9

Data Models and Data Quality: Problems and Prospects

MICHAEL F. GOODCHILD

GIS can bring enormous benefits because of its ability to handle spatial data in ways that are precise, rapid, and sophisticated. The previous two chapters have provided perspectives on the current state of GIS functionality and on the range of data types currently recognized by GIS. Many of these have direct relevance to the needs of environmental modeling and policy formulation, as later chapters in this book demonstrate. However, much of the data that have found their way into spatial databases over the past two decades have originated from map documents, and the uses to which the data are put in a GIS may differ substantially from the traditional uses of maps. Indeed, GIS may be leading to uses for which the data were never intended when they were collected and mapped.

Like all measurements, spatial data have levels of accuracy that are limited by the measuring devices used. For example, the spatial resolution of an imaging sensor is limited by its pixel size. Similarly, the spatial resolution of mapped data is determined by the inherent limitations of mapmaking technology. It is hard to draw a line of less than 0.2 mm width, and to prevent paper documents from stretching and shrinking with changes in humidity. However, such limitations have a way of disappearing in the digital domain. It is easy to print single-precision real numbers with seven or eight significant digits, even though these may be entirely unjustified by the quality of the input data. Spatial databases have no explicit scale, and rarely inform the user of their accuracies in any useful manner.

So while digital processing of geographic data brings immense benefits in the form of rapid, precise and sophisticated analysis, at the same time it reveals in data weaknesses that 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 also 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 that 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, and draws heavily on arguments presented by Goodchild (1991). An annotated bibliography and taxonomy of the literature in spatial data error and uncertainty have been provided by Veregin (1989a,b).

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. *Accuracy* refers to the relationship between a measurement and the reality 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 that 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 that can be usefully shown on a paper map. This limit is often assumed to be 0.5 mm as a rule of thumb; so the effective resolution of a 1:1000 map is about 1000 x 0.5 mm, or 50 cm, although the standards of most mapping agencies are substantially better. The effective resolutions of some common map scales, assuming 0.5 mm resolution, are shown in Table 9-1.

As noted earlier, because digital spatial data sets are not maps, and have no similar physical existence, they do not have an obvious or explicit 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 are ever resolution free. The term *scale* is often used loosely as a surrogate for database resolution.

Table 9-1: Scale and resolution for some common map scales

Scale	Effective resolution	Minimum resolvable area
1:1,250	62.5 cm	
1:10,000	5 m	
1:24,000	12 m	
1:50,000	25 m	0.0625 ha
1:100,000	50 m	0.25 ha
1:250,000	125 m	1.56 ha
1:500,000	250 m	6.25 ha
1:1,000,000	500 m	25 ha
1:10,000,000	5 km	2500 ha

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 0.5 mm can be assumed to be typical for the resolution of an input map document. 0.5 mm 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 0.25 mm 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 100 cm by 100 cm, then a precision of 0.5 mm 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 GIS operate at double 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. Double precision of 15 decimal digits in latitude and longitude is sufficient to resolve location to the size of a single atom on the surface of the Earth; single precision of 7 decimal digits resolves to about a meter.

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 meters. Processing of data taken from large-scale maps, perhaps 1:1000, 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 on both 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 9-2, 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.

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. A suitable set of rules to define spurious polygons might be:

Table 9-2: Polygons by area for five CGIS layers and overlays

Acres	1	2	3	4	5	1+2+5	1+2+3+5	1+2+3+4+5
0–1	0	0	0	1	2	2640	27566	77346
1–5	0	165	182	131	31	2195	7521	7330
5–10	5	498	515	408	10	1421	2108	2201
10–25	1	784	775	688	38	1590	2106	2129
25–50	4	353	373	382	61	801	853	827
50–100	9	238	249	232	64	462	462	413
100–200	12	155	152	158	72	248	208	197
200–500	21	71	83	89	92	133	105	99
500–1000	9	32	31	33	56	39	34	34
1000–5000	19	25	27	21	50	27	24	22
>5000	8	6	7	6	11	2	1	1
Totals	88	2327	2394	2149	487	9558	39188	90599

- Small in area;
- Long and thin in shape (high ratio of perimeter to square root of area);
- Composed of two arcs only (an arc is the common boundary between two polygons, or a topological edge of the boundary network);
- 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. So the detection of one spurious polygon provides additional evidence that nearby polygons are also spurious.

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 each other 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.

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. In a raster representation, the pixel size has the same effect of establishing an upper limit to the spatial resolution of GIS processing operations; thus there is no equivalent of the spurious polygon problem in raster-based systems. But in vector-based systems, the ambiguity that arises when the processor has a level of spatial resolution that is inconsistent with the positional accuracy of the data leads to problems where the known topological properties of the data conflict with geometric properties.

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 it purports to represent. In most cases reality is not the source document but the ground truth that 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 that 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), and the reverse is also sometimes observed to be true. All of these errors result from representing spatial variation using objects (Chapter 2); 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 that 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 from 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 the continuous variation of fields, 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 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.

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 no longer be 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 that follow roads, rivers or shorelines and again are subject to error processes that 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, or to entire datasets. 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. Moreover there seems little doubt that if such a system were available, it would reveal frightening levels of uncertainty in many GIS analyses and much GIS-based modeling. In Chapter 43, Englund shows the consequences of simulating actual levels of uncertainty in analyses based on kriging, using conditional simulation techniques. Very few techniques exist for displaying known levels of uncertainty on maps (Beard et al., 1991); so such information is commonly hidden from the user. Finally, it seems clear that the nonlinear characteristics of many of the models described in this book, and the ability of many of them to propagate between scales, must lead to high and perhaps disastrous sensitivity to uncertainty in spatial data inputs.

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 in GIS.

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 spatial data. This is not a simple and straightforward extension; as we will see, spatial data require a somewhat different and more elaborate approach.

The objective is to find an appropriate model of the errors or uncertainties that occur in spatial 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. For simple scalar measurements, conventional error theory is based on the Gaussian distribution. In effect, what we seek is the equivalent of the Gaussian distribution for the ensembles of measurements referred to here as spatial databases.

Points

The closest analogy between the classical theory of measurement and the problem of error in spatial 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 100 m, or 2 mm on a 1:50,000 sheet.

If the errors in both coordinates are represented by Gaussian or normal distributions, then we can represent accuracy in the form of a set of ellipses centered on the point. If the accuracies are the same in each coordinate and if the errors are independent, then the ellipses become circles, again centered 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 can be visualized as a probability density forming a bell-shaped surface centered over the point. 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. A commonly used summary measure is the Circular Map Accuracy Standard, defined as the distance within which the true location of the point lies 90% of the time, and equal to 2.146 times the standard deviation of the circular normal distribution.

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 = 0.5 mm (e.g., 12.5 m at 1:25,000)
B: CMAS = 1.0 mm (e.g., 25 m at 1:25,000)
C: CMAS determined but greater than 1.0 mm
D: CMAS not determined.

Many other mapping programs use CMAS as the basis for standards of horizontal or planimetric accuracy, and use similar values. Note, however, that CMAS is meaningful only for well-defined point features. Fisher (1991a) reviews many map accuracy standards, and Chrisman (1991) describes the background to the proposed U.S. Federal Spatial Data Transfer Standard (DCDSTF, 1988).

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. 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 that 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 that exist between the true line and the digitized representation of it.

Second, the errors between the true and digitized line are not independent, but instead tend to be highly correlated (Keefer, Smith, and Gregoire, 1988). 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 modeled 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, often 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.

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 that 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.

Consider, for example, the problem of placing confidence limits on the area of a polygon, and assume that uncertainty in the position of the boundary is described by an epsilon band. It is easy to compute an estimate of the polygon's area, and to place upper and lower limits on area by taking the area inside the inner and outer edges of a band of width epsilon on either side of the boundary. However, these limits will greatly overestimate the true uncertainty, because in reality the observed position of the boundary will sometimes be inside the true position, and sometimes outside; the two types of distortion tend to cancel in a calculation of area. In order to estimate confidence limits correctly, we would need information on the pattern of distortions, such as a measure of their spatial dependence.

Alternatives to the Perkal band

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 that fall on the diagonal of the table are correctly classified; pixels that fall off the diagonal in a column are termed "errors of omission," since the true value is omitted from the assigned class; and pixels that fall off the diagonal in a row are termed "errors of commission," 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).

Goodchild, Sun, and Yang (1992) describe a stochastic process that might be used to model error in soil, vegetation, or forest cover maps, and to propagate its effects. The model is based on a raster representation, in which every cell is assumed to have a vector of probabilities of membership in each of n classes. Such information might come from the use of a fuzzy classifier on a remotely sensed image, or from the misclassification matrix described above, or from the legend of a soil map from statements such as "class A is a mixture of 90% sand and 10% clay."

Fisher (1991b) has simulated the uncertainty implied by this model. If we interpret the probabilities in each pixel as indicative of the proportion of times a given pixel would be assigned a given class by the mapping process, then different realizations correspond to the range of uncertainty inherent in the process. However, it is unrealistic to assume that outcomes are independent in each pixel, since pixels are an artifact of the mapping process. Goodchild, Sun, and Yang (1992) introduce a level of spatial dependence between neighboring outcomes, and by varying the level are able to simulate a range of realistic scenarios regarding the spatial distribution of uncertainty. They use the model to propagate error from the database into confidence limits on the measurement of area, and the results of overlay.

Although suitable estimates of the probability vectors required by the model are readily available from fuzzy classifiers and many types of land cover maps, the estimation of the spatial dependence parameter presents more of a problem. While a single value of the parameter may be sufficient in some circumstances, in others it is likely that the parameter varies by class, and even regionally. Moreover, there is no simple intuitive interpretation of the parameter, and thus it seems unreasonable to expect the average GIS user to participate in its estimation. However, the parameter has a simple visual expression in the texture of each realization. Suppose, for example, that a large patch is predominantly class A, but is known to have inclusions of class B. At low values of spatial dependence, class B will appear in the realizations of this process as small, pixel-sized patches; as spatial dependence increases, the inclusions of B become larger and larger; and in the limit the entire patch is uniformly A in most realizations, but sometimes B.

This visual connection to texture provides a simple way of obtaining crude estimates of the value of the spatial dependence parameter(s). It also illustrates a more general point, that the visual nature of GIS may provide an intuitive access to some of the more sophisticated ideas of spatial statistics discussed in Part VI of this book.

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 that 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:

$$1.03(kn)^{1/2}(SD)^{-3/4} \qquad (9\text{-}1)$$

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 ($k \sim 1$ 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:

$$1.96(1 - S/S_0)^{1/2}(S/D)^{1/2} \qquad (9\text{-}2)$$

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 that 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.

PROSPECTS

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, the manner in which spatial variation is represented in GIS makes a significant difference. Descriptors of uncertainty such as the Perkal band are convenient and intuitive, but are unable to satisfy the requirements of error models, and thus do not lead to methods for assessing uncertainty in analyses or models based on GIS. For example, an epsilon band description cannot be used to estimate uncertainty in the area of a polygon, because it contains no information on the spatial dependence of errors around the polygon's boundary.

In general, certain GIS data models are much more suitable than others as bases for the modeling of uncertainty. Digitized contours are highly unsuitable, because it is virtually impossible to connect the information on the accuracy of contour positioning provided by standard descriptions of topographic map quality with the parameter(s) of a suitable error model. Similarly, we have seen above how the Perkal epsilon band description of positional uncertainty in the polygon model fails to provide the basis of an adequate error model. In general, it appears that the raster data models are much better bases for error modeling in GIS. Representation of spatial variation through the use of irregular objects acts in many cases to make error modeling more difficult, simply because of the nature of these data models. But unfortunately choices among data models for spatial data representation are bound in strong traditions, implying that progress on the accuracy issue will necessarily be slow.

Because of interest in error models for spatial data, 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 35) 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 that populate our maps and databases, but from the processes that 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 that currently exist in our understanding of spatial differentiation and the processes that produce it.

It seems unlikely that the ideal accuracy-sensitive GIS that was outlined earlier in this chapter will ever become a reality. Its analog in simple, aspatial measurement, the statistical analysis system that 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 that 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 a variable attribute of spatial information, and as a com-

ponent of its value to decision-makers. 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.

The measurement of data quality in spatial data products is dealt with in numerous publications and standards. Perhaps the most comprehensive view of quality standards is that developed in the U.S. by the Digital Cartographic Data Standards Task Force (DCDSTF, 1988), which provides a fivefold 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, and with certain very limited exceptions, current GIS still regard the contents of the database as perfectly accurate, and make no attempt to estimate the uncertainties that 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

Arbia, G., and Haining, R.P. (1990) Error propagation through map operations. *Unpublished paper*, Sheffield: University of Sheffield.

Beard, M.K., Buttenfield, B.P., and Clapham, S. (1991) Initiative 7 specialist meeting: visualization of spatial data quality. *Technical Paper 91-26*, Santa Barbara, CA: National Center for Geographic Information and Analysis.

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

Bracken, I., and Martin, D. (1989) The generation of spatial population distributions from census centroid data. *Environment and Planning A* 21(8): 537–544.

Chrisman, N.R. (1982) Methods of spatial analysis based on errors on categorical maps. *Unpublished PhD Dissertation*, Bristol: Department of Geography, University of Bristol.

Chrisman, N.R. (1991) The error component in spatial data. In Maguire, D.J., Goodchild, M.F., and Rhind, D.W. (eds.) *Geographical Information Systems: Principles and Applications*, London: Longman, pp. 165–174.

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

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.

DCDSTF (Digital Cartographic Data Standards Task Force) (1988) The proposed standard for digital cartographic data. *The American Cartographer* 15(1): 9–140.

Fisher, P.F. (1991a) Spatial data sources and data problems. In Maguire, D.J., Goodchild, M.F., and Rhind, D.W. (eds.) *Geographical Information Systems: Principles and Applications*, London: Longman, pp. 175–189.

Fisher, P.F. (1991b) Modeling soil map-unit inclusions by Monte Carlo simulation. *International Journal of Geographical Information Systems* 5(2): 193–208.

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

Franklin, W.R., and Wu, P.Y.F. (1987) A polygon overlay system in PROLOG. *Proceedings, AutoCarto 8*, Falls Church, VA: ASPRS/ACSM, pp. 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, pp. 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, pp. 107–114.

Goodchild, M.F. (1991) Issues of quality and uncertainty. In Muller, J.-C. (ed.) *Advances in Cartography*, New York: Elsevier, pp. 113–139.

Goodchild, M.F., Sun, G., and Yang, S. (1992) Development and test of an error model for categorical data. *International Journal of Geographical Information Systems* 6(2): 87–104.

Greenland, A., Socher, R.M., and Thompson, M.R. (1985) Statistical evaluation of accuracy for digital cartographic data bases. *Proceedings, AutoCarto 7*, Falls Church, VA: ASPRS/ACSM, pp. 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, pp. 81–90.

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.

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

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

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, pp. 69–78.

Mark, D.M., and Csillag, F. (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.

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

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*, Ann Arbor, MI: Michigan Inter-University Community of Mathematical Geographers.

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.

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

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

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, pp. 3–18.