# 15 INTRODUCTION TO VISUALIZING DATA VALIDITY

M. GOODCHILD, B. BUTTENFIELD & J. WOOD

## ...BACKGROUND

First and foremost, effective visualization of data in GIS requires the use of an appropriate method of representation. In general, the validity of a graphical display may be ensured by application of sound principles of graphical design, or evaluated for cognitive effectiveness, and each of these options is discussed in other sections of this book. In addition, geographical data are generally incomplete, and almost always have some degree of uncertainty. Thus while much of this book is concerned with the issue of valid representation, its converse, the representation of the validity of the data, is of particular significance to spatial data. This forms the focus of this section.

Information on validity affects the reliability of all aspects of analysis, spatial inference and reasoning. It also affects the credibility attached to decisions supported by geographic information systems (GIS). Collection and maintenance of valid spatial data has long been a goal of data producers, but only recently has the incorporation of data validity become a priority. In the context of GIS, the term 'validity' encompasses concepts related to measurable validity, including, but not limited to, measurement by deductive estimates, inferential evidence, or comparison with independent sources. Measurement and evaluation may include indices and processes, and visualization of these two types will be presented in the chapters in this section.

Dictionaries define the term 'valid' using phrases such as 'sound or defensible', 'well-grounded', 'executed with proper formalities (or formalisms)' and 'legally acceptable'. A search for the word 'validate' recovers words like 'ratify', 'confirm' or 'verify', implying that validity can be codified in formalized expressions, which in the course of geographical analysis may be numerical, verbal or graphical. Validity can thus encompass both the accuracy of data and of the procedures applied to those data, and becomes especially important when basing interpretation, inference or decision-making upon GIS data. The validity of data may be ratified by testing or by illustration, or by replication. Unfortunately, GIS operations are rarely replicated and this complicates the issues raised in the chapters of this section.

The term 'data quality' encompasses more than testable elements subsumed here under validity, incorporating aspects of lineage that are often monitored or tracked in database operations. Trackable elements include chronological

reporting of data collection and processing, algorithms used to process the data, and tests applied to the evaluation. The need for tools to facilitate description, storage and representation of such information has been recognised by the US National Center for Geographic Information and Analysis (NCGIA), which in June 1991 began a research initiative 'Visualizing the Quality of Spatial Information' (Beard, Buttenfield and Clapham, 1991). Priorities for the initiative include development of frameworks linking spatial data quality components with visualization methods, implementation of prototype displays, and database management to support error tracking and to facilitate graphical depiction of error propagation during GIS operations. This section reports research in all three areas.

## ...VALIDITY IN GEOGRAPHICAL INFORMATION

By definition, reality is continuous, while the observation of reality is discrete. Technology discretises measurement, as for example in satellite image 'snapshots' taken at regular intervals in comprehensive scanning paths. Perception also occurs in discrete 'chunks', is selective and easily masked or distracted. Digital organisation of data requires that models be fitted to observations and measured phenomena, and, as each of these chapters demonstrates, the model chosen for digital data will to a large extent drive and even constrain the type of visualization that may be appropriate to its subsequent representation. Because it is not possible to represent continuous phenomena completely, they must be approximated, or sampled as subsets. In numerical analysis, this may take the form of discrete approximation formulae, while in modelling, it may occur through processes of sampling, estimation and inference. In mapping, we achieve the same fundamental objectives by processes of abstraction, selection, and generalisation.

The types of validity that may be discovered or measured in geographical data most commonly include error and accuracy. As discussed by Buttenfield and Beard in Chapter 16 these are not identical concepts. Accuracy measures the discrepancy from a modelled or an assumed value, while error measures discrepancy from the true value. One consequence of the discretisation of reality is that true values for geographical phenomena are rarely determinable. It is impossible to collect a complete dataset, so that a 'true' value may be inferred but it is not often determined with full confidence. For this reason, many standards for data quality assessment include indices of accuracy rather than error, and require explicit reporting of the models applied. A common situation in spatial analysis is the measurement of root mean square error (RMSE), which is in fact an accuracy measure of perpendicular distance using a Euclidean model of geometry as reference.

A different aspect of validity relates to replicability, or consistency, of processes and realisations. GIS operations are rarely replicated in practice. The selection of an optimal site or of an optimal route is not usually repeated hundreds of times with slight variations in the values of coefficients or input parameters to determine the sensitivity of the solution to varying inputs. In soil mapping, for example, it is widely accepted that mapped boundary lines contain an element of subjectivity and uncertainty, but rarely have many photoint-

erpretations been made of the same area to evaluate variation between interpreters, or to determine overall 'average' boundary line locations for soil parcels. Soils mapping proceeds using the educated interpretive skills of a single interpreter, using source data (photos and maps) drawn from multiple scales at many dates and with varying accuracy. When the 'true' soils boundaries cannot be determined, it is difficult to select a numeric model against which to compare the interpreted map boundaries, and determination of soils mapping accuracy remains problematic.

## ...THE REPRESENTATION OF UNCERTAINTY

Clearly, it is important to convey an appreciation of the uncertainty present in the data we use in GIS. Moreover, when using visualization as an aid to this appreciation we are actually visualizing the validity of our data. The measurable aspects of data uncertainty will probably have a spatial component, since quality will be higher in some areas than others, or higher for certain features. Graphic representation provides an ideal mechanism for the communication of this spatial variation of quality.

Although data validity may be measurable, data quality has been defined as a point in the three-dimensional space of data goodness, application and purpose (Beard, Buttenfield and Clapham, 1991). A significant consequence of this is that the data themselves do not contain all of the information required to locate data quality. The two-way interaction which is a characteristic of ViSC has enormous potential in allowing the user's knowledge of purpose to be combined with the characteristics of the data.

Despite the many reasons, if not the imperative, to consider data quality issues, many users do not consider the implications of uncertainty in their use of GIS. Visualization can act as an effective communication medium, conveying the effects of error on modelling the real world. This may perform an important role in the education of users in 'data-quality awareness'.

## ...MODELS OF UNCERTAINTY

As argued earlier, the validity of spatial data derives from many potential sources and can be subject to a bewildering array of influences. However, visual representations of validity must be controlled by well-defined sets of rules. One such set of rules frequently used in discussions of uncertainty, particularly in the context of error or inaccuracy, is provided by statistics, and is the basis for Chapters 17 and 19 by Goodchild et al. and by Fisher.

In essence, statistics provides a rigorous mathematical framework for describing and manipulating uncertainty. While the term often conjures up images of tossing dice and drawing coloured balls from urns, the application of statistical methods does not imply an assumption that uncertainty in spatial data similarly derives from such processes. Instead, the assumptions behind statistics are much broader, and relate more to the availability of information than to any mechanical conception of process. Randomness simply reflects lack of knowledge, and statistics provides a means of dealing systematically with

incomplete information. For example, suppose we are uncertain about the true position of an object because we cannot determine which datum was used when its position was measured, or which projection was used for the map from which its position was digitized. These possibilities reflect different uncertainty-causing processes. Statistics provides a means of handling uncertainty that is responsive to its form but independent of the process by which it was generated. From this viewpoint, uncertainty may be temporary, since we might later discover the missing information on datum or projection.

From this perspective, the statistical paradigm is potentially useful for all sources of uncertainty and all kinds of validity but its applications are nevertheless limited and we would not want to suggest that it is the only possible framework for visualizing validity. There are circumstances where errors are sufficiently large or unique, and where the generalizations of statistics would have no relevance. The use of statistics also implies that uncertainty is quantifiable, whereas the validity of spatial data is often recorded qualitatively.

Statistics provides a framework for the description and modelling of uncertainty, and thus for its visualization. Uncertainty is described using such concepts as standard error, or root mean square error (RMSE), by measuring the differences between observation and a model, and summarising them as a standard deviation. Commonly used quantitative error descriptors for spatial data include the Circular Map Accuracy Standard (CMAS) (Goodchild, 1991), the Perkal epsilon band (Blakemore, 1984), the misclassification matrix (Chapter 12) or the root mean square error measure used to describe the quality of digital elevation models (US Geological Survey, 1987).

These measures describe the average case, and they can be used as the basis for methods of visualization. For example, a contour map might be accompanied by a map showing the variation of the root mean square error of elevation inherent in the data used to draw it. Many applications of geostatistics (Cressie, 1991) include such displays because Kriging, a standard method of contouring or spatial interpolation, provides estimates of the uncertainties associated with its predictions. Nevertheless, displays of measures of the average uncertainty are of much less value in applications where the focus is on specific errors. In such cases it may be much more valuable to illustrate the range of possibilities, but to do so we must be able to model uncertainty in addition to describing it.

An error model is defined here as a stochastic process capable of simulating the range of possibilities known to exist in spatial data. These possibilities may exist because vital information, such as datum or map projection, is missing, or because the data represent a subjective interpretation and would therefore vary from one interpreter to another, or because measuring instruments are known to be imprecise, or for a host of other reasons. Thus rather than describe the magnitude of uncertainty, an error model permits its simulation. The best known error model is the Normal or Gaussian distribution, which simulates the possible values of a simple measurement by sampling a bell curve, and describes the magnitude of uncertainty as the distribution's standard deviation. The Circular Normal distribution, a two-dimensional version of the bell curve, is commonly used in spatial databases to describe the uncertainty in the position of a point's location, or as the basis of map accuracy standards.

In the context of spatial data, an error model provides a range of possible maps, each of which might represent the truth. The chapters in this section by

Fisher and by Goodchild et al. present simple error models for a particular class of spatial data, the chorochromatic or 'area class' type exemplified by soil or land cover maps, and use it to create a range of possible maps or realisations. Visually, the impact of one or more realisations may be much greater than the more abstract concept of an error descriptor, since the user is exposed directly to the range of possibilities available rather than to a measure of the range.

This distinction between visualization of error descriptors on the one hand and realisations of an error model on the other is illustrated somewhat dramatically by Englund (1993) in the context of Kriging estimation. When the technique of Kriging is used to create an interpolated surface between sample points of known value, the result is both a surface and a map of uncertainty. In fact the surface is the estimated mean, and the map of uncertainty shows estimated variance around the mean. Englund deviates from common practice by showing not the map of estimated means, but sample maps from the distribution of possibilities defined by the means and variances. These, rather than the estimated mean, are then used in GIS processing. As a result, Englund is able to provide visually dramatic illustrations of the uncertainty expressed by the estimated variances, but normally ignored in analyses based on estimated means.

## ...UNCERTAINTY IN THE DECISION-MAKING PROCESS

Uncertainty is an inevitable characteristic of spatial data, because it is impossible to measure positions on the Earth's surface to infinite precision, reality changes faster than maps, and a host of other common reasons. Regrettably, the traditions of map-making include remarkably few methods for displaying uncertainty. In part this is because it may not be visible at the scale of the map. But spatial databases have no scale, and force us to deal explicitly with uncertainty. In part also the lack of attention to uncertainty is attributable to the more aesthetic aspects of cartographic tradition. Uncertainty would devalue the map, and confuse its aim of communicating an interpretation of the landscape to its readers. However, in GIS, neither of these arguments is valid, and the lack of methods for visualizing uncertainty is a major drawback of systems intended to support objective analysis, scientific investigation and spatial decision-making.

We have little experience to date of spatial decision-making under uncertainty (Leung, 1988) or of the potential impact of visual displays of uncertainty on the process. Information on uncertainty is often available to map users in the form of text descriptions, but these are often published separately and may be hard to find. A visualization of uncertainty in a GIS would immediately draw the user's attention to a potentially problematic aspect of the project. Increasingly, digital maps and GIS are used for enforcement of regulations, particularly in environmental management, and direct access to the uncertainty inevitably present in the data could have dramatic effects on the regulatory process. Should maps of US wetlands derived by classifying remotely sensed images, and therefore inevitably subject to uncertainty, be used to enforce wetland regulations? Should imperfect maps of Spotted Owl habitat be used to regulate old growth logging in the Pacific Northwest?

The approach used to visualization, whether it be suppression of uncertainty, display of error descriptors, or display of alternative realisations of an error model, clearly will affect the decision-making process. Moreover its effects will differ depending on context. In scientific research, statistical methods can be used to estimate the uncertainty in a GIS product based on uncertain data. In a legal context, where a court is attempting to resolve conflict, the approach will be very different and based more on process than on demonstrable truth. A court may hold that a decision based on uncertain data is valid if it can be demonstrated that reasonable care and control was exercised over the creation and processing of the data, and provided that these meet agreed professional standards. In other words, the fact that it is impossible in principle to ensure perfect representation of the truth is not important. The court is concerned instead with whether adequate procedures were in place to ensure a reasonable level of accuracy. The role of visualization in this legal debate is far from clear.

## VISUALIZATION TOOLS FOR DISPLAYING VALIDITY

The degree to which uncertainty has been recorded will affect the way in which it can be visualized. Many spatial databases currently contain little or no embedded information on accuracy, certainty or other components of validity. Those data formats that do embed capabilities for storing metadata (for example, Vector Product Format) homogenise validity information across data layers or coverages. Spatial variation in validity may occur within data layers as a result of data collection, or may result from particular algorithmic processes applied to the data. Cressie (1991) provides an example produced as a by-product of Kriging.

In many cases, graphical displays of data have the potential to reveal the variation in spatial or temporal accuracy. Map animation should be of utility for cleaning any dataset characterised by high serial correlation. Spatial inconsistencies may also be detected efficiently by visual inspection, when graphical methods are chosen to bring inconsistencies to the attention of the viewer.

Propagation of error during GIS operations is expected, although specific models for monitoring and compensation have not been developed in all data domains (but see Fisher, 1991). In some cases, combining separate but interrelated data sources can reveal inconsistencies between them. For example, overlay of digitised vector transport networks over a rectified satellite image might point out conflation in the vector layer caused by digitising from multiple map sources. Of course, relative discrepancies do not indicate which data source contains inaccuracies in every case. And conversely, failure to detect errors by visual inspection does not indicate a clear absence of uncertainty.

Two modes of visualization can be applied to display the validity of GIS data. Traditional cartographic displays provide static snapshots of data surfaces and objects, affording visualization at specific or selected phases of data processing. For example, visual displays might be used to determine drifts and trends during iterative modelling and simulation. Dynamic visualization tools are emerging as an alternative mode for cartographic data display, and may also prove worthwhile for viewing and exploring information on data validity.

In the static mode, cartographic conventions are guided by application of Bertin's (1985) visual variables, including size, shape, texture, value, colour, orientation and arrangement. The human information processing system has strongest acuity for distinguishing symbol size, value (relative darkness or gray-tone) and colour (hue). With respect to validity, the visual variables seeming to offer best opportunities for display include value, colour (saturation or intensity) and texture. The visual metaphor of fog has been proposed (Beard, Buttenfield and Clapham, 1991) to alert viewers to uncertainty in data position or attribution. Fog may be created graphically by manipulating value or by de-saturating colours. MacEachren (Chapter 13) cites Woodward's idea of de-focusing a display for visual impact, which is an earlier presentation of a similar concept. Other visual variables such as texture may be used to indicate invalid data. For example, a coarse-texture screen may introduce visual noise as a metaphor for noisy data. The use of colour in graphical display is recognised in empirical research to be especially complicated (Brewer, 1989). Recent empirical research at NCGIA indicates that viewer recognition of the concept 'uncertainty' is not tied to linear progressions of either value or saturation in the RGB colour cube.

The visual variable Bertin refers to as 'arrangement' incorporates two-dimensional position and apparent three-dimensional position, simulated in a CRT display by means of hidden surface and hidden line removal, shading and size variation (distant objects appear to be smaller). The arrangement of objects in a display may be used equally well to imply logical inconsistencies in a data structure. For example, polygon fill can be used to check efficiently for closure of digitised polygons. Where topological inconsistencies occur, the fill tone or pattern will not match the polygon boundaries. Other graphical methods for strengthening or deconstructing the holism of image objects may also be applied, relying on many principles of Gestalt psychology. Viewers try to make sense of nonsensical data, interpreting compound geometric relations by visual isolation of regular shapes, and searching for familiar patterns in ambiguous visual fields. To make sense of an image, viewers will tend to formulate a mental model of what that image represents, and make comparisons to interpret the image. The greater the level of image abstraction, the more difficult the comparison. Application of visual variables to the display of abstractions about data validity probably requires development of skills beyond traditional map use exercises in current GIS curricula.

The alternative to static modes of display is called dynamic cartography by some, map animation by others, and is discussed in detail earlier in this book. Dynamic displays provide several advantages to viewers and designers of GIS displays, particularly in the context of validity and metadata. One can imagine a data sequence in which data and metadata displays are toggled, either automatically or by viewer choice; at some speeds, the data quality may appear to become embedded in the data display, while other speeds will isolate the two images. Another advantage is that for positional accuracy, symbol movement overcomes the problem of viewers attaching too much weight to particular cartographic representations. Fisher (Chapter 19) demonstrates this principle for soils maps.

Advances in display refresh speeds and CPU speed have greatly improved functionality for interactive graphics, and this can be especially useful in the

context of data validity. Viewers can use point-and-click interface functions to query specific map objects in a display. One can imagine a hypermedia image in which clicking on a map object (a soil polygon, for example) will bring up a dialogue box, statistical chart or tabular display of lineage information to alert viewers to the date of data collection, scale of the data source, types of processing algorithms applied to generalise and symbolise the map image, and so forth. These capabilities lie somewhat beyond the scope of the current chapters, but there is every reason to expect that these types of functions can and will begin to be incorporated soon in upcoming versions of GIS software.

## ••• SUMMARY

This introduction has offered three levels on which one might approach issues and display of data validity: first, the development of frameworks linking spatial data quality components with visualization methods; second, the implementation of prototype displays; and third, error tracking and monitoring of error propagation during GIS operations. These interrelated concepts may be incorporated into a variety of measures of validity, or mapped using a variety of illustrative techniques. In Chapter 16, they are generalised into a triad of concepts: accuracy, resolution and consistency. Chapters 17 and 19 discuss multiple stochastic realisations of soils data and remotely sensed imagery, and present applications graphically. Chapter 18 by Jo Wood presents several methods to illustrate the validity of digital terrain data.

Research issues that remain to be addressed include refined methods for estimating error propagation rates, particularly in environmental monitoring where policy decisions are often based upon visual display, such as wetlands protection and climatic change studies. Additionally, the potential for cognitive overload in displays where data and data validity are embedded forms an important area for evaluation. Finally, the application of innovative graphical techniques to displays of validity has only recently begun to be addressed. Animation, multimedia and hypermedia may expand GIS viewer capabilities for exploration of data pattern and data validity. Considerations for design and evaluation of these media remain to be uncovered in further research.

## ... KEY READING FOR THIS SECTION

Overviews of the data quality issue in spatial data can be found in M.F. Goodchild and S. Gopal (1989) *The Accuracy of Spatial Databases* (London: Taylor and Francis); in the chapter by Goodchild (1991b) in J C Muller (ed.), *Advances in Cartography* (New York: Elsevier, pp. 113–40); or in the chapters by Fisher and by Chrisman in D.J. Maguire, M.F. Goodchild and D.W. Rhind (eds) (1991), *Geographical Information Systems: Principles and Applications* (Harlow: Longman).

A useful discussion of the accuracy of map analysis techniques can be found in D.H. Maling (1989) *Measurements from Maps* (Oxford and New York: Pergamon). The Beard, Buttenfield and Clapham report listed in the references

is a comprehensive overview of visualization of spatial data quality, and includes a prioritised research agenda.

## ... REFERENCES


Beard M K, Buttenfield B P, Clapham S (1991) *Visualizing the Quality of Spatial Information: Scientific Report of the Specialist Meeting*. National Center for Geographic Information and Analysis Technical Report 91–26, Santa Barbara, Calif.

Bertin J (1985) *Graphical Semiology*, University of Wisconsin Press, Madison, Wisconsin

Blakemore M (1984) Generalization and error in spatial databases. *Cartographica* 21(2–3): 131–139

Brewer C (1989) The development of process-printed Munsell charts for selecting map colors. *The American Cartographer* 16(4): 269–278

Cressie N A C (1991) *Statistics for Spatial Data*, Wiley, New York

Englund E (1993) Spatial simulation: environmental applications. In: Goodchild M F, Parks B O, Steyaert L T (eds) *Environmental Modeling with GIS*, Oxford University Press, New York

Fisher P F (1991a) First experiments in viewshed uncertainty: the accuracy of the viewshed area. *Photogrammetric Engineering and Remote Sensing* 57(10): 1321–1327

Goodchild M F (1991) Issues of quality and uncertainty. In: Muller J-C (ed.) *Advances in Cartography*, Elsevier, New York, pp. 113–140

Leung Y (1988) *Spatial Analysis and Planning under Imprecision*, Elsevier, New York

US Geological Survey (1987) *Digital Elevation Models – Data Users Guide*, US Geological Survey, Reston, Virginia