NINE

~-Errors and Quality Control

This chapter examines how errors occur in spatial data and the effects that they may have on data analysis and modelling. Errors include blunders and gaffs, but they are also an intrinsic part of the choice of data models and computational models. Statistical uncertainty and spatial variation are critical aspects of any error analysis in spatial data. Methods are presented for estimating errors in the entity domain in vector-raster conversion, digitizing, and polygon overlay.

Spatial data, costs, and the quality of GIS output

The quality of GIS products is often judged by the visual appearance of the end-product on the computer screen, plotter, or video device, and computer cartographers are devising ever more appealing techniques for communicating visual information to people. Quality control by visual appearance is insufficient, however, if the information presented is wrong or is corrupted by errors. Uncertainties and errors are intrinsic to spatial data and need to be addressed properly, not swept away under the carpet of fancy graphics displays. There can be a false lure about the attractive, high-quality cartographic products that cartographers, and now computer graphics specialists, provide for the users of GIS. In the 1980s Chrisman (1984*a*) pointed out that 'we have developed expectations, such as smooth contour lines, which are not always supported by adequate evidence' and Blakemore (1984) drew attention to the naïve claims of some adherents of computer cartography that computer-assisted cartographic products are necessarily accurate to the resolution of the hardware used to make them. He noted that only a few critical authors, such as Boyle (1982), Goodchild (1978), Jenks (1981), arid Poiker (1982), had drawn attention to the problems of errors in geographic information processing but in 1996 even after twenty-five years of development there is still inadequate attention to how errors arise and are propagated. Most studies on errors are, still at the research level (Fisher 1995, Goodchild and Gopal 1989, Heuvelink 1993, Lodwick et al. 1990) though systematic studies of spatial data quality are now being published (Guptill and Morrison 1995).

In a recent study, Wellar and Wilson (1995) conclude that though GIS has had an impact on the qualitative, quantitative, and/or visualization procedures of spatial theorizing, it has had little impact on the process of spatial theorizing, and hence on a better understanding of natural variation and errors. This is surprising given the costs of data acquisition and the investments that are linked to the use of GIS. In the fields of geostatistics and spatial statistics, however, there have been many theoretical and practical studies on how to deal with uncertainty in the spatial variation of attributes that can be treated as continu-

ous fields (e.g. Isaaks and Srivastava 1989, Journel 1996, Deutsch and Journel1992, Cressie 1991) and it is time to link the ideas developed in these areas to provide a sound basis for understanding the role of uncertainty in spatial data and spatial data analysis.

Data accuracy is often grouped according to ***thematic accuracy, positional accuracy,*** and ***temporal accuracy*** (Aalders 1996) but errors in spatial data can occur at various stages in the process from observation to presentation. Errors in perception (improper identification) can occur at the conceptual stage. Errors and approximations in determining the geographical location depend on surveying skills, the provision of hardware (GPS satellites, laser theodolites, etc.) and the choice of map projections and spheroids. Errors in the measurement of attributes, due to variation in the phenomenon in question, the accuracy of the measurement device, or observer bias can occur during the recording of the primary data. For phenomena treated as continuous fields, the density of samples, their support (physical size of the sample), and the completeness of the sampling are all sources of uncertainty.

Errors can creep in when data are stored in the computer because too little computer space is allocated to store the high-precision numbers needed to record data to a given level of accuracy. Some data may be so expensive or difficult to collect that one must make do with a few samples and rely on inexact correlations with other, cheaper to measure attributes, and so inevitably uncertainties arise. The logical or mathematical rules ('models' or interpolations) used to derive new attributes from existing data may be flawed or may involve computational methods that lead to rounding errors. When data that have been measured on different entities, or sampled on different supports are combined, the differences in spatial resolution may be so great that simple comparisons cannot be made.

Finally, in the visual presentation of results, users can obtain erroneous impressions if the semiotic language is not clear, if colours and shading are inappropriate or if displays are too crowded, or if it is just too difficult to get a clear result.

The usual view of errors and uncertainties is that they are bad. This is not necessarily so, however, because it can be very useful to know how errors and uncertainties occur, how they can be managed and possibly reduced, and how knowledge of errors and error propagation can be used to improve our understanding of spatial patterns and processes. Linking a good understanding of spatial uncertainties to numerical methods of modelling and interpolation can provide useful tools for optimizing sampling (and thereby improving value for money) and for identifying the weak and strong parts of spatial analysis. A good understanding of errors and error propagation leads to active quality control.

Many GIS users conduct data analysis using the techniques presented in Chapters 7 and 8 under the implicit assumption that all data are totally error free. By 'error free' is meant not only the absence of factually wrong data caused by faulty survey or input, but also statistical error, meaning free from variation. In other words, the arithmetical operation of adding two maps together by means of a simple overlay implies that both source maps can be treated as perfect, completely deterministic documents with uniform levels of data quality over the whole study area. This view is imposed to a large extent by the absence of information about data quality, the exact concepts embodied in most databases and retrieval languages (though it can be otherwise), a lack of understanding of how errors are propagated, and the absence of GIS tools for error evaluation.

Many field scientists and geographers know from experience that carefully drawn boundaries and contour lines on maps are elegant misrepresentations of changes that are often gradual, vague, or fuzzy (Burrough and Frank 1996). People have been so conditioned to seeing the variation of the earth's surface portrayed either by the stepped functions of choropleth maps (sharp boundaries) or by smoothly varying mathematical surfaces (see Chapter 2) that they find it difficult to conceive that reality is otherwise. Besides the 'structure' that has been modelledby the boundaries or the isolines, there is very often a residual unmapped variation that occurs over distances smaller than those resolvable by the original survey. Moreover, the spatial variation of natural phenomena is not just a local noise function or inaccuracy that can be removed by collecting more data or by increasing the precision of measurement, but is often a fundamental aspect of nature that occurs at all scales, as the proponents of fractals have pointed out (see Mandelbrot 1982, Burrough 1983*a*,*b*, 1984, 1985, 1993*a*, Goodchild 1980, Lam and De Cola 1993).

It is very important to understand the nature of errors in spatial data and the effect they may have on the quality of the analyses made with GIS.

The first part of this chapter explores the sources of errors in spatial data, and the factors affecting their quality, both' with respect to entity-based and continuous field-based models of spatial phenomena. The second examines the factors that affect the quality of spatial data, while the third covers the development and understanding of errors associated with transforming entity and field-based data from one representation to another (vector-raster), byline digitizing, and through polygon overlay. Chapter 10 presents a statistical approach to the understanding of error propagation in numerical modelling in the context of the kinds of spatial analysis presented in Chapters 6 and 7 and shows how a proper understanding of uncertainties can be used for optimizing sampling and spatial analysis.

Sources of errors in spatial data

Box 9.1 shows the main factors governing the errors that may be associated with geographic information processing. The word 'error' is used here in its widest sense to include not only 'faults' but also to include the statistical concept of error meaning 'variation'. The 'errors' include faults that are obvious and easy to check on but there are more subtle sources of error that can often only be detected while working intimately with the data. The most difficult sources of 'errors' are those that can arise as a result of carrying out certain kinds of processing; their detection requires an intimate knowledge of not only the data, but also the data models, the data structures, and the algorithms used. Consequently they are likely to evade most users. Many of these aspects of 'error', or more correctly 'data quality', are being addressed through international agreements (cf. Aalders 1996).

Accuracy of content

The accuracy of content is the problem of whether the attributes attached to the points, lines, and areas of the geographic database are correct or free from bias. We can distinguish between qualitative accuracy, which refers to whether nominal variables or labels are correct (for example, an area on a land use map might be wrongly coded as 'wheat' instead of 'potatoes') and quantitative accuracy which refers to the level of bias in estimating the values assigned (for example a badly calibrated pH meter might consistently estimate all pH values 1 unit high). Ensuring accuracy is a matter of having reliable, documented input and transformation procedures.

Measurement errors

Poor data can result from unreliable, inaccurate, or biased observers or apparatus. The reader should clearly understand the distinction between accuracy and precision. Accuracy is the extent to which an estimated value approaches the true value and is usually estimated by the standard error. In statistical terminology, precision is a measure of the dispersion (usually measured in terms of the standard deviation) of observations about a mean, Precision also refers to the ability of a computer to represent numbers to a certain number of decimal digits.

Field data

The surveyor is a critical factor in the quality of data that are put in to many geographical information systems. Well-designed data collection procedures and standards help reduce observer bias. The human factor is most important in data collection methods relying on intuition such as in soil or geological survey where an interpretation is made in the field, or from aerial photographs or seismographs, of the patterns of variation in the landscape or subsurface. The user should realize that some observers are inherently more perceptive or industrious than others-'the quality of soil surveyors varies from the two minute job of an irresponsible aerial photo interpreter to that of the surveyor whose sampling plan suggests that he is planting onions' (Smyth, quoted in Burrough 1969). Very large differences in the appearance of a map can result from differences in surveyor or from mapping methods as studies by Bie and Beckett (1973) and

BOX 9.1. FACTORS AFFECTING THE QUALITY OF SPATIAL DATA

1. Currency
   - Are data up to date?
   - Time series
2. Completeness
   - Areal coverage—is it partial or complete?
3. Consistency
   - Map scale
   - Standard descriptions?
   - Relevance
4. Accessibility
   - Format
   - Copyright
   - Cost
5. Accuracy and Precision
   - Density of observations
   - Positional accuracy
   - Attribute accuracy—qualitative and quantitative
   - Topological accuracy
   - Lineage—When collected, by whom, how?
6. Sources of errors in data
   - Data entry or output faults
   - Choice of the original data model
   - Natural variation and uncertainty in boundary location and topology
   - Observer bias
   - Processing
     - Numerical errors in the computer
     - Limitations of computer representations of numbers
7. Sources of errors in derived data and in the results of modelling and analysis
   - Problems associated with map overlay
   - Classification and generalization problems
   - Choice of analysis model
   - Misuse of logic
   - Error propagation
   - Method used for interpolation

more recently Legros et al. (1996) for soil survey and Salome et al. (1982) for geomorphology have clearly demonstrated.

In large survey organizations it should be possible to determine and record the qualities of each surveyor, an extra attribute that could be stored with the data themselves. Such procedures might be resisted by the staff as a slur on professional expertise but the best method for improving observer quality is to improve all aspects of the data-gathering process, such as standardizing observational techniques and data recording forms and by developing a joint commitment between survey management and staff to work to the highest possible standards.

The increasing use in many field sciences of automated sampling devices linked to electronic data loggers means that if all is operating properly then the accuracy and the precision of the data are good. Data from electronic sampling devices collectors can be automatically screened for extreme values indicating

malfunctioning. New sampling devices in areas such as geoengineering and pollution science mean that observations can be made in situ of materials that otherwise must be analysed in the laboratory (Rengers 1994).

Laboratory errors

Intuitively, one expects the quality of laboratory determinations to exceed those made in the field. Although determinations carried out within a single laboratory using the same procedure may be reproducible, the same cannot be said of analyses performed in different laboratories. The results of a major worldwide laboratory exchange program carried out by the International Soil Reference and Information Centre in Wageningen (van Reeuwijk 1982, 1984) showed that variation in laboratory results for the same soil samples could easily exceed ±1 1 per cent for clay content, ±20 per cent for cation exchange capacity (±25 per cent for the clay fraction only), ±10 per cent for base saturation, and f0.2 units for pH. The implications for the results of numerical modelling are enormous! Laboratory analyses should be improving in reproducibility thanks to the wider use of automated laboratory equipment, but no amount of laboratory technology will make up for poorly collected or poorly prepared samples.

Locational accuracy

The importance of the locational accuracy of geographic data depends largely on the type of data under consideration. Topographical data are usually surveyed to a very high degree of positional accuracy that is appropriate for the well-defined objects such as roads, houses, land parcel boundaries, and other features that they record. With modem techniques of electronic surveying and GPS the position of an object on the earth's surface can now be recorded to millimetre accuracy. In contrast, the position of soil or vegetation unit boundaries often reflects the judgement of the surveyor about where a dividing line, if any, should be placed. Very often, vegetation types grade into one another over a considerable distance as a result of transitions determined by microclimate, relief, soil, and water regimes. Changes in slope class or groundwater regime are also unlikely to occur always at sharply defined boundaries.

Positional errors can result from poor fieldwork, through distortion or shrinkage of the original paper base map or through poor quality vectorizing after raster scanning (Dunn et al. 1990, Bolstad et *al.* 1990). Local errors can often be corrected by interactive digitizing on a graphics work station, while general positional errors can be corrected by various kinds of transformation, generally known as 'rubber-sheeting' techniques, that have been described in Chapter 4. The combination of modern hardware and software for error detection has greatly improved the quality of digitizing in recent years.

The success of rubber-sheeting methods for correcting geometrical distortion depends largely on the type of data being transformed, and the complexity of the transformations. Many methods work well for simple linear transformations but break down when complex shrinkages must be corrected. The methods do not necessarily work well when the original map consists largely of linked, straightline segments. For example, some years ago, attempts were made at the Netherlands Soil Survey Institute to use rubber-sheeting methods to match a digitized version of an early nineteenth-century topographic map on to a modem 1: 25 000 topographical sheet for the purpose of assessing changes in land use. The road pattern of the area in question was similar to that of a rigid girder structure. When submitted to the rubber-sheeting process, the road lines were not stretched but the structure crumpled at the road intersections, in much the same way that a bridge or crane made from meccano might crumple at the joins!

Natural spatial variation

Many thematic maps, particularly those of natural properties of the landscape such as soil or vegetation, do not take into account local sources of spatial variation or 'impurities' that result from short-range changes in the phenomena mapped. This problem has been the subject of much research, particularly in soil survey, soil physics, and groundwater studies (e.g. Beckett and Webster 1971, Bouma and Bell 1983, Nielsen and Bouma 1985, Burrough 1993b). The problems are as much associated with paradigms of soil classification and mapping that are spatially simplistic as with the natural variation of the soil which was incompletely understood (Burrough et al. 1997).

Cartographic conventions forced soil scientists to map soils as crisply delineated, homogeneous areas. Information about gradual change within boundaries, and boundaries of varying width could not be represented on conventional chorochromatic maps. These maps have been diligently digitized and the digital soil polygon has been presented to GIS users as an entity

that is spatially as well-defined as a cadastral unit. Unfortunately, the truth is often otherwise-these crisp polygons are really crude, but convenient approximations, and a major problem concerns the lack of information about the difference between these models and reality.

Initially, conventional soil series maps at scales of 1: 25 000-l : 50 000 were characterized in terms of the 'impurities' within the units delineated (Soil Survey Staff 1951), which were supposed to be no more than 15-25 per cent. Impurities were defined as observations that did not match the full requirements as specified in the map legend. Many studies (e.g. see Beckett and Webster 1971 or Burrough 1993*b* for a review) have shown that not only was the 15 per cent a wild guess, but that the concept of 'impurity' had little meaning. By varying the legend, the definition of just what was a matching observation, and so the purity, could be manipulated at will. Subsequent work has demonstrated the natural variation of soil and shown its importance for understanding pollution problems or for optimizing soil fertilization in precision agriculture (e.g. Burrough 1993*b*, Goode 1997). There is increasing information on the variability of soil, and other natural phenomena such as water quality or species composition, which as yet may only be available to specialists.

It is important to realize that the unseen spatial variation of phenomena like soil, lithology, or water quality can contribute greatly to the relative and absolute errors of the results produced by modelling and map overlay. More details about how to estimate how these errors propagate through numerical models are given below in this chapter.

Factors affecting the reliability of spatial data

AGE OF DATA

It is rare that all data are collected at the same time for a given project, unless that project is a specific piece of research. Most planners and environmental agencies are forced to use existing published data in the form of maps and reports, filled in as necessary by more recent remote sensing imagery (including aerial photographs) and field studies. Mead (1982) comments that 'with the exception of geological data, the reliability of data decreases with age'. Although this may be broadly true in the sense that geology changes much more slowly than soil, water regimes, vegetation, or land use, it is also possible that old data are unsuitable because they were collected according to systems of standards that are no longer used or acceptable today. Many attempts to capture old data, using handwritten field sheets and out-of-date terminology have had to be abandoned simply because of the enormous costs involved.

AREAL COVERAGE

It is desirable that the whole of a study area, be it an experimental field or a country should have a uniform cover of information. If this is not so the resource data processer must make do with partial levels of information. Though global digital data are increasing in availability (e.g. the Digital Chart of the World on Internet) it is still common, even in developed countries, for there to be no complete cover of certain kinds of thematic information over a study area, except at scales that are too small for the purpose required. For example, many countries still have fragmentary coverage of soil maps at scales of 1 : 25 000-l : 50 000. Moreover, during the 30-40 years the concepts and definitions of thematic classes of soil, vegetation, and geology have changed as have the ways they should be mapped and the surveyors themselves have moved on. Historical facts can lead to inconsequential map units along map sheet boundaries that are difficult to resolve without further survey.

If coverage is not complete, decisions must be made about how the necessary uniformity is to be achieved. Options are to collect more data, to obtain surrogate data from remote sensing, or to generalize detailed data to match less detailed data. Note that it is extremely unwise to 'blow up' generalized or small-scale map data to obtain the necessary coverage.

MAP SCALE AND RESOLUTION

Most geographic resource data have been generated and stored in the form of thematic maps, and only in

recent years with the development of digital information systems has it been possible to have the original field observations available for further processing. Large-scale maps not only show more topological detail (spatial resolution) but usually have more detailed legends (e.g. a soil map of scale 1 : 25 000 and larger usually depicts soilseries legend units, while a soil map of scale 1: 250 000 will only display soil associations--see Vink 1963 for details). It is important that the scale of the source maps matches that required for the study-small-scale maps could have insufficient detail andlarge-scale maps may contain too much information that becomes a burden through the sheer volume of data. Many survey organizations provide their mapped information at a range of scales and the user should choose that which is most appropriate to the task in hand.

Density of observations

Much has been written about the density of observations needed to support a map or interpolation (e.g. Vink 1963, Burrough 1993*b*, Webster and Burgess 1984), yet there are still organizations that produce maps without giving any information whatsoever about the amount of ground truth upon which it is based. This attitude is changing-the Netherlands Winand Staring Institute provides its contract survey clients with maps showing the location and classification of all soil observations; the UK Land Resources Development Centre has published maps showing the density and location of sample points and transects in surveys (see for example the Reconnaissance Soil Survey of Sabah, Acres et al. 1976).

Although the actual density of observations may be a'reasonable general guide to the degree of reliability of the data, it is not an absolute measure, as statistical studies of soil variation have shown. A rough guide to the density of observations needed to resolve a given pattern is given by the 'sampling theorem' originating from electronic signal detection, that specifies that at least two observations per signal element need to be made in order to identity it uniquely. There has also been considerable work in photogrammetry to estimate the densities of observations that need to be made from aerial photographs on a stereoplotter in order to support reliable digital terrain models (Makarovic 1975, Ayeni 1982).

In short, sampling density is only a rough guide to data quality. It is also important to know whether the sampling has been at an optimum density to be able to resolve the spatial patterns of interest and this subject is treated in Chapter 6 and in the next chapter.

Relevance

Not all data used in geographic information processing are directly relevant for the purpose for which they are used, but have been chosen as surrogates because the desired data do not exist or are too expensive to collect. Prime examples are the electronic signals from remote sensors that are used to estimate land use, biomass, or moisture, or observations of soil series based on soil morphology that are used to predict soil fertility, erosion susceptibility, or moisture supply. Provided that the links between the surrogates and the desired variables have been thoroughly established then the surrogates can be a source of good information.

The calibration of surrogates is a major part of remote sensing technology. Briefly, a number of pixels on the image is selected for use as a 'training set'. The variation of reflectance of each frequency band recorded is displayed in the form of a histogram; the practice is to select a training set of pixels that return narrow, unimodal distributions. These training set pixels are calibrated by 'ground-truth' observations so that the set of pixels can be equated with a crop type, a soil unit, or any other definable phenomenon. The remaining pixels in the image are then assigned to the same set as the training set using allocation algorithms based on discriminant analysis (minimum distance in multivariate space of the original frequency bands), maximum likelihood or parallelopiped classifiers (see for example, Estes et al. 1983, Lillesand and Kiefer 1987).

Data format, data exchange, and interoperability

There are three kinds of data format of importance. First there is the purely technical aspects of how data can be written on magnetic media for transfer from one computer system to another. This includes aspects such as the kind of medium (digital tape, floppy disk, compact disk), the density of the written information (tape blocklengths, number of tracks and the density), the type of characters used (ASCII or binary), and the lengths of records. For data lines, it is essential that the speed of transmission of the two computers is matched, but most modems ensure that this is automatic.

The second kind of format concerns the way the data are arranged, or in other words, the structure of the data themselves. Do the data refer to entities in space, recorded as points, lines, and areas in a relational model, as objects in an object orientation system, or as discretized continuous fields coded as rasters? If the areas are coded in raster format, what is the size of each pixel? Is the organization of these data tied to a particular computer system that makes exchange difficult without conversion? For example, many commercial GIS have their own internal data structures (see Chapter 3) that may make direct data exchange difficult. The current moves to system interoperability and the availability of data sources on the Internet are driving people to develop generally acceptable, interchangeable data structures that conform to widely accepted industrial and international standards (Schell 1995*a*).

The third kind of format concerns the locational and attribute data, their scale, projection, and classification. Scale and projection conversions can usually be accomplished quite easily by using appropriate mathematical transformations on the coordinate data (e.g. Maling 1973). Matching classifications from different sources can be very difficult, and the problem is by no means confined to the problems of classifying soil profiles but also occurs in municipal applications of GIS where different administrative divisions may have completely different ways of recording essentially similar entities like roads or services.

To summarize, data exchange often requires that data be reformatted to a lowest common denominator format that can be read by many systems easily. These formats are not necessarily the most compact nor the most efficient but are expedient. There are data formats for satellite data, there are format standards for commercial vendors, there are within-country standards (e.g. in the UK for Ordnance Survey Maps, in the Netherlands and Germany for topographic mapping) and international standards for Geographical Information are now being developed. General standards for the encoding and exchange of spatial information have been set up by standards committees of the European Union (e.g. see Comité European Normalisation CEN Technical Committee 287—David *et al.* 1997, Salgé 1997), by the US Federal Data Standards Committee (National Research Council 1994), and by the recently formed Open GIS Consortium (Schell 1995*a*). Note that interoperability issues are forcing people to think of the conceptual problems of exchanging data as the first step, rather than solely concentrating on technical arguments.

ACCESSIBILITY

Not all data are equally accessible. Data about land resources might be freely available in one country, but the same kind of data could be a state secret in another. Besides the military aspects of data for geographic information systems (here one thinks immediately of digital terrain models) inter-bureau rivalries can also obstruct the free flow of data. Costs and format problems can also seriously hinder data accessibility. In recent years a new kind of middleman, the information broker, has sprung up to assist the seeker of data from digital archives. Details about information services can be obtained from government or international agencies (e.g. EUROGI, Euronet DIANE News, the newsletter of the Directorate General for Information Market and Innovation, Commission of the European Communities, Luxembourg). There is also much information to be found on the Internet and World Wide Web (see Appendix 2).

COSTS AND COPYRIGHTING

Collection and input of new data or conversion and reformatting of old data cost money. For any project, the project manager should be able to assess the costs and benefits of using existing data as compared to initiating new surveys. Digitizing costs may be especially high for inputting detailed hand-drawn maps or for linking attributes to spatial data. Scanners may offer savings for data input of contour lines and photographic images. It may be cheaper for a survey agency to contract out digitizing work to specialist service bureaux than to do the work in house using staff who can be better used for more skilled work Similarly, if an agency only occasionally needs to perform certain kinds of data transformations or to output results to expensive devices such as laser photo plotters of high quality, it may be cheaper to make use of service bureaux.

Copyright on published maps and spatial data varies from country to country and it is best to check on the legal situation in each case when digitizing maps or using spatial data for research or commercial applications (e.g. see Burrough and Masser 1997).

NUMERICAL ERRORS IN THE COMPUTER

As well as the problems inherent in the data, indicated above, there are other sources of unseen error that can originate in the computer. The most easily forgotten, yet critical aspect of computer processing is the abil-

BOX 9.2. ERROR CREATION BY COMPUTER WORD OVERFLOW

There is a simple, and very revealing test of calculation precision that can demonstrate just how computer word length can affect the results of calculation (Gruenberger 1984). The number 1.0000001 is squared 27 times (equivalent to raising 1.0000001 to the 134 217 728th power). The table shows the results of performing this calculation on a Personal computer with a 80486 processor using Microsoft Quick Basic with 4-byte or 8-byte precision. After 27 squarings the single precision result has acquired a cumulative error of more than 1300 per cent! Clearly, the programmer must avoid situations in which the results of a calculation depend on accuracies of representation that exceed the number of digits available for representing the numbers.

No. of squares	Single predsion	Double precision	Single/double per cent difference
1	1	1.0000002000000001	100.0000038418561
2	1	1.0000004000000006	100.0000076837062
3	1.000001	1.0000008000000028	100.0000153673913
4	1.000002	1.0000016000001201	100.0000307334694
5	1.000004	1.0000032000004962	100.0000614690337
6	1.000008	1.0000064000020163	100.0001229365
7	1.000015	1.0000128000081287	100.0002458676304
8	1.000031	1.0000256000326416	100.0004917125829
9	1.000061	1.0000512001308269	100.0009833944561
10	1.000122	1.0001024005237993	100.0019663060907
11	1.000244	1.0002048020962848	100.0039311601034
12	1.000488	1.0004096083877262	100.0078565185848
13	1.000977	1.0008195355595404	100.0157136544184
14	1.001965	1.0016397428294	100.0314297315305
15	1.003913	1.0032821744115347	100.0628687843718
16	1.007841	1.0065751213499587	100.1257171907418
17	1.015744	1.0131934753221909	100.2517048611989
18	1.031735	1.0265610168232249	100.5040404509439
19	1.064478	1.0538275241540432	101.0106167959664
20	1.133113	1.1105524506646417	102.0314517808015
21	1.283946	1.2333267456777186	104.1041728221037
22	1.648814	1.5210094861602679	108.3767906630347
23	2.717598	2.3137229577994073	117.4455787929624
24	7.385337	5.3533445600844633	137.9574445446612
25	54.54321	28.658297978987 73	190.3226945588952
26	2974.962	821.2980430524522	362.2268061013606
27	8850397	674530.4755217875	1312.082590848986

ity of the computer to be able to store and process data at the required level of precision. The precision of the computer word for recording numbers has important consequences for both arithmetical operations and for data storage.

Many people do not appreciate that use of computer variables and arrays having insufficient precision can lead to serious errors in calculations, particularly when results are required that must be obtained by subtracting or multiplying two large numbers. For example, the 'shorthand' method of estimating the variance of a set of numbers involves adding all the numbers together, squaring the result and dividing by the number of numbers. This 'constant' is then subtracted from the sum of the squares of all the numbers to obtain the sum of squared deviations. Box 9.2

explains that when many large numbers are involved there will almost certainly be large rounding errors occurring when the number of bits in the computer word is insufficient to handle the precision required.

Rounding errors Rounding errors are unlikely to be a problem when performing statistical calculations in large computers when the programming language allows double precision variables and arrays to be defined. They used to be troublesome in 16-bit micro computers, particularly if 'shorthand' methods of calculation were used. In the above example, it is much wiser to first calculate the average of the set of numbers, then to calculate the deviation of each number from the average and then sum the squared deviations. This method of estimating the sums of squares is not only closer to the original method of defining variation, but avoids rounding errors in the subtraction process.

In many systems used for image analysis data are coded as integers. The problems of accurately representing the areas and perimeters of polygons in raster format were noted in Chapter 3. Franklin (1984) explored the problem of data precision for other GIS operations, such as scaling and rotation, when the results of arithmetical operations are truncated to the nearest integer. As Figure 9. la shows, scaling a simple triangle by a factor of three results in the point P being moved outside the triangle. Rotating point P (Figure 9. 1 *b*) moves it inside the circle.

The obvious way to avoid&s problem is to increase the precision with which the computer represents numbers, i.e. to work with real numbers with a decimal representation.

As Franklin demonstrated, this merely pushes the problem to another level; it does not go away. The problem is one that is intimately linked with the way the computer represents numbers. It is possible to find real numbers for which computer implementations of simple arithmetic violate important real number axioms of distributivity, associativity, and commutativity.

For example, associativity:

$$(A + B) + C = A + (B + C)$$

This rule is violated in a computer that stores fewer than 10 significant digits for $A = 1.E10$, $B = -1.E10$, C = 1. Franklin showed that these problems can be corrected by using different methods of computation, which themselves bring extra problems of complexity and the need to develop or use special subroutines for arithmetical operations.

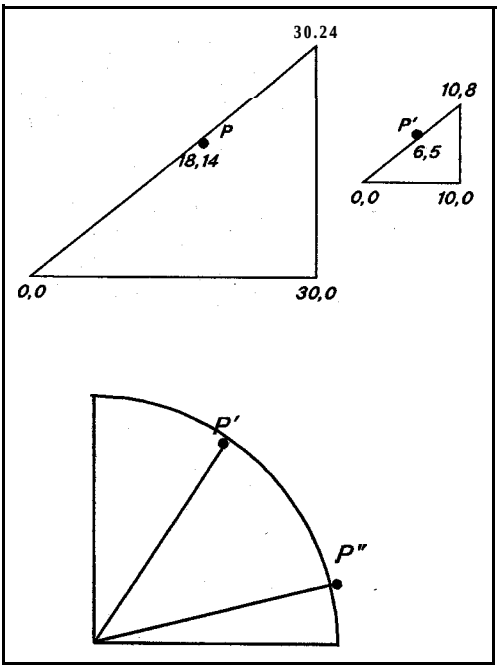

Figure 9.1. With integer arithmetic, scaling or rotation can cause points near boundaries to be rounded off inside or outside a polygon

Geographical coordinates and **precision** Chrisman (1984*b*) examined the role of hardware limitations **on** another problem in geographical information systems, namely that of storing geographical coordinates to the desired level of precision. Whereas 16-bit machines have presented few problems for storing the coordinates of low-resolution, single scene LANDSAT images, the high accuracy required by cadastral systems, or the sheer range of coordinates required to cover a continent result in numbers that are too large to be recorded in a single 16-bit computer word and 32-bit words or even 64 bits are necessary (Table 9.1). Fortunately this is no longer a serious problem. The 32-bit word used in many computers currently used for GIS, allows spatial dimensions to be recorded with the following precision:

Maximum dimension (metres)	Maximum precision attainable
10 000.00	dddd.dx
100 000.0	ddddd.x
1000 000	dddddx

where d means good data, and *x* is the excess precision needed to avoid most of the topological dilem-

Table 9.1. The relation between computer word length and digital range and precision

Variable type	Number of significant digits (decimal)	Approximate decimal range
16-bit integer (2 bytes)	4	$-32768 \leq x \leq +32767$
Short real 32 bits (single precision 4 bytes)	6–7	$-3.37 \times 10^{38} \leq x \leq -8.43 \times 10^{-37}$ through true zero to $8.43 \times 10^{-37} \leq x \leq +3.37 \times 10^{38}$
Long real 64 bits (double precision 8 bytes)	15–16	$-1.67 \times 10^{308} \leq x \leq -4.19 \times 10^{-307}$ through true zero to $4.19 \times 10^{-307} \leq x \leq 1.67 \times 10^{308}$

mas of the kind shown in Figure 9.1. While it is unlikely that a user will require a precision better than 10 m for an area of 1000 x 1000 km, data from sensors like the French satellite SPOT with its 10 m resolution, which may be used to supply data for the resource inventory, mean that the 32-bit floating point representation in the GIS is stretched to the limit. Moreover, it may be necessary in the inventory to refer to ground control points that have been located' with much greater precision.

Chrisman (1984b) and Tomlinson and Boyle (1981) have pointed out that locational precision is critical when the user wishes to interface different kinds of data sets that have been acquired at different scales and to different levels of precision. These problems are greater when working with established inventories that may have been geometrically using old 16-bit systems than when all data must be collected for specific projects, because often in the latter case, the data are collected from scratch.

Through national and international agreements and improvements in hardware and software, information on the quality of digital data is becoming an important part of the data itself.

Faults stemming from assumptions concerning the exactness of spatial entities

As already noted, most procedures commonly used in geographic information processing assume implicitly that (a) the source data 'are uniform, (*b*) digitizing procedures are infallible, (c) map overlay is merely a question of intersecting boundaries and reconnecting a line network, (*d*) boundaries can be sharply defined and drawn, (*e*) all algorithms can be assumed to operate in a fully deterministic way, and (f) class intervals defined for one or other 'natural' reason are necessarily the best for all mapped attributes. These ideas result from the traditional ways in which data were classified and mapped. They have presented large technical difficulties for the designers of geographical information systems but rarely have these problems been looked at as a consequence of the way in which the various aspects

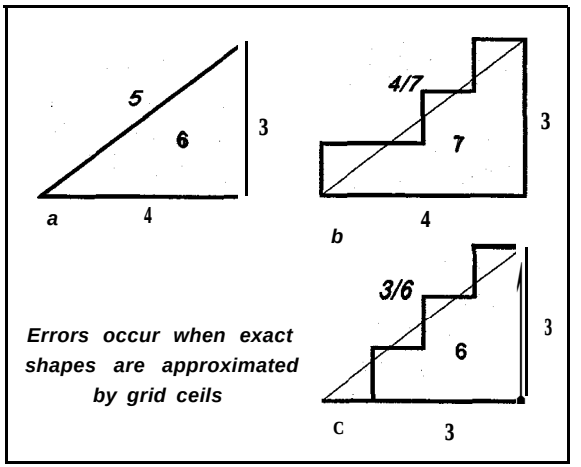

Figure 9.2. Errors occur when exact shapes are approximated by grid cells

of the world have been perceived, recorded, and mapped.

Many operations in geographic information processing require one or more spatial networks to be combined. The spatial networks may be composed of lines, of regular grids, or of irregular polygons. The overlay may be for' the purposes of data conversion, such as converting a vector representation of a polygon net to raster form by overlaying a grid of given resolution, or for the purposes of data combination or modelling, such as when two polygon networks are intersected, or when the boundary of a watershed is used to cut out areas from an overlay of administrative units, or when data from soil polygons are input to a crop yield model.

This section covers the errors that can result from (a) converting spatial entities such as polygons from a vector to a raster representation, and (b) from overlaying and intersecting two polygon networks under the assumptions that spatial entities are exactly defined.

ERRORS RESULTING FROM RASTERIZING A VECTOR MAP

Grid cells are only approximations? As Figure 9.2 shows, converting a vector triangle to unit pixels results in a serious loss of information. The area of the triangle should be 7 units, but could be taken to be 6 or 7 units depending on how the cells and their sides are counted. The hypotenuse could be 7 cell sides long if 4 cells are taken as an approximation of the

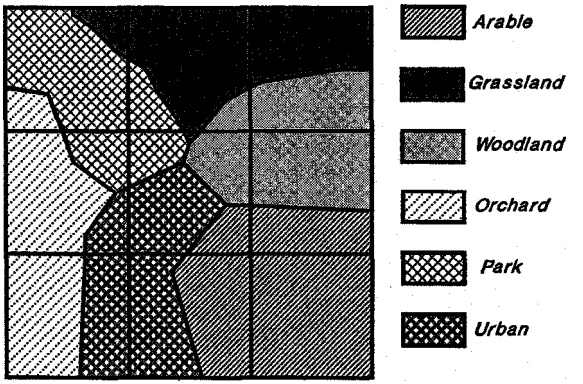

Figure 9.3. The mixed pixel problem occurs when grid cells are too large to resolve spatial details

diagonal, but only 6 if we opt for 3 cells-both are overestimates. Today, approximation errors with rasters are less of a problem because we have much larger and faster computer storage. In cartographic applications such as digital orthophoto maps (Plate 1) and on many laser and ink jet printers the grid cell is much smaller than the finest line drawn by a pen on a vector plotter-in fact, most plotters used today use raster technology. Only when large grid cells are used as basic database entities need we consider the different accuracies of a vector and a raster representation of space.

Mixed pixels Errors can arise in two ways when spatial phenomena are represented by an array of grid cells. The first and most obvious source of error is the problem of 'mixed pixels'; because each grid cell can only contain a single value of an attribute it is only the mean value that is carried in the cell. In the original LANDSAT imagery, in which each cell had a size of some 80 x 80 m, the signature of the pixel was a mean value of the reflectance averaged over the area of the whole cell, for SPOT the cells are 20 x 20 m so the spatial averaging is less. The differences in cell size mean that if part of the LANDSAT cell covered a highly reflecting surface such as water, this could so weight the mean reflectance as to give an over-representation of the area of 'water' compared to SPOT which might record other land cover types within the 80 x 80 m area. These kinds of classification error can occur whenever the size of the grid cell is larger than the features about which information is desired. It is a problem particularly when large-area grid cells are used to record many features in a complex landscape (Figure 9.3). In vector-raster conversion, the mixed

pixel problem leads to the dilemma of whether to classify a cell according to the class covering the geometric reference point of the cell (the centre or the south-west corner) or according to the dominant class occurring in the cell. In remotely sensed and other scanned imagery, the problem is complicated because the cell value is a weighted average of the information reaching the sensor from the area not only covered by the pixel but from nearby surrounding areas.

Vectors to line **rasters** Converting polygons from vector to raster representation when using grid cells smaller than the **polygons** (see Chapter 4) brings with it the problem of topological mismatch when the smooth polygon boundaries are approximated by grid cells. Although high-quality raster scanners and plotters have largely removed the problem of loss of information through rasterizing from the visualization area of GIS, there are still many instances where thematic data, originally in vector polygon form, need to be rasterized to match data on regular grids, such as those obtained by remote sensing, or for some of the analysis examples in Chapter 7. Therefore it is necessary to estimate the seriousness of the problems of mismatch caused. Piwowar et *al.* (1990) examined several algorithms for vector to raster conversion for the quality of the results, the accuracy, the lateral displacement of boundaries, and their speed of operation. They concluded that not all algorithms worked equally **well;** some are fast but cause distortion, while others take more time but produce better results.

Note that in the following discussion of vector to raster conversion, the polygons are regarded as exact entities with precisely located boundaries; the errors of conversions are therefore merely the result of representing a geographic area by one geometry or another. Errors of misidentification or of the inability to define exactly what the area comprises are not treated here.

Statistical approaches to estimating the errors of vector to raster conversion Frolov and Maling (1969) considered the problem of error arising when a grid cell is bisected by a 'true' boundary line. They assumed that the boundary line could be regarded as a straight line drawn randomly across a cell. The mean square area of the cut-off portion of each bisected boundary cell ***i (the*** error variance) can be estimated by

$$V_i = aS^4 \qquad 9.1$$

where V is the error variance, S is the linear dimension of the (square) cell, and a is a constant. Frolov and Maling calculated the value of a as 0.0452 but subsequent work reported by Goodchild (1980) suggests that a better value is $a = 0.0619$.

The error variance in an estimate of area for any given polygon is given by a summation of all the errors from all the bounding cells. If m cells are intersected by the boundary, the error variance will be

$$V = maS^4 \qquad 9.2$$

with standard error

$$\mathrm{SE} = (ma)^{1/2}S^2 \qquad 9.3$$

assuming that the contributions of each cell are independent. Goodchild (1980) suggests that this assumption should not always be regarded as valid.

The number of boundary cells m, can be estimated from the perimeter of the polygon. Frolov and Maling (1969) showed that m is proportional to $\sqrt{N}$, where N is the total number of cells in the polygon. The standard error of m is estimated by $kN^{1/4}a^{1/2}S^2$. Because the estimate of polygon area $A = N.S^2$, the standard error as a percentage of the estimate is proportional to $N^{-3/4}$ (Goodchild 1980), i.e.

$$SE = ka^{1/2}N^{-3/4} \qquad 9.5.$$

If the variable is cell side S instead of cell number N, the percentage error depends on $S^{3/2}$. Goodchild (1980) reports studies that have verified these relationships empirically.

The constant k depends on the polygon shape, long thin shapes having more boundary cells than a circular form of the same area. Frolov and Maling (1969) give values of k for various standard shapes, using the independent straight line hypothesis.

Switzer's **method** Switzer (1975) presented a general solution to the problem of estimating the precision of a raster image that had been made from a vector polygon map. His analysis does not deal with observational or location errors, but assumes that error is solely a result of using a series of points located at the centres of grid cells to estimate an approximate grid version of the original map. Switzer's method deals essentially with ideal choropleth maps, i.e. thematic maps on which homogeneous map units are separated by infinitely thin, sharp boundaries. The method assumes that a 'true' map exists, against which an estimated map obtained by sampling can be compared. Realizing that the 'true' map is often unknown or unknowable, Switzer showed that by applying certain assumptions and by using certain summary

BOX 9.3. SWITZER'S METHOD

The $P_{ij}(n^{-1/2})$ and $P_{ij}(2n^{-1/2})$ probabilities are estimated from a frequency count as follows:

1. Estimate the total number of cell pairs at distance d = 1 cell width. The total number of pairs at a given distance is equal to

$$NPAIRS = 4*(P*Q) - 2*d*(P+Q)$$

where P = number of rows, Q = number of columns in the grid, and the second term is a correction for the cells on the edge of the grid.

2. For each pair of mapping units i and j, count the number of cell pairs along grid up and down the grid that lie in different mapping units ($TALLY_{ij}$).

3. Compute $P_{ij}(n^{-1/2})$ as $TALLY_{ij}/NPAIRS$.

4. Repeat steps 1–3c with d = 2 cell widths to estimate $P_{ij}(2n^{-1/2})$.

5. Calculate O_{ij} from equation (9.5).

6. Calculate total mismatch for each mapping unit O_i as the sum of the O_{ij}s, remembering that the mismatch of $O_{ii} = O_{ii}$.

7. Calculate total mismatch as the sum of the O_is.

statistics, errors of mismatch could be estimated from the estimated or gridded map itself.

The analysis begins by assuming that a map M has been partitioned into k homogeneous map units, or **colours.** Each of the k map units may be represented on the map by one or more sub-areas. This 'true' map is estimated by laying an array of n basic sampling cells over the map. Here we shall only consider the situation where the array of sampling cells is regular and congruent, and each cell is defined by a single sampling point at the cell midpoint. The map units on the 'true' map are denoted Ml, M2,. . . Mk, and on the estimated map by Ml, M2, . . . Mk. Each cell on the estimated map is allocated to a map unit Mi if the sampling point in the cell falls within map unit Mi on the 'true' map. This is the procedure commonly used when converting a vector polygon network to raster format. For the purposes of this analysis we shall, like Switzer, assume the total area of the map is scaled to unity, i.e. A(M) = 1.

The degree of mismatch of the estimated map is a function of two independent factors, (a) the complexity of the true map, and (b) the geometrical properties of the sampling net. Considering first the complexity of the map, we can define a quantity $P_{ij}(d)$ as the probability that a random point is in true map unit i and that the cell centre point is in true map unit j when the points are separated by distance d. Switzer derived the following expression for the percentage overlap O_{ij} for each pair of rasterized map units i,j using square grid cells,

$$0, = 0.60P_{ij}(n^{-1/2}) - 0.11P_{ij}(2n^{-1/2}) \qquad 9.5$$

(Note that the values of the coefficients differ from Switzer's published formula; the corrections are given by Goodchild (1980)).

The total error for all k map units is given by

$$o = \sum_{i=j}^{k} O_{ij} \qquad 9.6$$

Box 9.3 shows how 0 can be estimated in practice.

An example. Figure **9.4***a* shows the boundaries of a simple thematic map depicting soil or geological units. Assuming that they form a true map, what will be the relative mismatch errors arising from digitizing it using grid rasters of different sizes, as shown in Figures **9.4***a,b* for 16 x 16 and 32 x 32 grids, respectively?

Table 9.2 gives the results. For a grid measuring 16 x 16 cells, there are 960 cell pairs at distance d = 1. The total number of cell pairs straddling a boundary lead to frequency estimates for each mapping unit. For a distance d = 2, the number of pairs is 896. Entering the frequency estimates into equation (9.6) leads to an estimated mismatch of 9.5 per cent. Using the 32 x 32 cell grid leads to an estimate of 4.1 per cent,

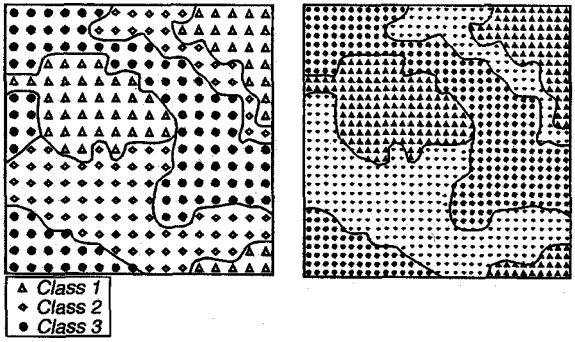

Figure 9.4. Rasterizing a vector map at two grid sizes to estimate rasterizing errors

demonstrating that a factor **4** increase in the number of grid cells is needed to reduce the estimation error by half. Both these estimates of mismatch compare favourably with the estimates of mismatch obtained by measuring the areas of the mapping units on the original map. Note that this means that mismatches with a printer of 600 dpi are approximately half those of one using 300 dpi.

Bregt et ***al.*'s method** Bregt et ***al.*** (**199** 1) developed an elegant method for estimating the error associated with vector-raster conversion called the ***double-conversion method,*** because it involves rasterizing the map twice. First the vector to raster conversion is carried out using the desired target raster size, this produces what they call the base raster. The map is then rasterized to a very much smaller grid and the two are compared. Those cells in the fine raster differing from those on the base raster provide an estimate of the error in the base raster.

Bregt ***et al.*** compared the errors so obtained with a parameter called the ***boundary index*** (*BI*), which is defined as the boundary length in centimetres per square centimetre of the map. The ***BI*** is calculated by dividing the total length of the polygon boundaries by their total area. They found that for a given cell size the rasterizing error (as a percentage mismatch) is a linear function of ***BI.*** They distinguish two situations, (a) that where the cell on the base raster is classified according to the polygon in which its central point falls, and (b) cell classification by the polygon that dominates its area. Table 9.3 presents the results.

Bregt ***et*** al. compared their method with Switzer's and demonstrated that it provides easier and better estimates of the rasterizing error since only the ***BI*** needs to be computed. *BI* values are independent of the units used. The disadvantage is that the regression equations need to be worked out for all possible situations, whereas Switzer's method is completely general and requires no previous work

ERRORS ASSOCIATED WITH DIGITIZING A MAP, OR WITH GEOCODING

As already noted, the methods of Switzer, Goodchild, and Bregt et al. to estimate mismatch assume implicitly that a 'true' map exists that has homogeneous (uniform) mapping units, and infinitely sharp boundaries. In practice, however, even the best-drawn maps are not perfect, and extra errors are introduced by the digitizing process as authors such as Blakemore (1984), Bolstad ***et al.*** (1990), Dunn ***et*** al. (1990), and Poiker (1982) have pointed out. Consider the problem of boundary width and location (the problem of within-map unit homogeneity will be dealt with later) on a

Table 9.2. The results of using Switzer's method on the map in Figure 9.3

Estimates/grid size	16 × 16 grid			32 × 32 grid		
Mismatch per polygon pair	L_{12} 0.020	L_{21} 0.020	L_{13} 0.008	L_{12} 0.010	L_{21} 0.010	L_{31} 0.004
	L_{13} 0.008	L_{23} 0.014	L_{32} 0.014	L_{13} 0.004	L_{23} 0.006	L_{32} 0.006
Mismatch per polygon	L_1 0.028	L_2 0.034	L_3 0.022	L_1 0.014	L_2 0.016	L_3 0.011
Total mismatch (%)	8.46			4.11		

Table 9.3. Relations between rasterizing method, cell size, and rasterizing error

Rasterizing method	Raster cell size (mm)	Regression equation	Variance explained (%)
Central	1 × 1	$L = 2.40I$	99.8
Dominant	1 × 1	$L = 2.48I$	99.8
Central	2 × 2	$L = 4.78I$	99.8
Dominant	2 × 2	$L = 4.68I$	99.8
Central	4 × 4	$L = 9.08I$	99.4
Dominant	4 × 4	$L = 8.78I$	99.0

Source: Bregt *et al.* 1991

digital map of polygons in vector format. The digital map will almost certainly have been derived by digitizing a paper version of the map. There are two sources of potential error-(a) errors associated with the source map, and (*b*) errors associated with the digital representation.

(a) Apart from the potentially correctable errors of paper stretch and distortion in the printed map or source document, errors arise with boundary location simply because drawn boundaries are not infinitely thin. A **1** mm line on a **1** : 1250 map covers an area 1.25 metres wide; the same line on a 1 : 100 000 map covers an area 100 m wide. A detailed 1: 25 000 soil or geological map measuring 400 x 600 mm may have as much as 24 000 mm of drawn lines covering an area of 24 000 sq. mm or 10 per cent of the map area! Common sense suggests that the true dividing line should be taken as the midpoint of the drawn line, but it is not being cynical to state that the area of the map covered by boundary lines is simply an area of uncertainty, and possibly, confusion. When these boundary lines are converted by digitizing, extra errors arise because with hand digitizing the operator will not always digitize exactly the middle of the line, and with scanners, errors will arise with the data reduction algorithms used.

(b) The representation of curved shapes depends on the number of vertices used (Aldred 1972: 5). Consequently, the relative error of digitizing straight lines is much less than that resulting from digitizing complex curves. Translating a continuous curved line on a map into a digital image involves a sampling process: only a very small proportion of the infinity of possible points along a curve is sampled (see in Figure 9.5; Smedley and Aldred 1980).

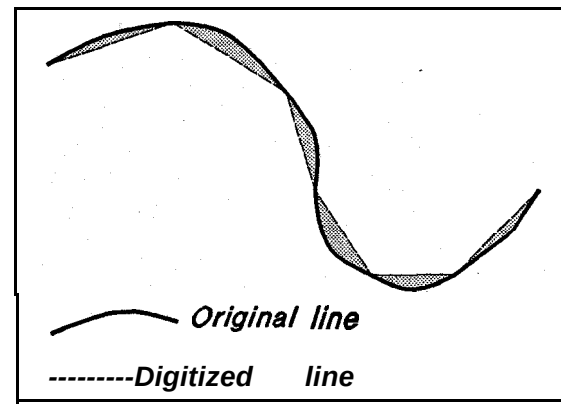

Figure 9.5. Digitizing a line is a sampling process

Clearly, boundaries on thematic maps should not be regarded as absolute, but as having an associated error band or confidence interval. MacDougal(l975) suggested that the total boundary inaccuracy could be estimated by

$$H = \sum_{i=1}^{N} (h_i l_i)/T \qquad 9.7$$

where h_i is the horizontal error (in standard deviations) of line i, length l_i, N is the number of boundary lines, and T is the total area of the map. If all boundary lines are the same type (e.g. they are all soil boundaries or all land use boundaries) equation (9.7) simplifies to

$$H = (hL)/T \qquad 9.8$$

The total line length L was originally estimated by placing a grid over the map and counting the number of crossings, K, and using the formula

$$L = (TK)/0.6366 \qquad 9.9$$

where 0.6366 is a constant described by Wentworth (1930), but today the total length can easily be computed from the database.

In an empirical study, Bolstad et al. (1990) report that the errors due to the manual digitizing of 1 : 25 000-1 : 50 000 soil maps were quite small, and for the United States, of less importance than positional errors due to uncertainty in georegistration.

Error bands around a digitized line: problems for point-in-polygon searches and when combining raster and vector databases

Perkal (1966) suggested that an 'epsilon' distance should be defined around a cartographic line as a means of generalizing it objectively. Blakemore (1984) reversed the concept to indicate the possible confusion associated with width of an error band about a digitized version of a polygon boundary in relation to an application of the well-known 'point-in-polygon' problem. He showed that the question 'Does point P lie within polygon A?' returns at least five possible answers, illustrated in Figure 9.6.

1. Definitely outside the target polygon A.
2. Probably outside A, but could be inside. Variants are $2'$ in which the point is probably inside a neighbour B, but could be in A, and 2" in which the point is probably outside A but could be in either of two neighbours B or C.
3. On the boundary-indefinite.
4. Probably inside A, but could be outside; other variants are $4'$ where the point is probably in A but could be in a neighbour C, and 4" in which the probably 'in' point could also be in one of two neighbours B or C.
5. Definitely in A.

'Definitely in' records the core area within the error band; 'possibly in' records a point that falls within the overlap of the inner half of the confidence band and the polygon. 'Possibly out' records a point that falls in the outer half of the confidence band; technically speaking the point would be returned as falling outside the polygon, but it could actually be inside the 'true' polygon if it had been erroneously digitized or geocoded. An ambiguous point has coordinates that coincide exactly with a point on the digitized boundary-such points are rare, but do occur.

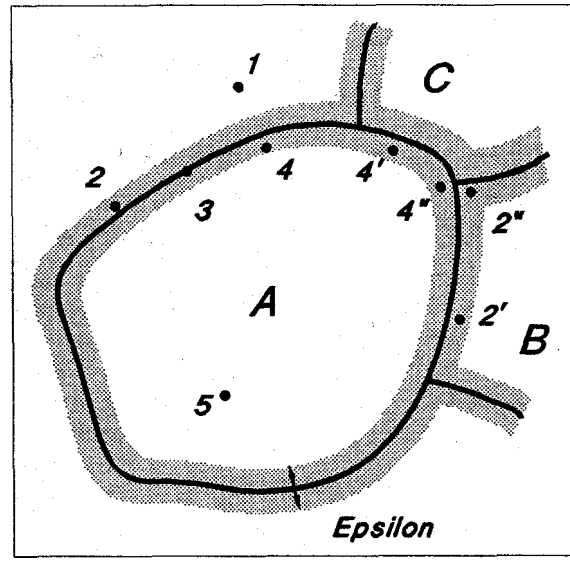

Figure 9.6. Perkal's concept of an epsilon error band around a digitized line

Blakemore (1984) illustrated the effects of these kinds of errors when dealing with problems of combining a vector polygonal net with a square grid cell network. The problem he chose was that of overlaying a UK Department of Industry 1 km square grid data base of industrial establishments on a polygonal map of 115 North-West England employment office areas. A total of 780 entries in the database geocoded to a 1 km square grid resolution were tested for their inclusion in the polygon network. The 1 km square grid leads to an epsilon or confidence band of 0.7071 km. Table 9.4 presents Blakemore's results.

The 'possibly out definite' class includes data points that fell outside the polygon network of employment office areas altogether. 'Possibly in' refers to points falling within the inner half of the error band in polygons on the edge of the polygon net. 'Unassignable' refers to points that fell outside the error band of the 'outer boundaries of the outer polygons. In some circumstances the point-in-polygon routine suggested that the industry was located in the sea! 'Possibly in/out 2 polys.' refers to points that were flagged as being possibly in and possibly out of two adjacent polygons; 'possibly in/out > 2 polys' refers to points that were possibly in or out of more than 2 polygons. 'Ambiguous' refers to those points actually occurring on the digitized polygon boundaries. The implication of the study was that only 60 per cent of the workforce in the industries in the database could definitely be

Table 9.4. Epsilon error results

Category	%
Possibly out definite	1.5
Possibly in	4.4
Unassignable	1.4
Possibly in/out 2 polys.	29.8
Possibly in/out > 2 polys.	6.7
Ambiguous	1.2
Subtotal	45.0
Definitely in	55.0
Total	100.0

associated with an employment office area. The mismatch errors and ambiguities were relatively larger for long, thin polygons and for employment areas having narrow protuberances or insets than for large, broadly circular areas. The study resulted in a considerable amount of validation and checking of the data bases to ensure that the errors brought about by the grid-cell point geocoding were removed. Perkal's epsilon assumes the boundary is real, the problem merely being one of knowing its location. Sometimes it is not the location but the *existence* of the boundary that is in doubt, and then other methods must be used-see Chapter **11.**

ERRORS ASSOCIATED WITH OVERLAYING TWO OR MORE POLYGON NETWORKS

Spatial associations between two or more thematic maps of an area are commonly displayed or investigated by laying the polygonal outlines on top of one another and looking for boundary coincidences. Before the days of digital maps, the process was achieved using transparent sheets, and the boundary coincidences were established using fat marker pens to trace the results. The onset of the digital map promised better results because all boundaries were supposed to be precisely encoded, but in fact the result of the new technology was to throw up one of the most difficult and most researched problems in computer cartography. Not only did a solution of the problem in technical terms cost many years' work but investigations have shown that the results of overlay throw up more questions about data quality and boundary mismatching than they solve.

McAlpine and Cook (1971) were among the first to investigate the problem when working with land resources data and their method still holds. They considered two maps of the same locality containing respectively m_1 and $m_2 (m_1 >= m_2)$ initial map segments (polygons) that are overlaid to give a derived map having n segments. To simplify the problem, they experimented by throwing a single hexagon with random orientation and displacement over a mosaic of hexagons. The trials were done using a hexagon of side 0.5, 1, 2, and 3 times the sides of the mosaic hexagons. They found that the number of derived polygons n on the derived map could be estimated by

$$n = m_1 + m_2 + 2 \cdot \{m_1 m_2\}^{1/2} \qquad 9.10$$

for two maps, which for k maps can be generalized to

$$n = \left[\sum_{i=1}^{k} m_i \right]^2 \qquad 9.11$$

McAlpine and Cook (1971) showed that map overlay gave rise to a surprisingly large proportion of small polygons on the derived map. They applied their analysis to a case-study of overlaying three maps of scale 1: 250 000 of census divisions, present land use intensity and land systems from Papua and New Guinea containing 7, 42, and 101 initial polygons respectively. The overlay of the three maps gave 304 derived polygons (Equation (9.11) estimates 368 derived polygons, but McAlpine and Cook regard this as satisfactory). The overlay process resulted in 38 per cent of the area being covered by polygons having areas of less than 3.8 sq. kilometres.

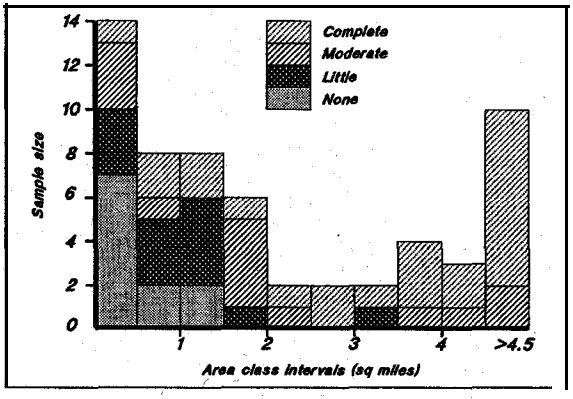

Figure 9.7. Measure of agreement between initial and derived polygon descriptions after polygon overlay

The results of the overlay were evaluated by classifying the derived polygons by size and boundary complexity (i.e. polygons bounded solely by initial mapping segments, those bounded only by land use and land system boundaries, and those bounded by all three types of boundaries). A 10 per cent random sample of derived polygons was evaluated by three colleagues to determine the measure of agreement between the initial and the derived polygon descriptions. As Figure 9.7 shows, the lack of agreement was substantial for the smallest derived polygons, and some 30 per cent of the area of the derived map was represented by polygons that had little or no agreement with the initial descriptions.

Goodchild (1978) extended the discussion of the polygon overlay problem to show that the number of derived polygons is more a function of boundary complexity than the numbers of polygons on the overlaid maps. He showed that an overlay of two polygons having respectively vl and v2 vertices could produce any number of derived polygons from three to v1.v2 + 2 when all Boolean operations including .NOT.A.AND.NOT.B are used. Moderate numbers of derived polygons are produced when, as in McAlpine and Cook's example, the overlaid maps show statistical independence. When the boundaries of polygons on the source maps are highly correlated, however, serious problems arise through production of large numbers of small, 'spurious' polygons. Prominent and important features, such as district boundaries or rivers, may occur as part of polygon boundaries in several maps. These several representations of the same boundary will have been separately digitized, but because of digitizing and other errors will not exactly coincide.

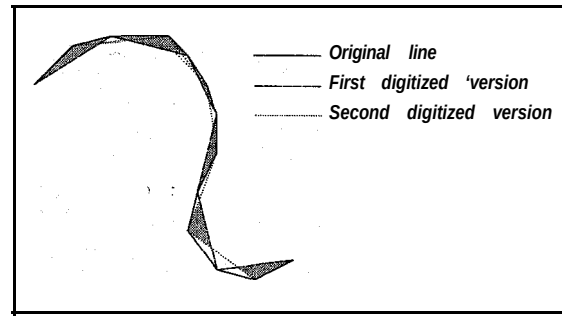

Figure 9.8. How spurious polygons occur in map overlay when the same line is digitized twice

The spurious polygon problem contains two apparent paradoxes. First, the more accurately each boundary is digitized on the separate maps, and the more coordinates are used, the larger the number of spurious polygons produced. Second, subjective methods of map drawing, designed to assist the eye in generalizing when using manual methods of overlay, result in large problems when working with digital maps.

Goodchild (1978) analysed the situations in which spurious polygons were most likely to occur through the conjunction of two digitized versions of the same arc, with n_1 and n_2 vertices respectively (Figure 9.8). Goodchild, using the statistics of runs of binary symbols, showed that the number of spurious polygons S generated by conjunction of two arcs having n_1 and n_2 vertices ranges from

$$Smin = O \qquad 9.12$$

to

$$Smax = 2 \min(n_1, n_2) - 4 \qquad 9.13$$

with a random expectation of

$$E(S) = [2n_1, n_2/(n_1 + n_2)] - 3 \qquad 9.14$$

if symbols occur randomly in sequence along the conjoined arcs. The minimum value of S occurs when the overlap is of maps having symbols of one type occurring together; the maximum value of S occurs for maximum intermixing. By simulating five possible situations in which arcs were conjoined, Goodchild showed that equation (9.14) overestimates the average number of spurious polygons that can occur by some 17 per cent. The actual number of spurious polygons found never exceeded 71 per cent of Smax. The more carefully a map is digitized, however, the larger the values of n_1 and n_2, and so the larger the number of spurious polygons will become.

Spurious polygons are in fact equivalent to the mismatch areas resulting from rasterizing a polygon. Their total area should decrease as digitizing accuracy increases, but the greater problem is their removal to avoid nonsense on the final map. They can be removed by erasing one side on a random basis, after screening the polygon for minimum area, or the two end-points can be connected by a straight line and both sides dissolved. A more sophisticated approach is to consider all points within a given distance from the complex arc as estimates of the location,of a new line, and then fit a new line by least squares or maximum likelihood methods. Unless one version of the digitized boundary can be taken to be definitive, it is highly likely that the complex line will be moved from its topographically 'true' position. The net result of overlaying a soil map (having not very exact boundary locations) with a county boundary map (topographically exact boundaries) may be that the topographic boundaries become distorted unless the user specifies that they should remain constant.

Adopting exact paradigms of exact boundaries or smooth contour lines for spatial entities presents problems for converting from one representation to another and for entity overlay and intersection, but methods exist to estimate the errors that are involved in these actions. The degree of error caused by forcing of spatial phenomena into possibly inappropriate, exact, crisply defined entities has received less attention but may be a major source of unseen errors and information loss. Geographical phenomena with uncertain boundaries are covered in Chapter 11. (See also Burrough and Frank 1996.)

Summary: errors and mistakes

As in any manufacturing process, poor quality raw materials leads to poor quality products. Spatial information systems, however, also make it possible to turn good raw materials into poor products, if proper attention is not paid to the ways data are collected, modelled, and analysed. Conventionally, data quality has been linked to the precision of geographic coordinates, but today, exactness of location is but one aspect of data quality. The reader should also be aware that sometimes people expect a higher-quality product than is strictly possible, or even necessary. For example, for auto navigation it is extremely important that the database is geometrically and factually precise, but for marketing studies (e.g. Plates 2.5–2.8) extreme spatial accuracy is not only necessary, but threatens individual privacy.

Questions

1. Review the different methods that can be used to determine errors in spatial data. Consider a range of different GIS applications and assign appropriate error analysis techniques to each application.

2. Design a meta data system for recording the results of data quality and error propagation as active aspects of a spatial data set.

3. Review four practical situations where lack of information about errors could be critical for the acceptance of the results of GIS analyses.

4. Compile lists of the sources of errors for each of the examples of GIS analysis given in Chapters 6 and 7 and classify these errors using the terms given in this chapter. For

each example decide which source of error is most likely to be critical for successful analysis.

Suggestions for further reading

DAVID, B., **VAN DEN HERREWEGEN,** M., and SALGÉ, F. (1996). Conceptual models for geometry and quality of geographic information. In P. A. Burrough and A. U. Frank (eds.), Geographic *Objects with Indeterminate* Boundaries. Taylor & Francis, London.

GOODCHILD, M., and GOPAL, S. (1989). The Accuracy *of Spatial* Databases. Taylor & Francis, London.

GUPTILL, S., and **MORRISON,** J. (eds.) (1995). The ***Elements*** of Spatial Data ***Quality.*** Elsevier, Amsterdam.

THAPA, K., and BOSSLER, J. (1992). Accuracy of spatial data used in Geographic Information Systems. ***Photogrammetric Engineering and Remote Sensing, 58:*** 841-58.

Principles of Geographical Information Systems

Peter A. Burrough

AND

Rachael **A. McDonnell**

OXFORD UNIVERSITY PRESS
1998