SPATIAL DATA INTEGRATION

R FLOWERDEW

Data integration is the process by which different sets of data within a GIS are made compatible with each other. These data sets may or may not be defined in terms of the same geographical referencing system. Different data sets have different spatial coverage; many data collecting agencies have their own system of regions, and these regional systems are subject to boundary changes over time. Data for different regions may be collected in incompatible ways, may vary in reliability or may be missing or undefined. The larger the number of different data sources needing to be integrated, the more such problems will be encountered.

Other problems in data integration relate to incompatibilities between the spatial entities for which data are recorded. Sometimes these are a result of differences in dimension; often data exist for a set of points but are needed for a continuous area, and the appropriate process is interpolation. Sometimes data are available only for a set of zones, and are needed for a different set of zones, or for point locations. Methods exist based on the assumption that such data reflect an underlying smooth surface, and other methods are being developed which take account of other variables in transforming data between zonal systems.

INTRODUCTION

Data integration is the process of making different data sets compatible with each other, so that they can reasonably be displayed on the same map and so that their relationships can sensibly be analysed (Rhind *et al.* 1984). As such, it is one of the most important topics in the whole field of GIS. It is often an essential preliminary to the use of GIS for investigation of substantive questions. It is a problem which recurs in almost all applications of GIS; the more ambitious the application and the more data sets that are needed, the more likely it is that data integration will be a problem. This may be particularly so when both environmental and socio-economic data are involved. This chapter reviews the main issues involved, although some are considered in more detail in other chapters. Most of the examples are taken from socio-economic applications of GIS, because this is where the author's main experience lies.

Data integration has several different aspects. These can be summarized in terms of a number of straightforward questions (whose implications may be far from straightforward!):

- What type of data?
- Where do the data refer to?
- When do the data refer to?
- How accurate are the data?

It is assumed that the data are geographical, in other words that each observation to be included has two aspects – what was observed, and where it was observed. In many cases, it is also relevant to consider when it was observed. The first question can be regarded as being about the measurement scale of the data: does the observation refer to the presence or absence of something, to the category that something has been assigned to, or to some more quantitative measure of the size or intensity of whatever is being studied? The second question has two main aspects: does an observation refer to a

point, a line or an area (each will be treated differently in a GIS), and how is the location of the observation represented (in other words, what reference system is used to record the data)? The third question may refer to a specific point of time or period of time. The fourth can refer to error of several kinds, both measurement and locational, including mistakes, imprecision and estimation. All of these issues will be referred to in the following sections. For further discussion of the nature of geographical data see Unwin (1981), Fisher (1991 in this volume) and Gatrell (1991 in this volume).

WHAT TYPE OF DATA?

- *Dichotomous or presence/absence.* This measurement scale is obviously relevant when considering the presence or absence of a plant or animal species in an area; it also applies where places are classified into one of two categories – a country may or may not be a member of NATO, a road may or may not be a dual carriageway (divided highway), a city may or may not have a convention centre.
- *Categorical.* This measurement scale is used when a place can be classified into one of several categories – rock type, vegetation cover and system of government are examples.
- *Ranked.* There are two types of ranked data; ranked (or ordered) categories are used where a set of categories has a natural ranking associated with it – for example, grades of agricultural land; alternatively a set of places may be ranked from first to last according to some criterion, such as the rankings of urban residential desirability fashionable in the United States (Cutter 1985).
- *Count.* Data consisting of the number of items or the number of times something has happened in a place – population, the number of species, the number of television channels, the traffic count or the number of customers.
- *Continuous.* A measurement on a continuous scale, such as wheat production, average annual rainfall, height above sea level or the unemployment rate. Sometimes this may be reducible to the ratio or the sum of count variables, sometimes not.

WHERE DO THE DATA REFER TO?

- *Points.* Data may relate to sample points, either selected randomly (as in some soil or vegetation surveys) or for convenience (spot heights; rain gauges); they may also relate to real entities, like trees, factories or cities (which can be considered as points at some scales).
- *Lines.* Line data may also be obtained for sample lines, like transects, or for real linear phenomena, like rivers, railways or geological faults.
- *Areas.* Some areas used in GIS may be thought of as natural units in Unwin's (1981) terminology, that is, areas whose boundaries are defined by the value of the variable under consideration, such as rock outcrops, islands or marshes; others may be imposed units, where data have been collected for some artificially defined unit, such as a local government area.
- *Surfaces* (interpolated points). Many phenomena are defined everywhere but can only be measured at discrete points – height above sea level, annual rainfall and vegetation cover are in this category. Within a GIS, interpolation methods can be used to estimate values for other points and to construct a surface. The TIN (Triangulated Irregular Network) surface representation is an example of how a surface can be stored and displayed within a GIS (see Weibel and Heller 1991 in this volume).

Reference systems

One way in which data from two maps of the same region may be integrated is through relating the location of map features to a reference system. This is typically a pair of numbers defining the distance east and the distance north from a fixed point (see Maling 1991 in this volume). In the vector representation system (the most common used in cartographic applications), a line is represented as a

set of these number pairs defining the coordinates of points along the line. These numbers may be table coordinates, based simply on how the digitizing table was set up when points and lines were digitized with no other significance. Alternatively they may be unique to a particular map; commercially produced street maps may refer to locations in terms of their own specially designed grid.

It is more common, however, for map feature representations to be linked to one of a few standard referencing systems. The most general of these is the network of lines of latitude and longitude. Others include the Universal Transverse Mercator (UTM) system in common use in North America and the Ordnance Survey National Grid used in Great Britain. The fact that the world is (approximately) spherical and not flat, however, means that no two-dimensional coordinate system, and hence no two-dimensional GIS, can represent the earth's surface without distortion. Figure 24.1 illustrates the lack of conformity between latitude and longitude and the National Grid. This distortion increases in seriousness with the size of the area represented.

Maps based on latitude and longitude will not necessarily be compatible with each other because of the many different projections available for mapping. Even within the same map, problems arise because the length of a degree of longitude is not constant, changing quite dramatically approaching the poles. It is also the case that most projections do not represent lines of longitude as straight (often this is true of lines of latitude too). Although the US Geological Survey produces 7.5 minute quad sheets, defined by latitude and longitude boundaries, their curvature makes latitude and longitude unsuitable coordinates for a GIS (Aanstoos and Weitzel 1988). Digitizing a map can only be done with the aid of two orthogonal coordinate axes and points located with reference to curved lines cannot simply be integrated with data using an orthogonal system. Even those projections with straight-line graticules (of which Mercator is the best known), because of the distortions of shape and/or area involved, cannot easily be integrated with other data sources. Some fundamental GIS operations, like calculations of polygon areas, will of course be wrong if data have not been input from a map with an equal-area projection.

The British National Grid is only satisfactory

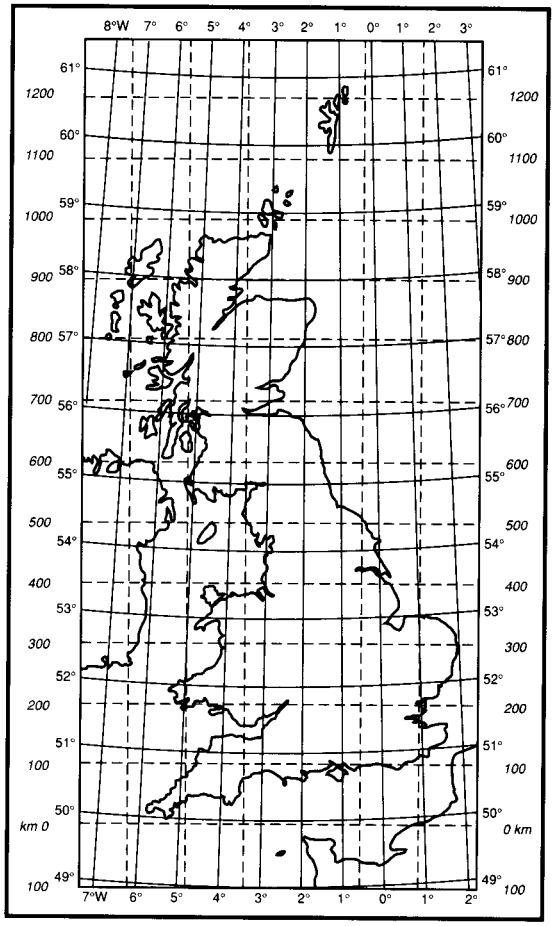

Fig. 24.1 Relationship of the Transverse Mercator graticule (solid) to the National Grid system (pecked) of the Ordnance Survey.

for use in GIS because Great Britain is small enough for distortion to be relatively minor; attempts to extend it, even if only as far as Ireland, rapidly become unacceptable. Other systems based on standard meridians or parallels are also subject to error increasing with distance from the centre of the map, leading to obvious errors when maps based on different standard lines are to be integrated. Mapping the state of Texas, for example, on the State Plane system (based on the Lambert Conformal Conic projection) would have necessitated using five separate coordinate systems, a problem which Aanstoos and Weitzel (1988) overcame by defining their own Lambert projection with parameters optimal for the entire state. Where a large area is mapped on a UTM system, integration problems arise for places equidistant

from two of the standard meridians used to define the system. In Canada, for example, mapping the area around Calgary on a UTM system is problematic, since the metropolitan area is split between coordinate systems defined around two different meridians (a second example is illustrated in Fig. 24.2).

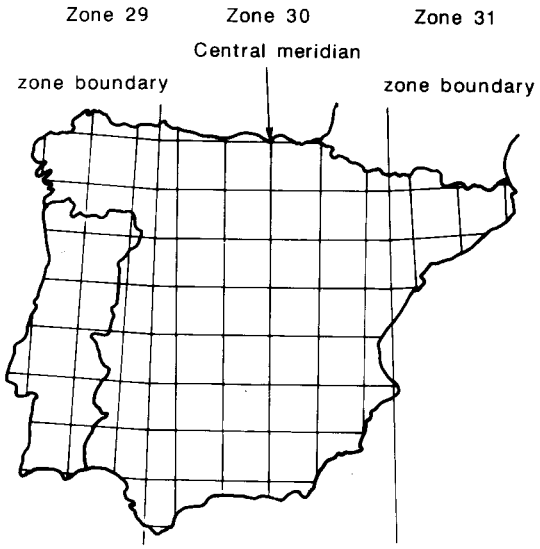

Fig. 24.2 Map of the Iberian peninsula showing divergence of UTM grids around zone boundaries.

A further type of reference problem occurs when photographic imagery is being input into GIS. If air photography is oblique, projection problems arise as a result of varying scale over the image. Even for vertical photographs, scale is not constant and distortions occur with increasing distance from the centre of the photograph (Maling 1989: 247–76). Such problems are particularly acute for satellite photography because the greater the altitude the greater the effect of the earth's curvature on image distortion.

Most of these projection problems are well known to surveyors and cartographers, and for many of them solutions exist and can be operationalized. Any GIS system should allow the conversion of table coordinates to a user-defined set (see Bracken and Webster 1990: 211–22), and many include routines for conversion between different projections. Algorithms exist for conversion between latitude and longitude and UTM, although they are complicated. It is not a simple matter, however, to combine within GIS, two maps drawn on different projections. Maling (1973) provides a good guide to problems of this nature.

Data set coverage

A very common problem in data integration is the difference in the area for which data are available for two different variables. The ideal would be for each variable needed in the GIS to be mapped separately at the same scale and for the same areal extent. In practice, map sheets will overlap and data may not be available for all the areas required.

If two or more map sheets are being input into the same GIS, problems may occur at the edges of map sheets, even if they are based on a common referencing system. Such problems are likely to be associated with linking up line or area phenomena which cross the boundary between the map sheets. A fundamental operation in any vector-based GIS system is polygon creation, in which the GIS operates on a set of line segments to produce a set of well-formed polygons. If two points (such as the western end of a line on one map and the eastern end on another) are intended to be the same but are actually digitized as being at slightly different locations, the system has major problems in deciding whether or not to treat the points as the same or different (Fig. 24.3).

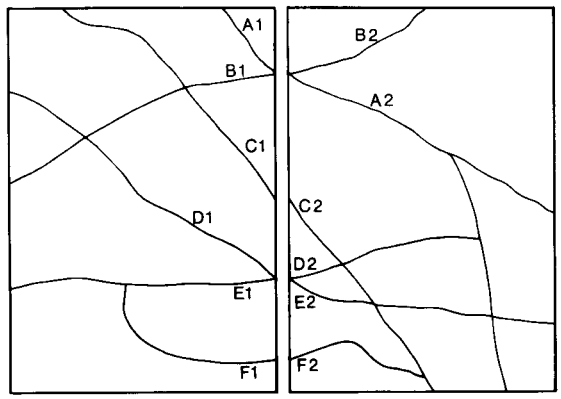

Fig. 24.3 Edge matching during data integration.

Both mapping agencies and data collection agencies organize their operations spatially. This might mean, first, that information must be acquired for a larger area than is actually needed because of the way map sheets, or data collection

units, have been divided up. The English city of Oxford, for example, is in the South East Standard Region, the Central television area, the Western Post Office region and has its own Regional Health Authority. These are all different sizes and shapes and, therefore, only partially overlap. Hearnshaw, Maguire and Worboys (1989) provide a systematic treatment of the range of data units relevant to the English county of Leicestershire. Integrating data from all these sources may mean that the GIS is confined to a very narrow area, where all these regions overlap, or that much information must be collected that will never be used (Fig. 24.4). Most map users have experienced the frustration of having to acquire and handle large map sheets of which only one small corner is actually relevant to their needs (Fig. 24.5).

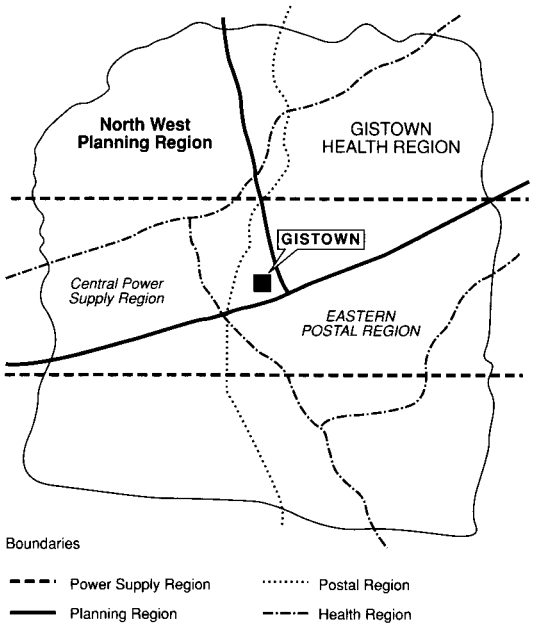

Fig. 24.4 Different regionalizations as a problem in data integration.

It may be that information is depicted or symbolized differently for different parts of the area being studied. A related problem may be that more detail is available for some parts than others. Little difficulty is caused if the phenomena mapped are equivalent but the symbolization is different – for example, if roads are drawn in red on one map and blue on another. Some digitizing problems may exist if they are depicted as a pair of parallel lines on

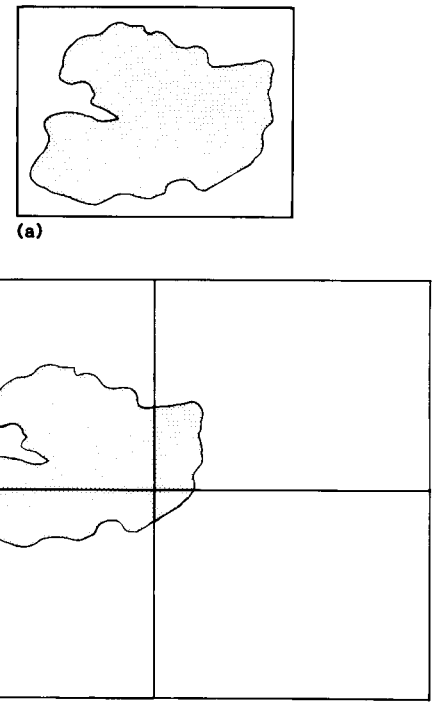

Fig. 24.5 (a) Ideal map sheet boundaries; (b) Actual map sheet boundaries.

one map and just as a centre line on another. If a different classification is used, the problem is greater – for example, one map (to give a British example) may distinguish between class A and class B roads, while another may distinguish between dual carriageways and undivided highways. A simpler example is that different maps may use different contour intervals – and hence a hill or depression of moderate size may be marked on one map and an exact equivalent omitted from another (Fig. 24.6). Integration is a still greater problem if the method of depiction is totally different. For example, a contour system for showing altitude cannot easily be compared with one reliant solely on spot heights; these systems can only be integrated by transforming one set of data. More realistically, a map showing cities as circles whose type depends on city size is not fully compatible with one showing the boundaries of their built-up area.

Data may not only be symbolized differently but also may actually have been collected differently. Regional offices of a national agency may have freedom to decide on how they collect

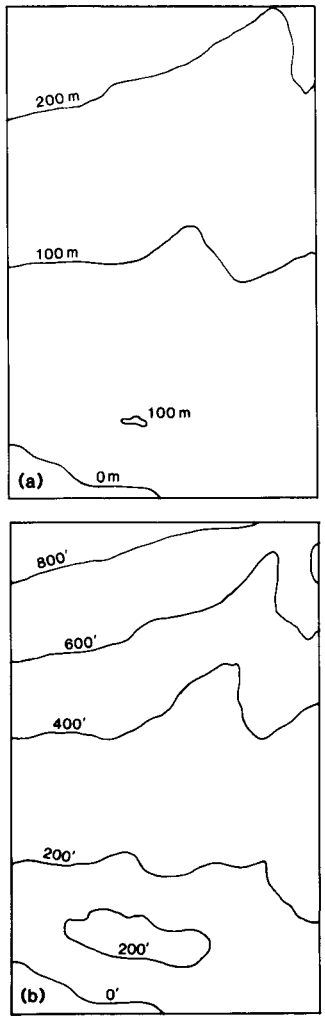

Fig. 24.6 The same landscape with (a) 100 metre contours; (b) 200 foot contours.

information, and may well make different decisions, perhaps for very good reasons. Even basic data sources like the British census include minor differences in the data collected between England and Wales, and Scotland (the definition of a room), while such differences are magnified where there is scope for subjectivity, as in geomorphological or soils maps.

Some of these problems are annoying rather than crucial, for example the overlapping regional data sets that must be assembled. Others are virtually insuperable, but may be tackled to some extent by trying to reconcile the data differences: usually this involves making intelligent guesses about what the data really needed would be like if they were available. Problems of this type can often be regarded as examples of missing data problems. Sometimes the obvious solution is to use only the lower quality or less detailed data if those are all that are available for the entire region of interest. However, if more detail is needed or data are absent altogether, then interpolation methods of various types may be used to try to guess what is going on. These methods are of two main types according to what information is used to guide the guesses. If information is available for variables related to the one required, an appropriate form of statistical estimation may be used to predict a likely value; for example regression or categorical data analysis (Williams 1984, 1986; Wrigley 1985). If the variable required is likely to show systematic spatial variation, spatial interpolation techniques may be appropriate. These methods are reviewed by Weibel and Heller (1991 in this volume) and, in different contexts, by Lam (1983), Schut (1976) and Tobler and Kennedy (1985).

WHEN DO THE DATA REFER TO?

Data for different places may be collected (or mapped) at different times; indeed, this is naturally to be expected where the process is expensive in time or resources. One map sheet may have been revised last year while its neighbour has long been out of date (Fig. 24.7). Changes in methods of symbolization may have occurred in the mean time as well as changes in the phenomenon concerned. More drastically, data may not yet have been produced for some areas. It may be that production is a time-consuming process and the agency concerned has not yet reached the areas in question. Alternatively the data may never be produced – it may not be cost effective to do so, people in the region may object to its collection or release, or there may just have been an oversight by the person responsible.

Comparison between places on the map may of course be made more difficult as a result of the data referring to different times. This problem is particularly acute for those producing an international atlas or statistical compendium. Sometimes maps may be available for one place for two dates and for another place at an intermediate

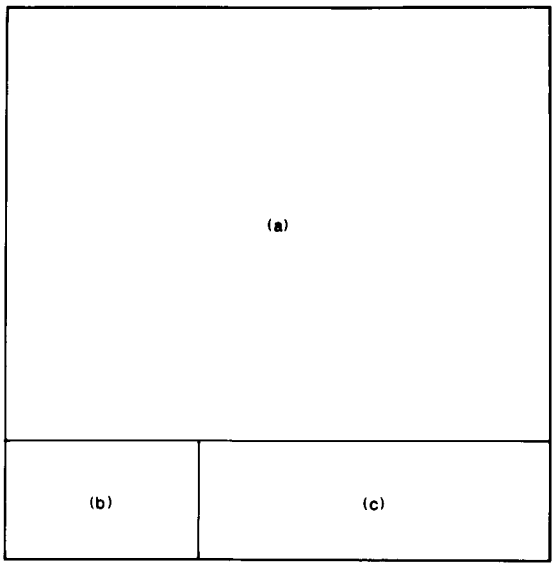

Fig. 24.7 Survey dates for revision for an imaginary topographic map (a) 1962–63; (b) 1963–64; (c) 1959. Revised for major roads and other significant changes 1974.

date; the best comparison may perhaps be based on estimating data for the first place at the intermediate date. To do this, however, requires some assumptions about the trajectory of change.

Another potential problem relates to shorter time scales. Many phenomena fluctuate on an annual or daily cycle, or perhaps more haphazardly. Comparisons may be distorted if observations for different places relate to different times in such cycles. It may be inappropriate to compare a vegetation survey of one place in July with one of another place in September. Even if surveys were done at the same time in different years, climatic fluctuations may distort the comparison.

Data may also represent accumulations of observations or averages over time. For example, the number of times relatively rare events (such as floods, earthquakes, power failures or mortgage defaults) have occurred may be of interest, but the figures are obviously sensitive to the period over which observations have been made. In addition, external events may affect such occurrences and, for example, mortgage defaults may look very different if they are observed over a period including a national or local economic recession than over a period of the same length characterized by boom conditions. Figures for average values of some fluctuating variable, such as rainfall, crop yields or disease occurrence, will also depend on the length of time (and the precise time period) for which data are collected. A 50-year rainfall average is likely to be more reliable than a 10-year average, and as usual the GIS user has the responsibility of deciding what to do if only the 10-year average is available.

ERROR AND ACCURACY

Data integration may also be affected by error in one or more of the maps incorporated into the GIS. This topic is discussed by Chrisman (1991 in this volume) and, with the related subject of accuracy in GIS, was the theme of the first of the major initiatives launched by the US National Center for Geographic Information and Analysis (NCGIA). The initiative's volume of position papers (Goodchild and Gopal 1989) is a goldmine of informed discussion and analysis of different aspects of these problems.

Veregin (1989) distinguishes between different types of error in two important respects. First, there is a distinction between 'cartographic' error, error in the positions of map features such as points, lines and areas, and 'thematic' error, error in the values of an attribute of map features. Second, he differentiates between 'measurement' error, or imprecision in the location or attributes of features, and 'conceptual' error, error associated with the process of translating real-world features into map objects. He also considers how these types of error are combined when two maps are overlaid (i.e. when data are integrated).

There is now a good deal of literature on the treatment of cartographic measurement error, which is usually thought of as arising from digitizing error, although error of this sort can also be generated during the original map production process. Burrough (1986), for example, discusses early work on this topic, while Maffini, Arno and Bitterlich (1989) provide one recent treatment of the issue. One problem arises simply from the level of precision possible on a paper map. As Goodchild and Gopal (1989: xii–xiii) point out, the precision with which map features are recorded on paper maps is generally less than that with which they are recorded in a vector-based GIS. The problem is accentuated because maps at different scales may be included in the same GIS; a reasonable level of

precision in recording data from a large-scale plan may be totally spurious if they are to be integrated with other data taken from a smaller-scale map. The spurious precision with which data may be recorded in a GIS makes it impossible for the best digitizing technician to digitize the same line twice in exactly the same way; and human error and inaccuracy magnify the problem. The results of this imprecision may not matter for a map of a single phenomenon, but problems arise when data integration takes place. Overlaying two zonal systems, for example, may result in the creation of a host of 'sliver polygons' and 'dangling chains', geometrical entities arising in the topological structuring procedure of a vector GIS if points do not lie exactly on the lines they should be on (Fig. 24.8).

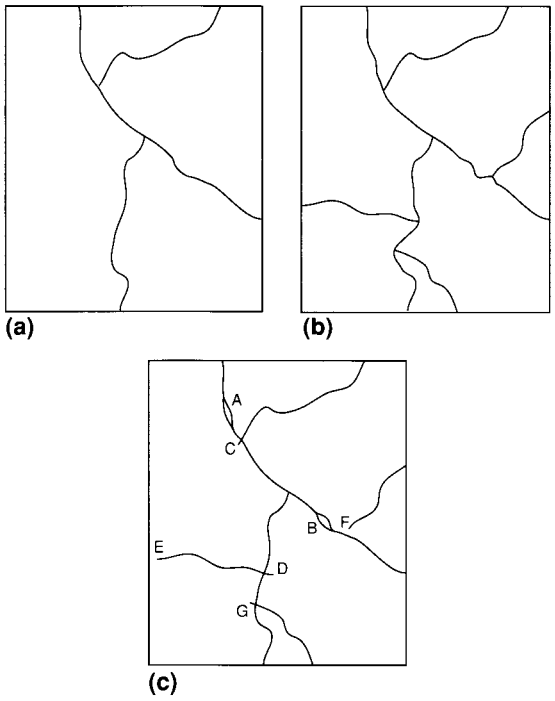

Fig. 24.8 Generation of sliver polygons (A,B) and dangling chains (C,D,E,F,G) from adding a new set of lines to an existing coverage. (a) Limited coverage; (b) New lines added; (c) result of adding (b) to (a).

Here it may be relevant to consider how digitizing can be done in a consistent manner. First, registration, the position of the map with respect to reference points with known coordinates, must be consistent between different digitizing sessions. Second, there must be consistency in the level of map generalization, as reflected in the degree of detail in line boundaries and the inclusion and exclusion of point and areal features. An important point may be whether topological relationships are preserved between two sets of phenomena digitized separately. For example, it may be important that a set of points is in one-to-one correspondence with a set of polygons (if they are to be used as 'label points'); a very small error in the position of either object can be a major problem if it results in a point being outside its polygon, while a much larger error may be unimportant if topological relationships are preserved. In Fig. 24.9, for example, a small error in locating point A might leave it in the wrong polygon, whereas a much greater error in locating B would be relatively harmless. Blakemore (1984) reviews these problems and suggests a technique based on the concept of 'epsilon' distance.

Most vector-based GIS have facilities for

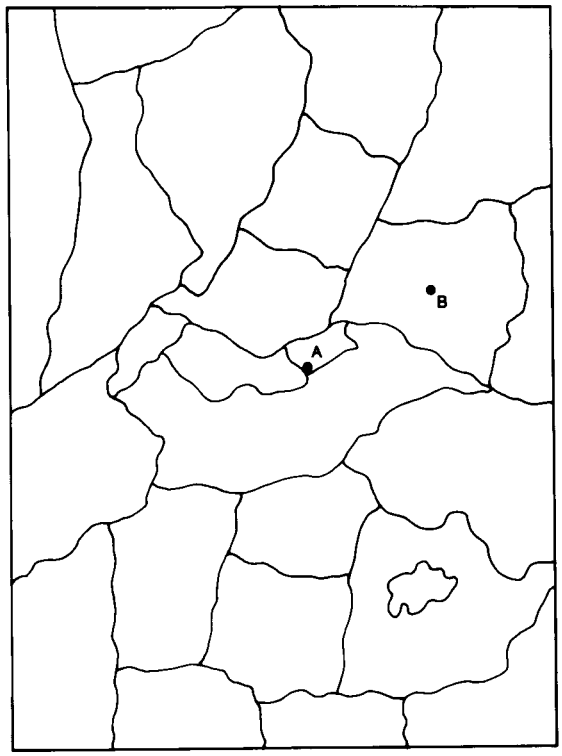

Fig. 24.9 The importance of topology when working with label points.

dealing with these problems usually based on the concept of 'tolerance', a critical distance within which two points are regarded as being identical. In theory, choice of a suitable tolerance level should remove the spurious lines and polygons and produce a correct map. In practice, at least in this author's experience, the process invariably leads to headaches. Either the tolerance is set too small and the mistakes are not removed, or it is set too large and distinct points are amalgamated; perhaps an isthmus or river meander is cut through, leading to the formation of new spurious polygons of a different type (Fig. 24.10). Flowerdew and Banting (1989) discuss problems of this type encountered when attempting to update a previously digitized map.

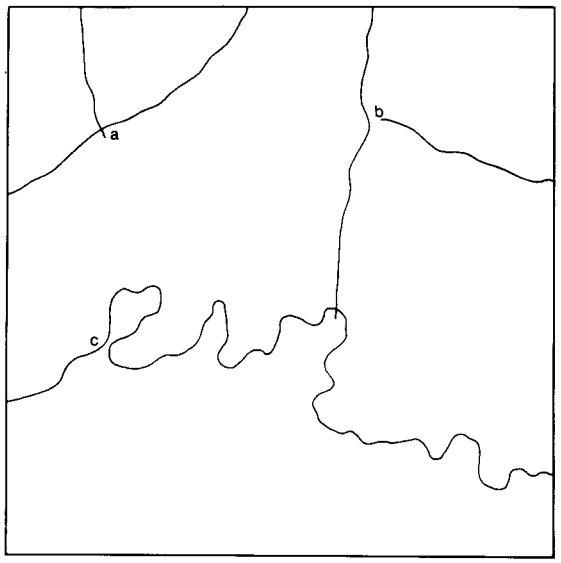

Fig. 24.10 Problems in setting tolerance levels when building topology. If the tolerance level is set large enough to correct the errors at a and b, the loop at c will also (incorrectly) be closed.

Thematic measurement error can itself be subdivided according to the scale at which the attribute is measured. If it is measured on a continuous scale, as with elevation or rainfall, for example, it can only be recorded to some specified level of accuracy. If it is a count or a categorical variable, exact recording is possible. However, whatever the scale of measurement, the map (or other data source) may not display the attribute to the level of accuracy possible. A point or line attribute (such as city size or pipeline diameter), for example, may be shown as one of a set of graduated symbols, and an area attribute (such as population density) may be depicted according to which of a set of class intervals it falls into. Of course, an important advantage of a GIS over a map is that attribute data can be stored accurately without having to tackle the problems of mapping them clearly, but accuracy in a GIS may not always be possible if it is not present in the component data sets.

Particular problems occur for data (such as the elevation and rainfall examples mentioned above) which are defined everywhere. Frequently these are mapped as isopleths or contours – even if these lines were totally reliable, a point between two lines could have any value within the range the lines define. An additional source of error here (perhaps better regarded as attribute conceptual error) is that the value for any point is likely to be based on an interpolation procedure and may be wildly out if the assumptions of the procedure do not hold.

Another important special case of thematic measurement error occurs with categorical attributes. The assignment of points, lines or areas to a category is based on a classification of some type, and many such classifications can be drawn up with differing degrees of detail. Zoning or land use maps are cases in point. Sometimes the categories apply to points in space rather than to pre-defined areal units – to categorical coverages in Chrisman's terminology (1989). Examples include geology, soils and vegetation maps. Such attributes raise particularly major measurement problems because of the varying level of detail to which classification is possible.

Cartographic conceptual error raises further difficulties in integrating data sets. There may be uncertainty about where to place a point symbol intended to represent a large city, but it is with line and area phenomena that the problem is greatest. The location of a coastline involves far from obvious decisions about high, low or median water levels, about generalization, and about historical change (Fig. 24.11). For a single map, this may not matter much but, when two data sets are overlaid, differences in the coastline can generate many sliver polygons and other incompatibilities. Decisions about the placement of property and other areal unit boundaries may raise similar difficulties. In addition to error of this type, perhaps attributable

to conceptual fuzziness, other errors due to incompetence, incorrect guesswork, low standards of accuracy or intention to mislead should not be ignored. As an example of the third of these, attempts to integrate Canadian census and postal zone coverages ran up against the inadequacy of postal zone maps, which were not true to scale (unimportant for post office purposes provided that the zone boundaries were clearly marked, but fatal for data integration purposes). Demko and Hezlep (1989) illustrate the fourth point with their illustration of how Soviet maps were purposefully distorted until recently.

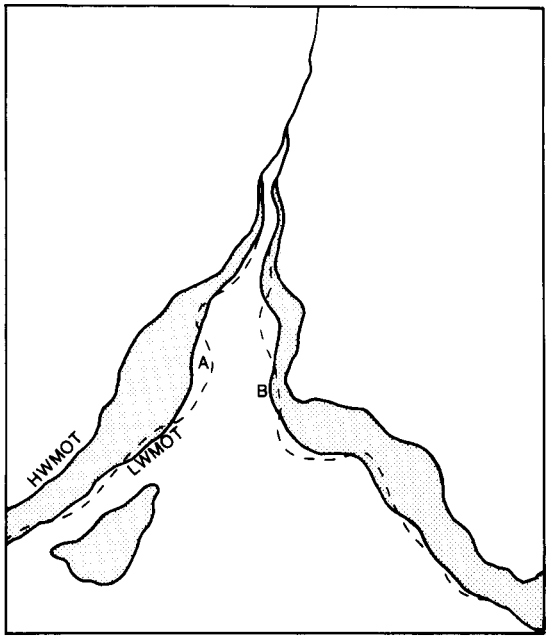

Fig. 24.11 Problems in digitizing coastlines. A, coastline in 1960; B, Coastline in 1969. HWMOT = High Water Mark, Ordinary Tides; LWMOT = Low Water Mark, Ordinary Tides.

Thematic conceptual error can likewise be a matter of incompetence, guesswork, inaccuracy or deliberate distortion. In addition it may arise from conceptual fuzziness. In the case of categorical coverages, like vegetation or soils, there may be difficult decisions to make about whether a small area, which differs from the surrounding region, should be picked out as constituting a separate patch. The concept of a transition zone also creates problems. Over space there may be a gradual change from a dominantly forest area to a dominantly grassland area, for example; constructing a categorical coverage forces the GIS user to split this transition zone into two distinct categories. Even if 'forest/grassland transition' is allowed as a separate category, the problem is not solved, because the user must then make a sharp boundary between the transition zone and each of the original categories.

Errors of all these types in individual maps create difficulties in GIS use, but what happens when several different maps are to be integrated? It might be claimed that the error involved in using any GIS is the sum of the errors in all the component data sets. A less stringent view would be that the error is at least as great as the error involved in the worst of the component data sets. A still more optimistic view is that errors in some data sets may cancel out, or that it may be possible to correct an error in one data set on the basis of the information contained in the others. In practice, of course, the effects depend on what the error is and how the maps are to be combined. Goodchild and Gopal (1989) and many of the contributors to their volume argue that one of the greatest research challenges in the GIS field is to track the effects of different types of error in GIS. They also argue (1989: xiv) that 'the objective should be a measure of uncertainty on every GIS product'.

Measuring the effects of error in the context of data integration is a major problem, and it is one tackled by Veregin (1989), Chrisman (1989), Maffini *et al.* (1989) and Lodwick (1989) among others. Openshaw (1989), in recognition of the fact that error in GIS databases is here to stay, advocates that there is a need to live with error. He suggests a general strategy based on Monte Carlo simulation to produce estimates of the likely extent and importance of error.

INCOMPATIBLE AREAL UNITS

Social, economic and demographic data are frequently collected for pre-defined zones of various types. Unfortunately for GIS users, however, the zonal systems used are not the same for different data sets, and may also be subject to change over time (see Flowerdew and Openshaw 1987 for a review of the problem). If one zonal system nests within another, the problem can be overcome by

aggregation, although it should be remembered that the results of geographical data analysis may depend on the areal units used (Openshaw 1984). More generally, however, zonal systems overlap: either attempts at comparison must be forgone, or a method of estimating the values for one set of zones on the basis of another must be devised.

One method of transforming data from one set of areal units to another is to assume that the data can in principle be represented as a smooth surface, and that the values for any areal unit can be calculated by integrating the surface for the zone defined by the areal unit boundaries. This approach was put forward by Tobler (1979) under the name 'pycnophylactic interpolation'. It appears to have merit to the extent that the surface being represented can reasonably be expected to be smooth. Much zonal data, however, like population, may vary abruptly, for example at the boundary between urban and rural areas. A somewhat similar approach has been developed by Martin (1989) for a related problem. British small-area population data are released for Enumeration Districts, whose centroids are generally available although the boundaries are not. Martin fits a surface to the centroid values in order to construct maps of the underlying distributions.

An alternative approach is to apportion the known data for a zone in one system (the source zone) so as to construct estimates for a zone in another system (the target zone). The obvious way of doing this is to assume that the data are distributed evenly within the source zone, and hence that a target zone constituting a certain proportion of the source zone should contain that same proportion of the data for the source zone. This areal weighting or areal interpolation method has been described by Goodchild and Lam (1980).

The problem with this method is that many kinds of data are most unlikely to be distributed evenly within source zones. Often a GIS user will have information for the target zones, or for other units which Goodchild, Anselin and Deichmann (1989) refer to as control zones, that makes it highly improbable that the distribution of target zone data will be even. Flowerdew and Green (1989) have investigated ways in which such information can be used to improve areal interpolation estimates; they suggest the use of the EM algorithm, originally devised to allow statistical analysis where some observations are missing, to derive these estimates. Langford, Maguire and Unwin (1991) illustrate the same approach with an application to population estimation from remote sensing data. Kehris (1990) has developed methods of linking a GIS to the statistical package GLIM to enable these methods to be operationalized.

CONCLUSION

This chapter has reviewed a number of considerations relevant to integrating different data sets within a GIS. There are several aspects of data integration that have not been treated. In particular, there is a large literature on the integration of cartographic and remote sensing databases (Bracken *et al.* 1989: 45–9; see also Davis and Simonett 1991 in this volume) and the related problem of integrating spatial data stored in raster and vector format (Devereux 1986). There have also been studies of the practical problems encountered in creating large integrated databases, for example in the CORINE project (Briggs and Mounsey 1989; see also Mounsey 1991 in this volume), in Belize (Robinson *et al.* 1989) and in the production of the BBC's Domesday videodisk (Openshaw, Wymer and Charlton 1986).

Two points should be made in conclusion. First, data integration is not a trivial or straightforward process; as with so many aspects of GIS use, the apparent ease and flexibility observable in a software demonstration obscures the necessity for a great deal of painstaking work. Further, the accuracy and tolerance levels of GIS may draw attention to problems that can be overlooked if cartographic comparison is all that is attempted. The process of data integration in a GIS may be salutary in that it forces GIS users to think explicitly about the comparability and accuracy of the different data sets they hold.

Second, and most important of all, data integration is at the very heart of GIS. The ability to combine together data of many different types and to display them in any combination is the main factor differentiating a GIS from mere database management systems on one hand and computer mapping systems on the other. The potential problems in data integration are many and fearsome, but it is well worth facing them, for that is the only way to get the full potential from GIS.

REFERENCES

Aanstoos R, Weitzel L (1988) Tracking oil and gas wells in Texas. *Proceedings of the Ninth Annual IASU Conference. Harnessing the hidden power of your system: maximizing your return-on-investment.* International Association of Synercom Users, Houston

Blakemore M J (1984) Generalisation and error in spatial data bases. *Cartographica* **21**: 131–9

Bracken I, Webster C (1990) *Information Technology in Geography and Planning: including principles of GIS.* Routledge, London

Bracken I, Higgs G, Martin D, Webster C (1989) A classification of geographical information systems literature and applications. *Concepts and Techniques in Modern Geography*, Vol. 52. Environmental Publications, Norwich

Briggs D, Mounsey H M (1989) Integrating land resource data into a European geographical information system: practicalities and problems. *Applied Geography* **9**: 5–20

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

Chrisman N R (1989) Modeling error in overlaid categorical maps. In: Goodchild M F, Gopal S (eds.) *Accuracy of Spatial Databases.* Taylor & Francis, London, pp. 21–34

Chrisman N R (1991) The error component in spatial data. In: Maguire D J, Goodchild M F, Rhind D W (eds.) *Geographical Information Systems: principles and applications.* Longman, London, pp. 165–74, Vol 1

Cutter S L (1985) Rating places: a geographer's view on quality of life. *Resource Publications in Geography.* Association of American Geographers, Washington DC

Davis F W, Simonett D S (1991) GIS and remote sensing. In: Maguire D J, Goodchild M F, Rhind D W (eds.) *Geographical Information Systems: principles and applications.* Longman, London, pp. 191–213, Vol 1

Demko G J, Hezlep W (1989) USSR: mapping the blank spots. *Focus* **39** (1): 20–1

Devereux B J (1986) The integration of cartographic data stored in raster and vector formats. In: Blakemore M J (ed.) *Proceedings of AUTOCARTO London 1.* Royal Institution of Chartered Surveyors, London pp. 257–66

Fisher P F (1991) Spatial data sources and data problems. In: Maguire D J, Goodchild M F, Rhind D W (eds.) *Geographical Information Systems: principles and applications.* Longman, London, pp. 175–89, Vol 1

Flowerdew R, Banting D (1989) Evaluating the potential role of GIS for a market analysis company. *North West Regional Research Laboratory, Research Report 2.* NWRRL, Lancaster

Flowerdew R, Green M (1989) Statistical methods for inference between incompatible zonal systems. In: Goodchild M F, Gopal S (eds.) *Accuracy of Spatial Databases.* Taylor & Francis, London, pp. 239–47

Flowerdew R, Openshaw S (1987) A review of the problems of transferring data from one set of areal units to another incompatible set. *Northern Regional Research Laboratory, Research Report 4.* NRRL, Lancaster and Newcastle-upon-Tyne

Gatrell A C (1991) Concepts of space and geographical data. In: Maguire D J, Goodchild M F, Rhind D W (eds.) *Geographical Information Systems: principles and applications.* Longman, London, pp. 119–34, Vol 1

Goodchild M F, Anselin L, Deichmann U (1989) A general framework for the spatial interpolation of socio-economic data. Paper presented at the Regional Science Association meeting, Santa Barbara California

Goodchild M F, Gopal S (1989) (eds.) *Accuracy of Spatial Databases.* Taylor & Francis, London

Goodchild M F, Lam N S-N (1980) Areal interpolation: a variant of the traditional spatial problem. *Geo-Processing* **1**: 297–312

Hearnshaw H M, Maguire D J, Worboys M F (1989) An introduction to area-based spatial units: a case study of Leicestershire *Midlands Regional Research Laboratory Research Report 1.* MRRL, Leicester

Kehris E (1990) Interfacing ARC/INFO with GLIM. *North West Regional Research Laboratory Research Report 5* NWRRL, Lancaster

Lam N S-N (1983) Spatial interpolation methods: a review. *The American Cartographer* **10**: 129–49

Langford M, Maguire D J, Unwin D J (1991) The area transform problem: estimating population using satellite imagery in a GIS framework. In: Masser I, Blakemore M J (eds.) *Geographic Information Management: methodology and applications.* Longman, London

Lodwick W A (1989) Developing confidence limits on errors of suitability analyses in geographical information systems. In: Goodchild M F, Gopal S (eds.) *Accuracy of Spatial Databases.* Taylor & Francis, London, pp. 69–78

Maffini G, Arno M, Bitterlich W (1989) Observations and comments on the generation and treatment of error in digital GIS data. In: Goodchild M F, Gopal S (eds.) *Accuracy of Spatial Databases.* Taylor & Francis, London, pp. 55–67

Maling D H (1973) *Coordinate Systems and Map Projections.* George Philip, London

Maling D H (1989) *Measurements from Maps: principles and methods of cartometry.* Pergamon, Oxford

Maling D H (1991) Coordinate systems and map projections for GIS. In: Maguire D J, Goodchild M F, Rhind D W (eds.) *Geographical Information Systems: principles and applications.* Longman, London, pp. 135–46, Vol 1

Martin D (1989) Mapping population data from zone centroid locations. *Transactions of the Institute of British Geographers.* NS **14** (1): 90–7

Mounsey H M (1991) Multisource, multinational environmental GIS: lessons learnt from CORINE. In: Maguire D J, Goodchild M F, Rhind D W (eds.)

Geographical Information Systems: principles and applications. Longman, London, pp. 185–200, Vol 2

Openshaw S (1984) The modifiable areal unit problem. *Concepts and Techniques in Modern Geography*. Vol. 38. Geo Abstracts, Norwich

Openshaw S (1989) Learning to live with errors in spatial databases. In: Goodchild M F, Gopal S (eds.) *The Accuracy of Spatial Databases*. Taylor & Francis, London, pp. 263–76

Openshaw S, Wymer C, Charlton M (1986) A geographical information and mapping system for the BBC Domesday optical disks. *Transactions of the Institute of British Geographers* NS **11**: 296–304

Rhind D W, Green N P A, Mounsey H M, Wiggins J C (1984) The integration of geographical data. *Proceedings of Austra Carto Perth*. Australian Cartographic Association, Perth, pp. 273–93

Robinson G M, Gray D A, Healey R G, Furley P A (1989) Developing a geographical information system (GIS) for agricultural development in Belize, Central America. *Applied Geography* **9**: 81–94

Schut G (1976) Review of interpolation methods for digital terrain models. *Canadian Surveyor* **30**: 389–412

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

Tobler W R, Kennedy S (1985) Smooth multidimensional interpolation. *Geographical Analysis* **17** (3): 251–7

Unwin D J (1981) *Introductory Spatial Analysis*. Methuen, London

Veregin H (1989) Error modeling for the map overlay operation. In: Goodchild M F, Gopal S (eds.) *Accuracy of Spatial Databases*. Taylor & Francis, London, pp. 3–18

Weibel R, Heller M (1991) Digital terrain modelling. In: Maguire D J, Goodchild M F, Rhind D W (eds.) *Geographical Information Systems: principles and applications*. Longman, London, pp. 269–97, Vol 1

Williams R B G (1984) *Introduction to Statistics for Geographers and Earth Scientists*. Macmillan, London

Williams R B G (1986) *Intermediate Statistics for Geographers and Earth Scientists*. Macmillan, London

Wrigley N (1985) *Categorical Data Analysis for Geographers and Environmental Scientists*. Longman, London