engineering and Remote Sensing 54(5): 587-592.

Congalton R G (1988b) A comparison of sampling schemes used in generating error matrices for assessing the accuracy of maps generated from remotely sensed data. Photogrammetric Engineering and Remote Sensing 54(5): 593-600.

Congalton R C, Oderwald R G and Mead R A (1983) Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques. Photogrammetric Engineering and Remote Sensing 49(12):

1671-1678.

Crapper P F (1980) Errors incurred in estimating an area of uniform land cover using Landsat. Photogrammetric Engineering and Remote sensing 46(10): 1295-1301.

Crapper P F, Walker P A and Manninga P M (1986) Theoretical prediction of the effect of aggregation on grid cell data sets. Geoprocessing 3(2): 155-166.

Curran P J and Williamson H D (1985) The accuracy of ground data used in remote sensing investigations. International Journal of Remote Sensing 6(10): 1637-1651.

Dahlberg R E (1986) Combining data from different sources. Surveying and Mapping 46(2): 141-146.

Digital Cartographic Data Standards Task Force (1988) The proposed standard for digital cartographic data. The American Cartographer 15(1).

Dutton G (1984) Geodesic modeling of planetary relief. Cartographica 21: 188-207.

Dutton G (1989) Modeling locational uncertainty via hierarchical tesselation. In: Goodchild M F and Gopal S (eds.) Accuracy of Spatial Databases. New York, Taylor and Francis: 125-140.

Flowerdew R and Green M (1988) Statistical methods for inference between incompatible zoning systems. In: Goodchild M F and Gopal S (eds.) Accuracy of Spatial Databases. New York, Taylor and Francis: 239-248.

Franklin W R (1984) Cartographic errors symptomatic of underlying algebra problems. In: Proceedings, International Symposium on Spatial Data Handling, Zurich. University of Zurich: 190-208.

Franklin W R and Wu P Y F (1987) A polygon overlay system in PROLOG. In: Proceedings, AutoCarto 8. Falls Church, VA, ASPRS/ACSM: 97-106.

Frolov Y S and Maling D H (1969) The accuracy of area measurements by point counting techniques. Cartographic Journal 6: 21-35.

van Genderen J L and Lock B F (1977) Testing land-use map accuracy. Photogrammetric Engineering and Remote Sensing 43(9): 1135-1137.

van Genderen J L, Lock B F and Vass P A (1978) Remote sensing: statistical testing of thematic map accuracy. Remote Sensing of Environment 7(1): 3-14.

Goodchild M F (1979) Effects of generalization in geographical data encoding. In: Freeman H and Pieroni G G (eds.) Map Data Processing. New York, Academic Press: 191-206.

Goodchild M F (1980) Fractals and the accuracy of geographical measures. Mathematical Geology 12: 85-98.

Goodchild M F (1989) Modeling error in objects and fields. In: Goodchild M F and Gopal S (eds.) Accuracy of Spatial Databases. New York, Taylor and Francis: 107-114.

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

Goodchild M F and Lam N S (1980) Areal interpolation: a variant of the traditional spatial problem. Geoprocessing 1: 297-312.

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

Goodchild M F and Wang M H (1989) Modeling errors for remotely sensed data input to GIS. In: Proceedings, AutoCarto 9. Falls Church, VA, ASPRS/ACSM: 530-537.

Goodchild M F and Yang S (1989) A hierarchical spatial data structure for global geographic information systems. Technical Report 89-5, National Center for Geographic Information and Analysis, University of California, Santa Barbara CA.

Greenland A, Socher R M and Thompson M R (1985) Statistical evaluation of accuracy for digital cartographic data bases. In: Proceedings, AutoCarto 7. Falls Church, VA, ASPRS/ACSM: 212-221.

Griffith D A (1989) Distance calculations and errors in geographic databases. In: Goodchild M F and Gopal S (eds.) Accuracy of Spatial Databases. New York, Taylor and Francis: 81-90.

Haining R P, Griffith D A and Bennett R J (1983) Simulating two-dimensional autocorrelated surfaces. Geographical Analysis 15: 247-53.

Hakanson L (1978) The length of closed geomorphic lines. Mathematical Geology 10(2): 141-167.

Hay A M (1979) Sampling designs to test land-use map accuracy. Photogrammetric Engineering and Remote Sensing 45(4): 529-533.

Hay A M (1988) The derivation of global estimates from a confusion matrix. International Journal of Remote Sensing 9(8): 1395-1398.

Heuvelink G B M, Burrough P A and Stein A (1989) Propagation of errors in spatial modelling with GIS.

International Journal of Geographical Information Systems 3(4): 303-322.

Holdridge L R, Grenke W C, Hathaway W H, Liang T and Tosi J A Jr. (1971) Forest Environments in Tropical Life Zones: A Pilot Study. Oxford, Pergamon.

Honeycutt D M (1986) Epsilon, generalization and probability in spatial data bases. Unpublished MS.

Hudson D and Ramm C W (1987) Correct formulation of the kappa coefficient of agreement. Photogrammetric Engineering and Remote Sensing 53(4): 421-422.

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

Lee Y C (1985) Comparison of planimetric and height accuracy of digital maps. Surveying and Mapping 45(4): 333-340.

Lloyd P R (1976) Quantisation error in area measurement. The Cartographic Journal 15(1): 22-25.

Lodwick W A and Monson W (1988) Sensitivity analysis in geographic information systems. Department of Mathematics, University of Colorado at Denver, Paper RP886.

Lodwick W A (1989) Developing confidence limits on errors of suitability analyses in geographical information systems. In: Goodchild M F and Gopal S (eds.) Accuracy of Spatial Databases. New York, Taylor and Francis: 69-78.

MacDougall E B (1975) The accuracy of map overlays. Landscape Planning 2(1): 23-30.

MacEachren A M (1985) Accuracy of thematic maps: implications of choropleth symbolization. Cartographica 22(1): 38-58.

Maling D H (1968) How long is a piece of string? The Cartographic Journal 5(2): 147-156.

Maling D H (1989) Measurement from Maps: Principles and Methods of Cartometry. New York, Pergamon

Mandelbrot B B (1967) How long is the coast of Britain? Science 156: 636-638.

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

Mark D M and F Csillag (1989) The nature of boundaries on area-class maps. Cartographica 21: 65-78.

Mead R A and Szajgin J (1982) Landsat classification accuracy assessment procedures. Photogrammetric Engineering and Remote Sensing 48(1): 139-141.

Merchant D C (1987) Spatial accuracy standards for large scale topographic maps. Photogrammetric Engineering and Remote Sensing 53(7): 958-961.

Moellering H (ed.) (1987) A Draft Proposed Standard for Digital Cartographic Data. Report No. 8, National Committee for Digital Cartographic Data Standards, Columbus.

Muller J-C (1977) Map gridding and cartographic errors: a recurrent argument. Canadian Cartographer 14(2): 152-167.

Muller J-C (1987) The concept of error in cartography. Cartographica 24(2): 1-15.

Newcomer J A and Szajgin J (1984) Accumulation of thematic map errors in digital overlay analysis. The American Cartographer 11(1): 58-62.

Openshaw S (1977a) A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling. Institute of British Geographers, Transactions 2(NS): 459-472.

Openshaw S (1977b) Optimal zoning systems for spatial interaction models. Environment and Planning A 9: 169-184.

Openshaw S (1978) An empirical study of some zone-design criteria. Environment and Planning A 10: 781-794.

Openshaw S and Taylor P J (1979) A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In: Wrigley N (ed.) Statistical Applications in the Spatial Sciences. London, Pion.

Openshaw S and Taylor P J (1981) The modifiable areal unit problem. In: Wrigley N and Bennett R J (eds.) Quantitative Geography: A British View. London, Routledge.

Perkal J (1956) On epsilon length. Bulletin de l'Academie Polonaise des Sciences 4: 399-403.

Perkal J (1966) On the Length of Empirical Curves. Discussion Paper No. 10, Michigan Inter-University Community of Mathematical Geographers, Ann Arbor.

Robinson J E (1973) Frequency analysis, sampling, and errors in spatial data. In: Davis J C and McCullagh M J (eds.) Display and Analysis of Spatial Data. London, Wiley: 78-95.

Rosenfield G H (1971) On map accuracy specifications: Part II. Horizontal accuracy of topographic maps. Surveying and Mapping 31(1): 60-64.

Rosenfield G H (1986) Analysis of thematic map classification error matrices. Photogrammetric Engineering and Remote Sensing 52(5): 681-686.

Rosenfield G H and Fitzpatrick-Lins K (1986) A coefficient of agreement as a measure of thematic map

accuracy. Photogrammetric Engineering and Remote Sensing 52(2): 223-227.

Smith G R and Honeycutt D M (1987) Geographic data uncertainty, decision making and the value of information. In: Proceedings, GIS '87, San Francisco: 300-312.

Stoms D (1987) Reasoning with uncertainty in intelligent geographic information systems. In: Proceedings, GIS '87, San Francisco: 693-700.

Story M and Congalton R G (1986) Accuracy assessment: a user's perspective. Photogrammetric Engineering and Remote Sensing 52(3): 397-399.

Switzer P (1975) Estimation of the accuracy of qualitative maps. In: Davis J C and McCullagh M J (eds.) Display and Analysis of Spatial Data. London, Wiley: 1-13.

Thompson M M (1971) On map accuracy specifications: Part I. Practical interpretation of map-accuracy specifications. Surveying and Mapping 31(1): 57-60.

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

Veregin H (1989a) Accuracy of spatial databases: annotated bibliography. Technical Report 89-9, National Center for Geographic Information and Analysis, University of California, Santa Barbara.

Veregin H (1989b) A taxonomy of error in spatial databases. Technical Report 89-12, National Center for Geographic Information and Analysis, University of California, Santa Barbara.

Veregin H (1989c) Error modeling for the map overlay operation. In: Goodchild M F and Gopal S (eds.) Accuracy of Spatial Databases. New York, Taylor and Francis: 3-18.

Veregin H (1989d) A review of error models for vector to raster conversion. The Operational Geographer 7(1): 11-15.

Vitek J D and Richards D G (1978) Incorporating inherent map error into flood-hazard analysis. Professional Geographer 30(2): 168-173.

Walsh S J, Lightfoot D R and Butler D R (1987) Recognition and assessment of error in geographic information systems. Photogrammetric Engineering and Remote Sensing 53(10): 1423-1430.

Webster R and Beckett P H T (1968) Quality and usefulness of soil maps. Nature 219: 680-682.

Wehde M (1982) Grid cell size in relation to errors in maps and inventories produced by computerized map processing. Photogrammetric Engineering and Remote Sensing 48(8): 1289-1298.

Yoeli P (1984) Error bands of topographical contours with computer and plotter (program KOPPE). Geoprocessing 2(3): 287-297.