15

Detecting and evaluating errors by graphical methods

M K BEARD AND B P BUTTENFIELD

Both uncertainty and errors are inherent in spatial databases. The processes of observing, measuring, interpreting, classifying, and analysing data give rise to systematic and random errors. Some errors may be quite large (blunders) and easily detectable. Other errors and uncertainties in spatial data are more subtle and are not easily detected or evaluated. Casual users of GIS may not be aware of their presence or even the possibility of their existence. These are the most problematic and the ones we must try hardest to illuminate. Graphical methods in conjunction with error analysis provide a means for identifying both gross and subtle errors and evaluating the uncertainty in geographical data. This chapter outlines a rationale for the use of graphical methods, highlights several historical and recent examples, develops a framework linking error analysis and graphical methods, and points to research challenges for the future and the potential for new techniques arising from technical innovations.

1 INTRODUCTION

Several other chapters in this volume discuss error and uncertainty in spatial databases and the importance of making these known to GIS users (Fisher, Chapter 13; Heuvelink, Chapter 14; Hunter, Chapter 45; Veregin, Chapter 12). This chapter focuses on revealing error in geographical data and GIS by graphical means, used in conjunction with error analysis methods.

1.1 Rationale for graphical methods

Almost 200 years ago William Playfair began the serious use of graphs for looking at data. Many of the same issues which motivated Playfair to develop graphical solutions then are present today. Graphical methods for error detection and evaluation are motivated by several factors including physiological, technical, and institutional concerns. First, it is generally accepted that the human information processing system has strong acuity for visualisation and an exceptional ability to recognise structure and relationships. Representing information in a form that matches our perceptual capabilities (mainly visual) makes the process of getting information and digesting it easier and more effective (Gershon and Brown 1996). Second, spatial structure is more easily expressed and grasped through graphic or cartographic representation. Third, graphical methods are a fast communication channel and one capable of carrying high volumes. These characteristics make graphical methods highly suitable for human comprehension of the complex, multidimensional aspects of spatial data quality.

In terms of technical motivations, the growing interest in digital libraries and the National Spatial Data Infrastructure (NSDI) provide new impetus for documenting spatial information reliability. More and more spatial data and geographical information processing resources are rapidly becoming accessible over the Internet. The implications of this for data evaluation are profound. Intelligent use of such resources requires a substantial investment in metadata (data describing the data; Guptill, Chapter 49) and the availability of sense-making tools to digest large volumes of metadata and data. Graphical methods may provide the most efficient means to evaluate the quality of large volumes of geographic data as they become available through digital libraries and NSDI.

Institutional motivations centre around several national and international standards efforts. The US standard (the Spatial Data Transfer Standard or SDTS; Morrison 1992) categorised an initial set of data quality components, and recent completion of a Metadata Content Standard (FGDC 1995) extended and expanded representation guidelines for various categories of spatial data error. Adoption of standards similar to SDTS in other nations demonstrates international recognition of the complexities encountered in detecting and managing errors. In France for example, the effort generated MEGRIN standards (Salgé, Chapter 50; Salgé et al 1992).

1.2 Limitations of graphical methods

Graphical methods are not always either an effective solution or a substitute for conventional numerical analytical tools. Some researchers have suggested that graphic design for data analysis and presentation is still largely unscientific (Cox 1978). Graphical methods like other communication channels are open to misinterpretation. The cartographic literature is full of examples of techniques and their possible misinterpretations (Monmonier 1991; Robinson et al 1985). As MacEachren (1994) suggests, data exploration tools allow us to identify potentially meaningful patterns that we might otherwise miss, but these tools cannot always determine the probability that the pattern we see is real. Despite limitations in graphic methods we can find several compelling examples of effective use of graphical techniques.

2 EXAMPLES OF GRAPHICAL METHODS

This section reviews several examples of graphic techniques used to detect, evaluate, and display errors. Several disciplines have contributed to these developments including cartography, spatial statistics, statistical graphics, scientific visualisation, and spatial error modelling.

2.1 Graphical methods in statistics

Tukey (1977) was the force behind exploratory data analysis (EDA) and many of the well-known graphical methods for exploring data. These methods highlight unusual values which may in fact

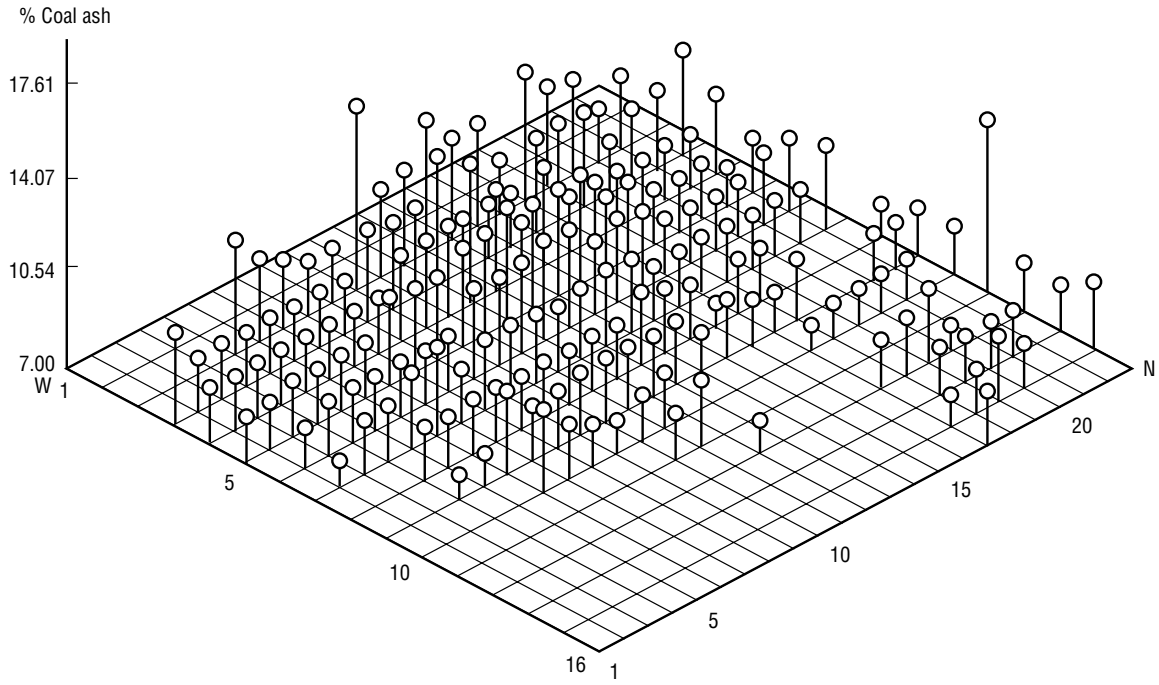

Fig 1. Three-dimensional scatter plot of core measurements of per cent coal ash.
Source: Cressie 1991

be errors. Tukey's work has been carried on and expanded by others (Becker et al 1987; Chambers et al 1983; Cleveland 1993). The shortcoming of most of these aspatial methods is that they do not consider spatial dependencies and as a result do not detect values which may be unusual in a spatial context. Cressie (1991) identifies some EDA methods which overcome this limitation (see Figures 1 and 2). These methods do not provide absolute indication of error but highlight potentially suspect values. Anselin (Chapter 17) describes exploratory methods for spatial data in more detail.

2.2 Graphical methods in cartography

The traditions of map making have included remarkably few methods for displaying uncertainty. Reliability diagrams were an early attempt to display variation in source documents used to compile maps (Wright 1942). More theoretical treatments were applied to projection distortion (Imhof 1964; Maling 1973; Tissot 1881).

2.3 Graphical methods related to GIS

Graphical means for detecting errors are now being explored and implemented in GIS software. Many recent examples illustrate expansion of Bertin's (1983) graphical framework (visual variables). New visual variables have been explored, including defocusing of features (MacEachren 1994; McGranaghan 1993) and development of multivariate symbols (Hancock 1993). Much of the recent change can be attributed to the emergence of new visualisation technologies (voxel-based 'true' 3-dimensional displays, animation, hypermedia).

Specific examples of tools developed for integration with GIS include those implemented by Fisher (1994a, 1994b), Goodchild et al (1994), Hunter and Goodchild (1995), MacEachren et al (1993), Mitasova et al (1995), and Paradis and Beard (1994). MacEachren et al (1993) developed a reliability visualisation tool (RVIS) which supports several options for viewing data and metadata (reliability). Reliability estimates in this system are based on Kriging residuals and cross validation. The display options

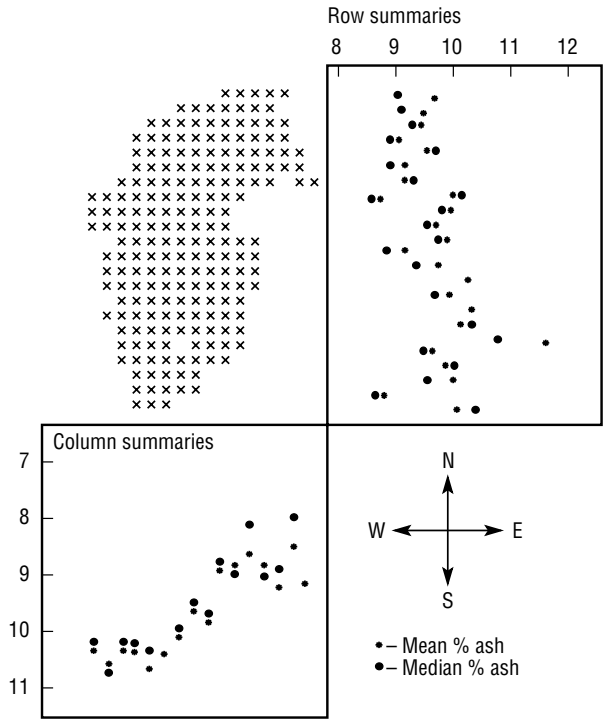

Fig 2. Comparison of mean and median summaries of non-stationarity. Units are in per cent coal ash. Comparison of mean and mean summaries for rows and columns can highlight atypical observations.
Source: Cressie 1991

include side-by-side, overlay and merged displays. The merged displays make use of several different bivariate mapping schemes. RVIS also includes a focusing tool which allows a user to specify interactively a subset of the data for further analysis. By interacting with a slider bar, an analyst can select a data range and reliability range to be displayed (Figure 3).

Fisher developed a technique referred to as error animation to view the reliability of classified imagery (Fisher 1994b) and soil maps (Fisher 1994c). For classified imagery the technique uses over-all, producer, and user accuracies as a foundation. The uncertainty inherent in the assignment of a pixel to a class is conveyed by making the value or colour of a pixel proportionate to the strength of it belonging to a particular class. In the case of soil maps the process uses randomisation to display either primary soil type or an inclusion at any particular pixel location on the map. The changing and random location of the inclusions is meant to convey the impression that the location of inclusions is not precisely known.

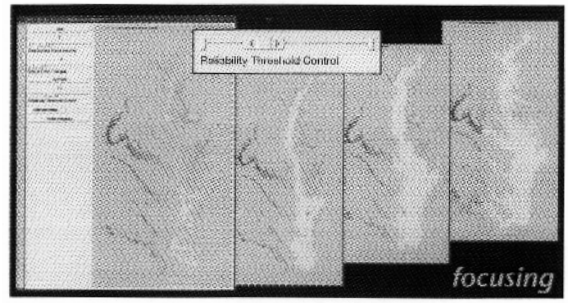

Fig 3. An implementation of the operation of focusing. The set of images illustrates the use of threshold settings for data reliability.
Source: MacEachren et al 1993

Goodchild et al (1992) use a fuzzy classifier as a foundation for creating multinomial probability fields. The fuzzy class memberships become

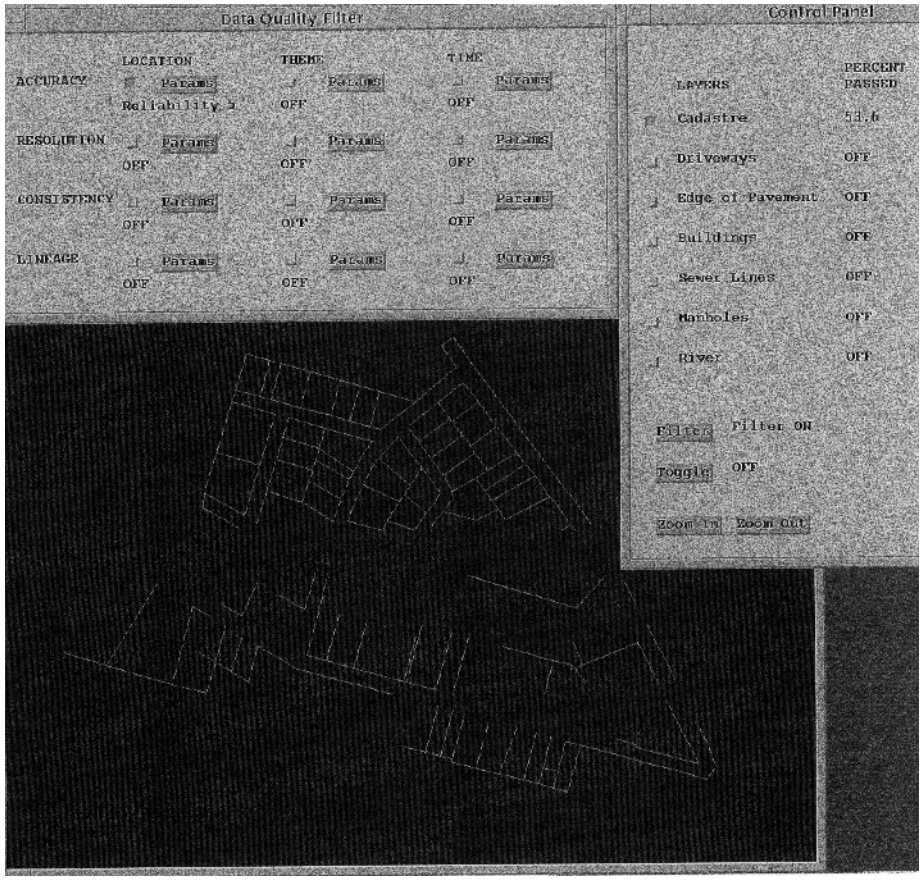

Fig 4. An implementation of the data quality filter showing parcels filtered for positional accuracy. Lines not shown indicate those not meeting the threshold set by the user.

parameters of an error model, and the range of possibilities is defined by realisations of the model. A display of realisations of the error model can inform users of the potential variation.

The data quality filter (Paradis and Beard 1994) allows a user to specify a data quality parameter (e.g. positional accuracy), a quality measure (e.g. root mean square error, RMSE) and a threshold value. The filter is applied to the data and only data meeting this threshold are displayed (Figure 4). Users are informed of how many data did not meet their specified threshold and they can toggle the display to see data which did not meet this cut-off level (Figure 5). Users can interactively adjust the threshold value in order to experiment with data that satisfy higher or lower constraints.

Hunter and Goodchild (1995) describe a probability mapping approach for representing the uncertainty of the horizontal position of a nominated terrain elevation value. They compute the probability of a cell exceeding or being exceeded by a threshold value. Once the probability for a cell has been computed it can be displayed according to different schemes. Plate 9 displays a colour ramp which shows the probability of a cell exceeding the chosen 350-metre elevation.

The visualisation tools developed by Mitasova et al (1995) incorporate multidimensional interpolation, visualisation of the resulting model, and predictive accuracy of model results using cross-validation. The visualisation tools allow cross-validation error to be viewed separately from the data, at a single depth, across different depths, at different times (Plate 10) and in combination with the data (Plate 11). These tools have been incorporated in the GRASS GIS software system. This work includes several novel visualisations but points out the many difficulties in trying to visualise reliability and data simultaneously particularly for 3-dimensional data representations (Plate 12).

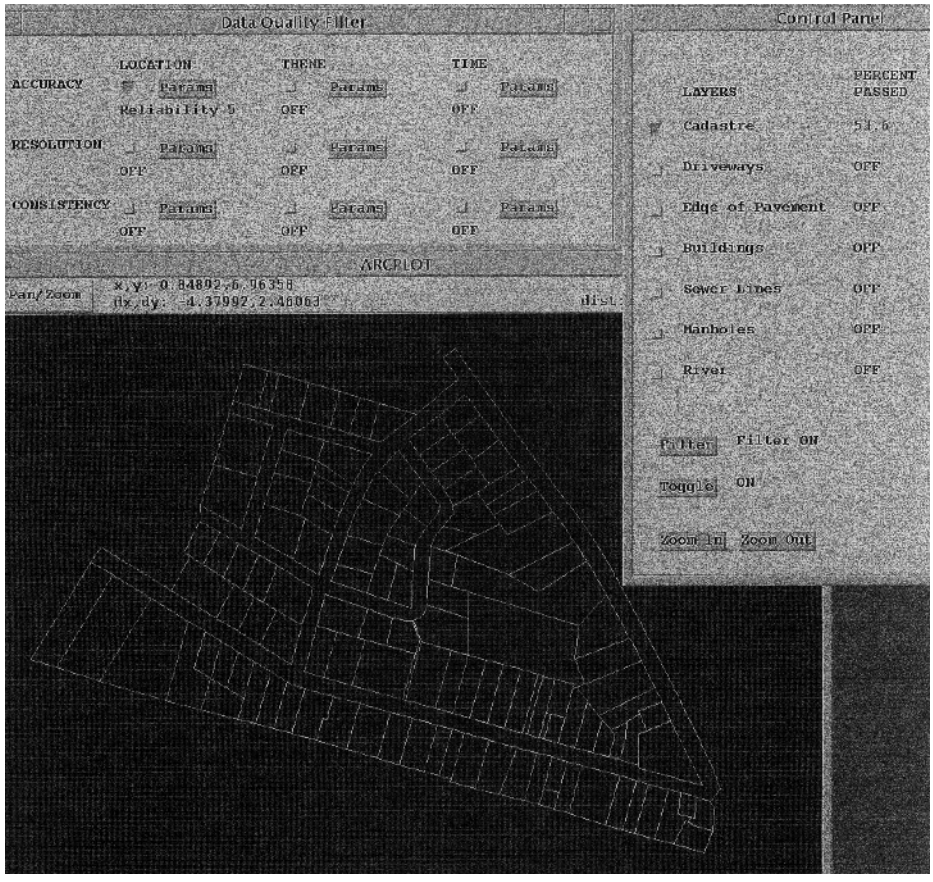

Fig 5. The data quality filter with the toggle turned on.

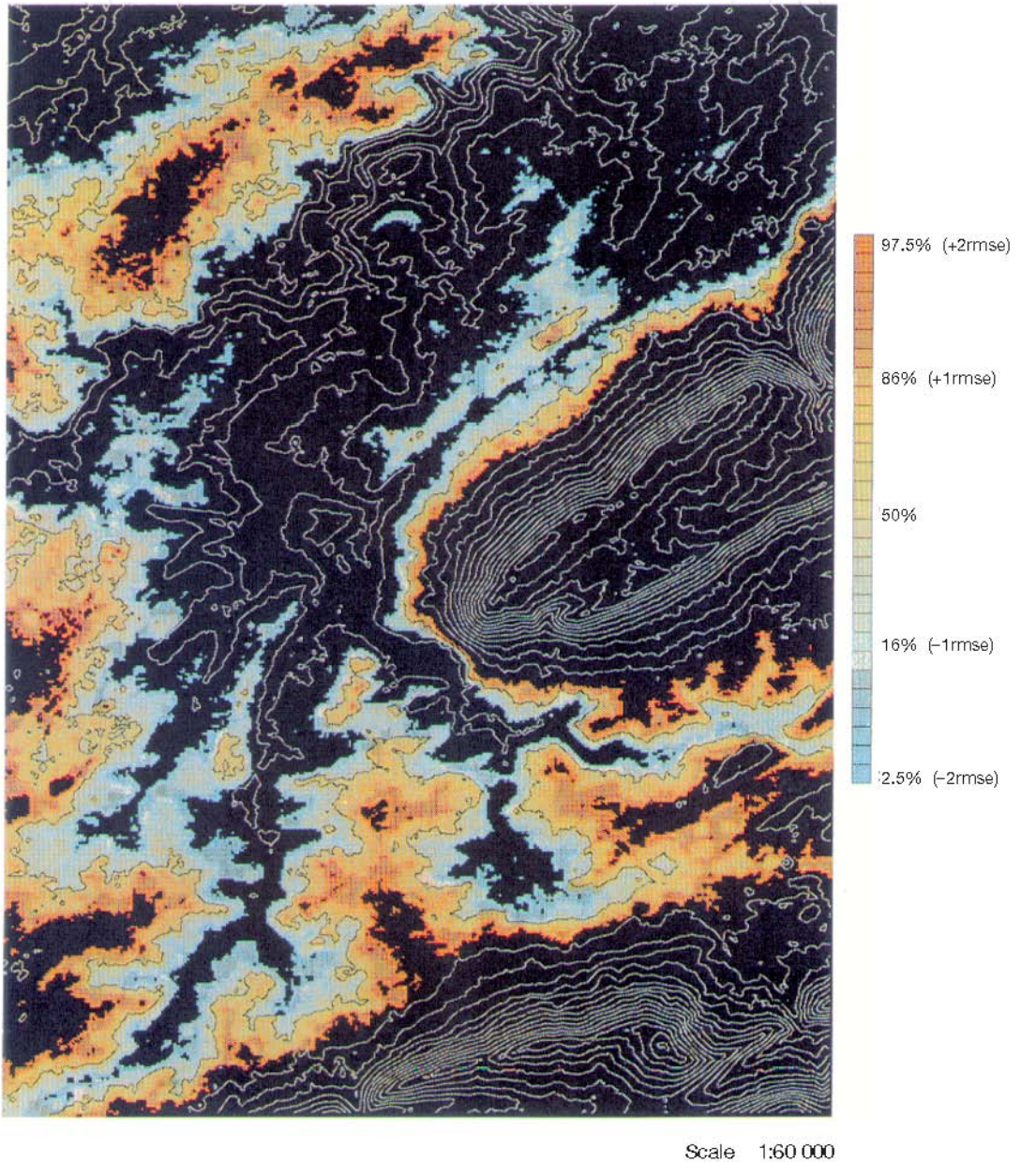

Plate 9 Graphic depiction of the error in a single-elevation value.
(*Source*: Hunter and Goodchild 1995a)

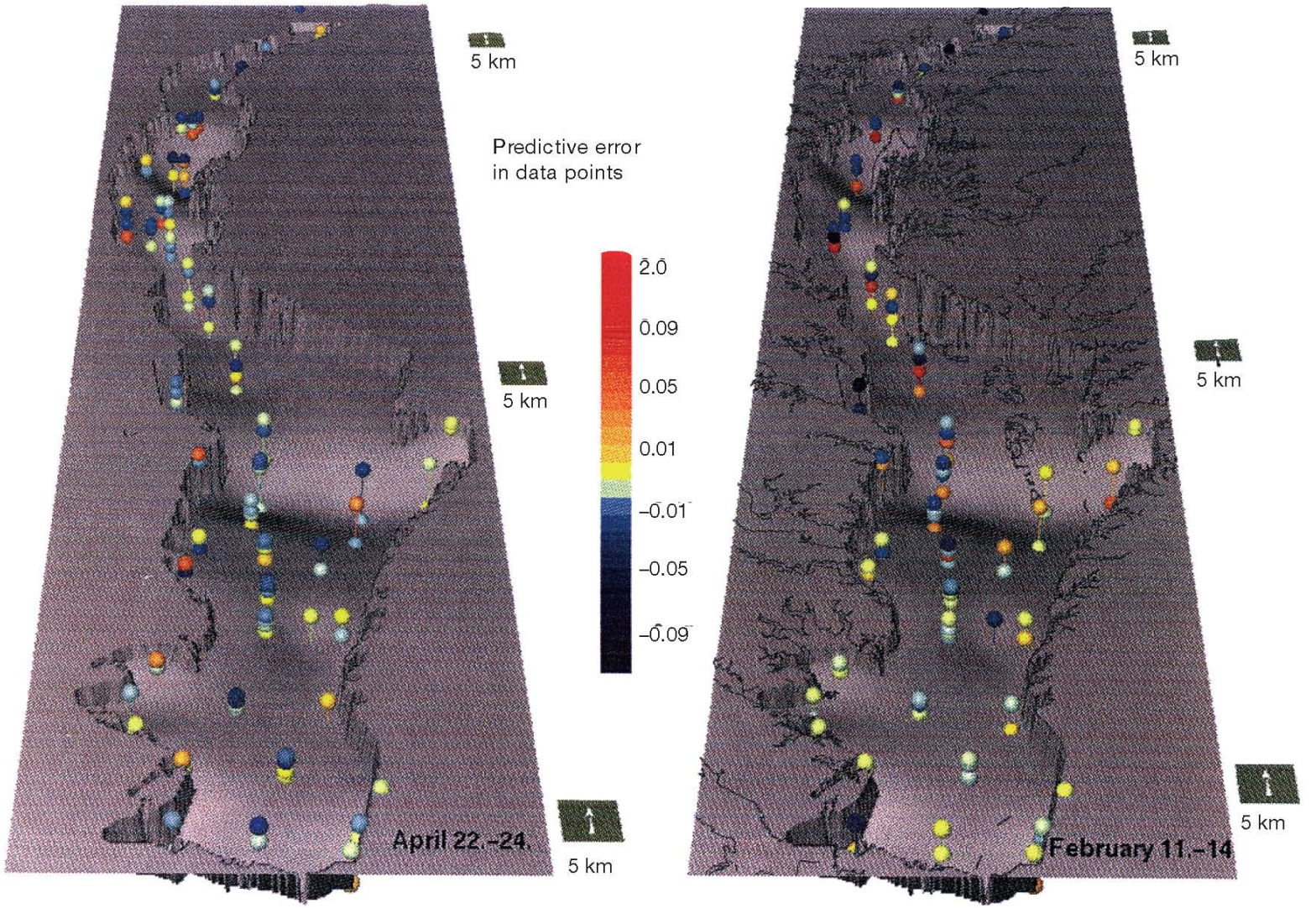

Plate 10 The cross-validation error shown separately from the data in a side-by-side display for two different periods. The cross-validation error is displayed as glyphs (coloured balls on pins) with the colour of each ball representing the error at sampled depths.

(*Source*: Mitasova et al 1995)

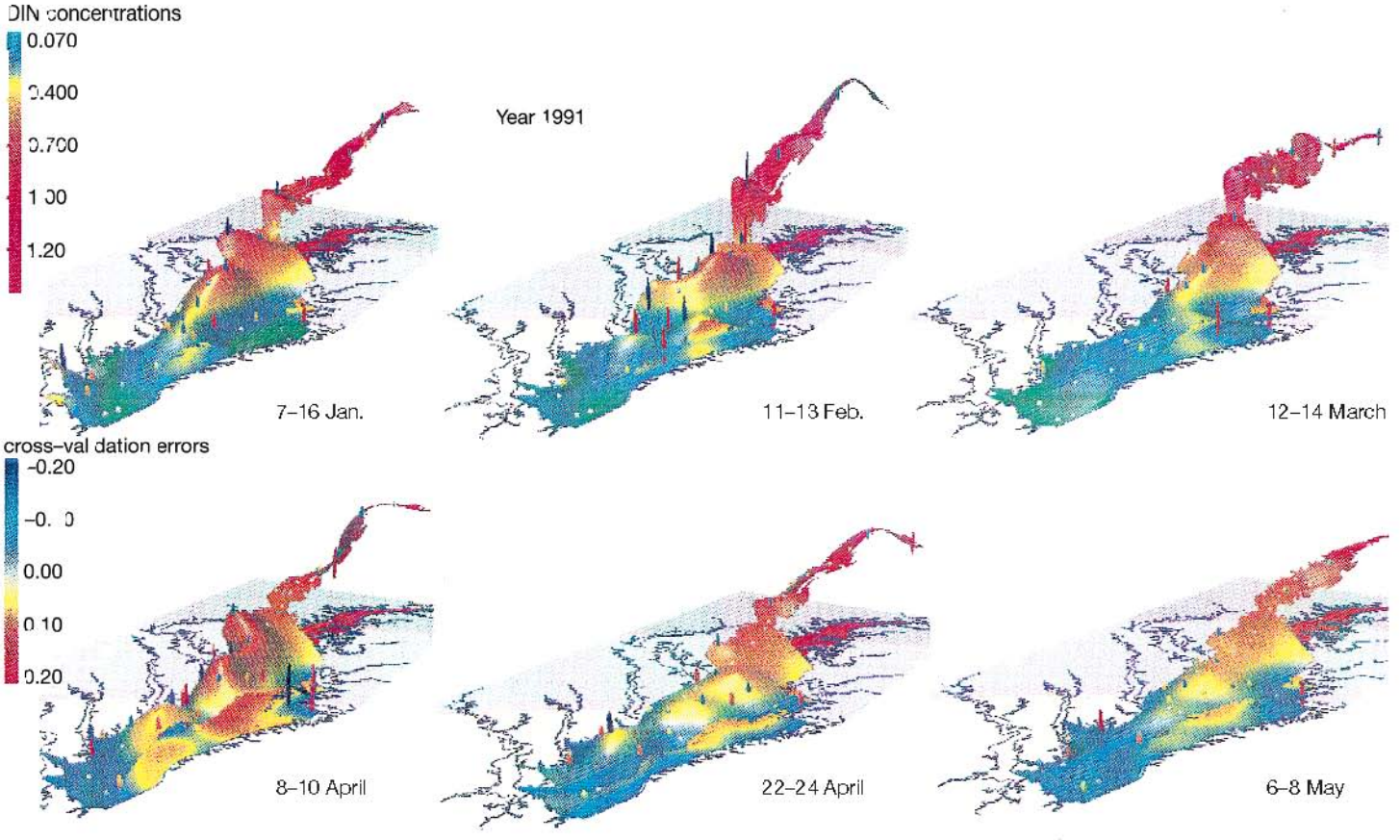

Plate 11

Images showing estimated surfaces of dissolved inorganic nitrogen for two time periods with cross-validation error at each observation station. Cross-validation error is displayed as a glyph. The height and colour of the glyph indicate the value of the cross-validation error.

(*Source*: Mitasova et al 1995)

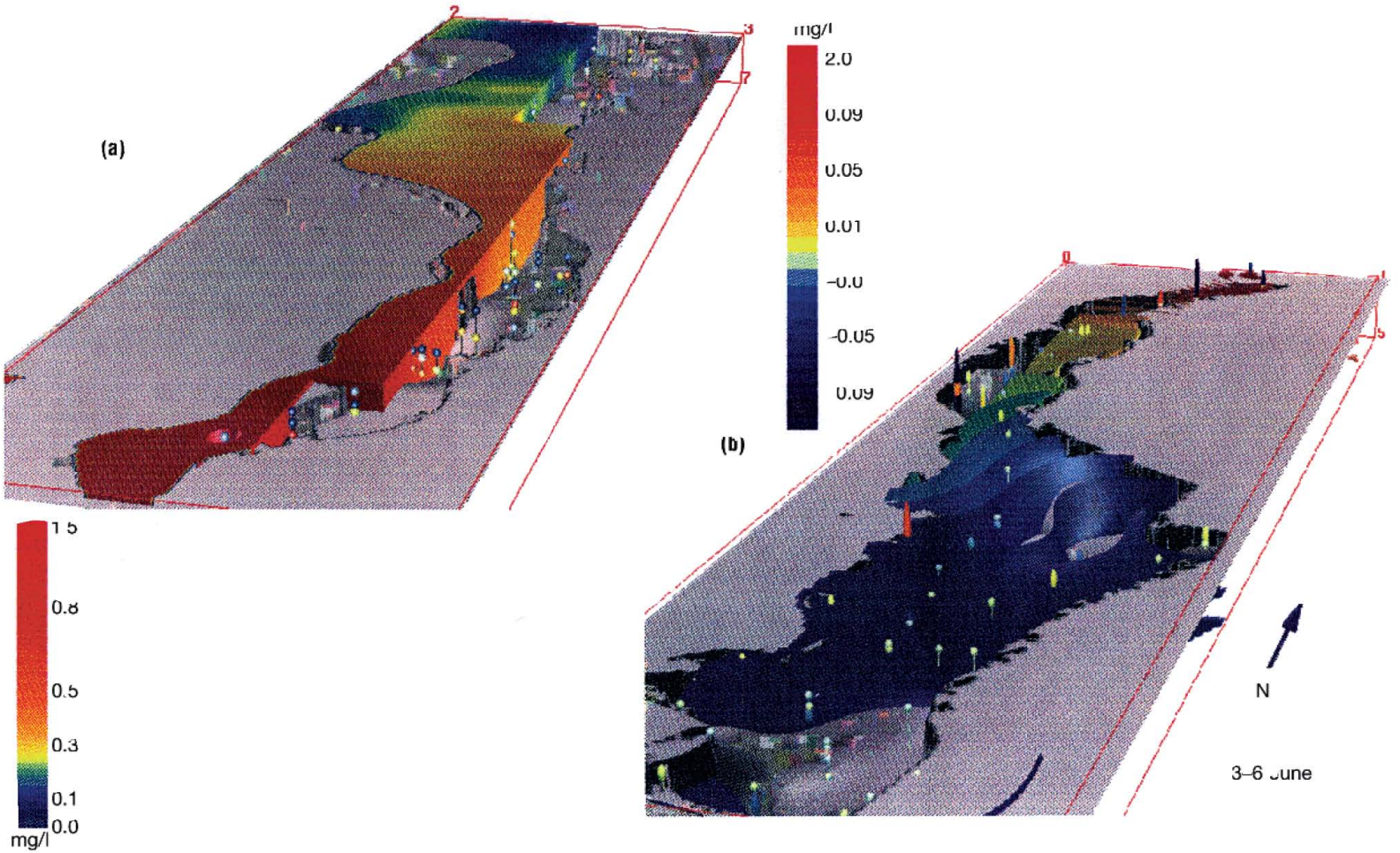

Plate 12

Images showing combinations of data display with error for three dimensions: (a) vertical cuts through the data to display predicted values along the cut in combination with glyphs showing prediction error; (b) isosurfaces in combination with the glyphs showing cross-validation error.

(*Source*: Mitasova et al 1995)

3 CHALLENGES IN GRAPHIC ERROR DETECTION AND EVALUATION

Visualisation of data is a 'demanding display problem' (Robertson 1991; see also Kraak, Chapter 11) which becomes even more demanding when the display must also include error and uncertainty and address the special characteristics of spatial data. The challenges include: (1) graphic design issues; (2) metadata issues; (3) error analysis issues; and (4) user satisfaction issues.

3.1 Graphic design issues

Graphic detection and evaluation of spatial data error and uncertainty create particular problems for graphic design. Firstly, they require a representation of space or linkage of aspatial displays to a spatial representation (Monmonier 1989) so users can be informed of the spatial distribution of errors or uncertainty. Spatial displays provide users with information on whether errors are regular, random, or clustered in space and may help users to comprehend the cause of errors. Depending on the dimensionality of the data the spatial representation may be 2- or 3-dimensional. Two-dimensional displays restrict views of the full 3-dimensional space but 3-dimensional displays add substantial cognitive and computational costs.

Second, graphic displays need to allow for both implicit and explicit displays of uncertainty. Uncertainty can be conveyed implicitly with visual variables which suggest uncertainty (e.g. fog, unfocused displays, unsaturated colours: McGranaghan 1993). The alternative is to quantify and display the uncertainty explicitly. Quantification of the uncertainty requires a solid understanding arrived at through error analysis. The graphic challenge lies in communicating the quantity visually. Cleveland and McGill (1984) provide a very helpful theoretical and empirical presentation of elementary perceptual tasks and their accuracy in judging quantitative values. For the most part visual variables convey only nominal and ordinal information. However interaction with a graphic can provide access to numerical values. A maxim of Tufte (1983)

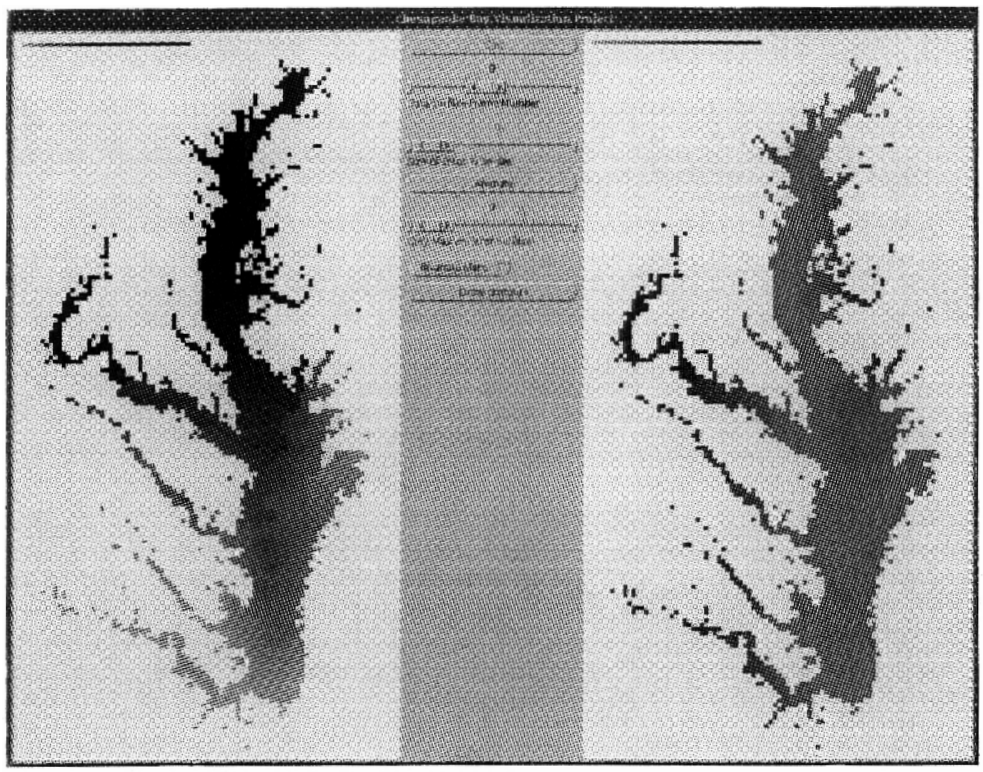

Fig 6. Side-by-side displays showing the same saturation/value range for both data and reliability. The representations for data and reliability only differ in hue and thus can be compared efficiently.

is that text and images can and should be combined. Thus we should not hesitate to incorporate numbers in displays when they may be appropriate.

Third, the graphic display should allow a data distribution and its reliability to be displayed independently or jointly. Complex graphic design issues arise in trying to display data and reliability together so users can observe correlations in the patterns. MacEachren (1994) offers three possibilities for joint display of data and reliability: (1) side-by-side images; (2) composite images; and (3) sequenced images. Each of these options presents its own graphic challenges.

In side-by-side displays the viewer must interpret two images simultaneously. Bertin's (1983) image theory applies in that we need to reduce the number of variables the viewer must process in relating the two images. This suggests that the image representing the data closely matches the image representing the data reliability. The images should be the same size and have the same coordinate scales. In addition, operations in side-by-side images should be linked. A zoom or pan operation in one image should be reflected in the other. MacEachren et al (1993) address some of these issues in their RVIS system. After experimenting with using value to represent the data and saturation to represent reliability they determined that an identical combined value/saturation scale for both data and reliability with dark equal to 'more' was the most effective representation (see Figure 6).

The composite mode of displaying data and reliability together requires overlay of contrasting visual variables, bivariate or multivariate mapping. Bertin (1983) proposes that discrimination of multiple variables in overlaid distributions can be maximised by expressing different data variables with symbols of different dimensions (point, line, area). Mitasova et al (1995) and MacEachren et al (1993) adopt Bertin's strategy by using point symbols to display reliability estimates in combination with coloured surface representations of data values. Brewer (1994) offers several useful strategies for effective use of colour in bivariate maps (see Plate 13).

The third method displays data and reliability images in sequence. When images are sequenced other considerations come into play such as the interval of time between display of the alternating images. The visual frame of reference must also remain constant between images so user attention can be devoted to the changes in the data.

Other options include linked displays and multiple version displays (Anselin, Chapter 17). For disjoint but linked images (Monmonier 1989) there must be common visual cues for the same variable in different contexts. In geographical brushing, a data value in one window is highlighted in the same colour as its representation in another view in order that users may pair up the representations. In multiple version displays we do not have separate visual variables or separate depictions for data and reliability. Instead we display multiple realisations which by their differences indicate a range of uncertainty in the data. They can be displayed as small multiples as described by Tufte (1983), or sequenced using animation (Dibiase et al 1992). Examples of geostatistical simulations are shown by Englund (1993). Uncertainty in this case is expressed implicitly by showing the range of possible variations.

Iteration (Cleveland 1993) is a particularly useful concept for viewing errors or uncertainty and it can serve several purposes. Displaying data in combination with error and uncertainty can quickly exhaust the available graphic variables. Very complex symbols can be designed (for an example see McGranaghan 1993) but these can easily become too complex and detract from interpretation of the intended message. The objective is to create more, but less complex, displays.

Since visual displays themselves can generate misleading information several iterations of a display can help to convey the uncertainty that has arisen out of map design decisions. MacEachren (1994) recommends generating multiple views rather than trying to pick a single best method of representation. In this case the iterations are over changes in visual variables, class breaks, and other map design components rather than the data. This is in contrast to viewing several simulated results in which the underlying data distribution changes, as discussed above. Viewers should be aware of possibilities for both. The challenge here is to create and display several graphic images quickly without the user losing track of his or her goal.

3.2 Metadata issues

Spatial data are frequently poorly documented. Information on how the data were collected, the sampling design, and whether any compilation or processing steps were performed on the data is usually minimal or missing. This is a serious problem for error analysis as without this information there is

Plate 13
This image shows a bivariate mapping scheme from RVIS that uses saturation to represent uncertainty and lightness to represent the data.
(*Source*: MacEachran et al 1993)

little basis on which to proceed. Fortunately the standards efforts mentioned above are having some impact and geographic datasets are appearing with more documentation (and see Salgé, Chapter 50). Several issues remain on how such metadata can be effectively stored with the data and maintained as the data evolve (see Goodchild and Longley, Chapter 40; Guptill, Chapter 49).

3.3 Error analysis issues

In most cases errors and uncertainty in spatial data are not simply detected by displaying the raw data (although examples of this are possible). As Tufte (1983) points out, graphics gather their power from content and interpretation beyond the immediate display of numbers. Thus good graphic design and, by association, effective detection and evaluation are highly dependent on effective error analysis.

Simply plotting the data can work as an error detection device because we often have some expectation about the pattern we will see. Deviations from this pattern suggest errors. Thus we can say that all error detection requires some model or reference framework, either implicit or explicit, from which departures can be determined. These may include (1) a known or postulated distribution for a set of observations; (2) a hypothesised or assumed relationship; (3) an expected set or range of values; or (4) an independent (and more accurate) set of observations. These models and frameworks can range from simple and inexpensive to complex and expensive.

Statistics provide one framework for describing and modelling error and uncertainty and thus for visualising it (Fisher, Chapter 13; Goodchild et al 1994). In standard statistics, errors and their significance are characterised by their distance from the central trend in values. This has some limitations as a method for detecting errors since outliers may in fact be unusual values and not errors. However statistical methods make clear that in order for detection to be possible we must first establish some expected distribution for values. In the case of spatial data we can add departures from assumed stationarity of mean or stationarity of dependence as the basis for detection of possible errors. For example, we should be suspicious of observations when they are unusual with respect to their neighbours.

In the case of raw data a range of exploratory techniques can be applied to identify outliers, detect blunders, and perform preliminary identification of data structure and statistical properties. Exploratory techniques are most appropriate where observational data are not obtained by formal means, where measures are not very precise (often made on nominal or ordinal scales), or where real repetition is neither feasible nor practical (Haining 1990). These are common characteristics of many spatial datasets. Cressie (1993) outlines some exploratory techniques for spatial data.

Detection can also rely on establishment of a set of consistency rules – rules indicating ranges of expected values or expected relationships between values. Many GIS apply a set of topological rules deriving from a map model as the basis for automated geometric error detection. Topological rules such as the requirement that all chains begin and end with a node, or that all polygons must close, are applied against the data and any geometric configurations which deviate from these rules are flagged. Most GIS editing packages support graphic highlighting of these inconsistencies for their easy visual detection as well as display of their spatial distribution. Bicking and Beard (1995) describe a formalisation for detection of inconsistencies in attribute data by analysing symbol encoding and symbol relationships.

Other error detection methods require ground truth data or other sources of higher accuracy for their computation (e.g. RMSE: Fisher, Chapter 13; Hutchinson and Gallant, Chapter 9; Veregin, Chapter 12). Root mean square error measures the error between a mapped point and a measured ground position. A limitation of this measure is that the error standard deviation is spatially invariant. There is no information about variation in positional accuracy at individual points. For survey measurements additional information contained as redundancies within the survey network can allow for computation of positional error for individual points within the network. Comprehensive ground checks are expensive, however.

Error and uncertainty in spatial data are not static. New error and uncertainty can occur as data are subjected to geographical information processing operations. Detection is thus an ongoing process which should be continually informed by metadata. Ideally, processes applied to the data should be known to utilise a specific graphic technique. For

example, before using Tissot's indicatrix to evaluate projection distortion, we must first know what projection was originally used. If this information is missing, then other techniques must be used.

In the more frequent case where processes are unknown, simulations can be applied to generate information for graphic display. Several researchers have looked at simulating error through geostatistical (Englund 1996), Monte Carlo (Openshaw et al 1991), or similar techniques (Heuvelink, Chapter 14; Goodchild et al 1992) to generate multiple realisations. The set of realisations generated by the simulation provides a distribution from which we can compute a variance and confidence limits. These simulations can be quite computationally demanding.

In the evaluation of errors or uncertainties we assume that having detected them, then users wish to assess their significance. For evaluation a new set of requirements arises. We need to know (1) the context of use, and (2) a model and possibly a hypothesis to determine significance. Cross-validation is one common method used to assess statistical prediction. In cross-validation, observations are iteratively deleted and the remaining data are used to predict deleted observations. Repeating this over many deleted subsets allows an assessment of the variability of prediction error.

Fuzzy classifiers have recently become sources of error descriptions and error models (Burrough 1989; Leung et al 1992). In a raster representation fuzzy classifiers provide a means of describing uncertainty by associating each pixel with a vector of class memberships. Goodchild et al (1992) describe an error model based on the vector of probabilities for a pixel's class membership. These methods can create quite large processing and/or large storage overheads.

In summary it should be evident that substantial costs and processing can be required to generate information that can be graphically displayed to identify error or uncertainty in data. The form and content of graphic displays will be highly dependent on the success of the error analysis. In most cases it cannot be assumed that such analyses will have been carried out on the data. The implications are that GIS or other visualisation software packages must either include error analysis tools or data producers must subject their data to these analyses and store the results alongside the data.

3.4 User satisfaction issues

User satisfaction issues relate to the packaging around the graphic and error analysis tools. The interface to these tools should, as always, be intuitive and easy to use. Users should be able to get the error information they require without losing sight of their original application goals. The ideal graphic displays are those which are simple, relevant, and unambiguous. Uncertainty in the data should not be mapped to an uncertainty in the graphics in a way which requires the user to search hard or spend a long time interpreting the results. For most users the evaluation of uncertainty and error is a step on the path to some further goal rather than an end in itself. Thus error analysis and graphic display should not be a long involved process for the user.

One pertinent issue is how much error information users are really interested in. For example, will users be satisfied with depiction of the existence (location of errors or uncertainty), or will they desire more extensive information such as the rank, magnitude, or significance of errors? Computing the magnitude and significance of errors or uncertainty requires more processing of the data while the display of quantitative values requires some very thoughtful consideration of the visual display.

4 FRAMEWORK FOR GRAPHICAL METHODS

Several frameworks for mapping data to graphic displays have been developed in the past by researchers from different fields (Bertin 1983; Mackinlay 1986; Robertson 1991). While error and uncertainty are inherent in spatial data, indicators of these are not sitting at the surface waiting to be displayed. They must be extracted and revealed through error analysis. Because of this dependency, error analysis and graphical methods are necessarily bundled together. The scope of the proposed framework thus deviates from previous frameworks in that it creates a two-phase mapping. The first mapping is between data, an application context, and a suite of appropriate error analysis methods. The second mapping is between the outcome of the error analysis and graphical display methods.

The framework organises information around three basic components: (1) the data; (2) the context of the analysis; and (3) error analysis/graphical

methods. Ideally this framework should support the categorisation of the above examples and also indicate what additional error analysis/graphical methods may be required for a given combination of data characteristics and context. In future it might also provide a foundation for an automated visualiser that, given certain inputs, could respond with an appropriate suite of error analysis and error visualisation methods.

4.1 Data characteristics

Characteristics of the data to be considered in the framework include: (1) status – whether the data are raw or processed and, if processed, what processes and parameters were applied; and (2) observed spatial, thematic, or temporal dimensions of the data.

4.1.1 Data status

Some indication of data status is necessary to assign an appropriate error analysis method. Data status is essentially a lineage report (FGDC 1995). It should describe how the data were collected, how they were compiled and what processes were applied along with any associated parameters, and whether field checks were carried out. For example it would be important to know that ground control information had been gathered if one wished to compute RMSE for positional accuracy.

4.1.2 Data dimensions

The error analysis and graphic display will also depend on what dimensions of the data were observed. The range of possible dimensions includes the three spatial dimensions x, y, and z, several attribute dimensions $A_1 \ldots A_n$ and time, t. An observation could be a 2- or 3-dimensional spatial observation in which only geometry was observed (a survey measurement), a single or multivalued spatial observation or estimate in which geometry and attributes were observed or estimated (e.g. soil colour and texture at location P), or a single or multivalued space-time observation (e.g. observations on surface temperature and precipitation at the same station at the same time intervals). The presence of these dimensions along with their level of measurement provides essential information for error analysis and display.

4.2 Context characteristics

The context description should indicate the environment in which the error analysis might be carried out, specifying at least (1) the task – error detection or evaluation; (2) the desired dimensions of the error analysis – spatial, thematic, temporal, or combination; and (3) the user types.

The first distinction is between detection and evaluation tasks. The error analysis could involve exploratory methods to detect blunders or unusual values. At the simplest level, detection may be accomplished by plotting the data and relying on the human eye to do the detection. Evaluation methods may also be exploratory but extend toward confirmatory data analysis. Evaluation tasks would include tests for the significance of the errors.

4.2.1 Dimensions of the analysis

The second specification of context is the type of error analysis a user wishes to conduct. The possible dimensions for error analysis are spatial, thematic, temporal, or a combination of two or more of these. For example, the only information that may interest a user may be the error or uncertainty in the location of an observation. With this information the error analysis can be restricted to those methods appropriate for detecting and evaluating positional errors or uncertainty. This information needs to be combined with the available data dimensions. The observed dimensions might restrict a desired positional error analysis to two dimensions rather than three.

As another example, a user may be interested in the accuracy of an estimated attribute in space, in which case another set of error analysis/graphical methods will apply (e.g. cross-validation). Additionally, if users are interested in the reliability of an attribute in space and time, a set of methods to support this analysis can be identified. In this latter situation, animation may be a logical method to detect errors or to evaluate the uncertainty of attributes in space and time.

4.2.2 User types

The user type should also influence the selection of error analysis and graphic methods, but user types are generally too broad to categorise. As an example of why user types can be important, take two prototypical users and their associated tasks for which a set of error analysis/graphic tools might require special consideration. These are the data producer/distributor and the data browser in a digital spatial library.

Data producers which are also data distributors need robust error detection and correction tools that

can operate quickly and effectively on large volumes of data. Data producers will be dealing primarily with raw data for which the major objective will generally be blunder detection and correction. The error detection and evaluation will be context-independent so a review of all dimensions of the data (space, theme, time) may apply. A goal for the data producers might be to save the results of the error, analysis and graphical displays as metadata for transfer with the data to end users.

Digital libraries may be the new setting for error analysis. Data browsers will likely search for data, and evaluate them to determine if they will be sufficiently reliable for particular applications. Detection and evaluation tasks may both apply, and the error analysis and graphical methods will need to be fast since users may be paying for connection time. Additional constraints and challenges arise in the context of a client-server environment. Both error analysis and graphics will need to be simple and efficient in order to work over the possible range of client configurations. Other technical issues arise such as: will the error analysis and graphic tools reside with the client or with the server?

4.3 Error analysis/graphic display options

To organise the error analysis techniques within the framework we assign them several attributes which relate them to the data and their context. These attributes include: (1) the data status level for which the techniques are applicable; (2) the data dimensions for which they are applicable; (3) the analysis tasks (detection, evaluation) for which they are applicable; (4) their computational complexity; (5) their cost; and (6) the applicable user groups.

The examples that follow distinguish only two data status levels: raw and processed. As more specific metadata becomes available and as error analysis methods evolve it should be possible to make this relationship more specific (e.g. error models linked to specific spatial operations). Some error analysis methods are general and can apply to the analysis of any dimension (e.g. plotting). Others may apply to a single dimension (e.g. adjustment computation for the spatial dimension). A collection of error analysis methods could address purely the accuracy or uncertainty of the thematic dimension of observations. More pertinent are methods that indicate errors in thematic variables as they are distributed in space. Users may be interested in errors in the observed value at these locations or in estimating values and determining the reliability of the estimates. Cross-validation, Kriging prediction error, and simulation would apply in this latter case (see Anselin, Chapter 17).

Other error analysis methods may also apply specifically to the temporal dimension. Temporal errors are difficult to detect because they can be easily confounded with other errors. They can be detected by an observation being out of sequence or in a future time, through inconsistencies among relationships to other known events, or through aberrations in a measured attribute value. For example a recorded January temperature of 80°F in Maine could be indicative of an error in the recorded time. Similar to thematic variables, space-time errors or uncertainties are of particular interest.

An example set of error analysis methods along with their attributes and the tasks they are suited for is shown in Table 1. The curly brackets used under 'applicable dimensions' indicate that the analysis method applies to combined dimensions rather than to dimensions individually. The underline indicates the dimension of primary interest. Computational complexity could be measured more precisely but it is classified here simply by category. Reading through the table, it can be seen for example that plotting as an error analysis technique applies to raw data, can be applied to the analysis of all dimensions, serves the detection task, and has low computational complexity.

Each error analysis method produces an output which can be characterised. Given that the goal is to display the output graphically, the characteristics of interest are: (1) the level of measurement of the result; and (2) the spatial object to which the result attaches (point, line, pixel, surface, etc.). The graphic problem here can be expressed as one of representing k variables in an n-dimensional field using a fixed set of spatial object representations (points, lines, pixels, surfaces). The range of possible variables which need to be displayed either separately or jointly includes: (1) the observed data values; (2) the errors in or reliability of the observed values; (3) estimated data values; and (4) the reliability of estimated values. Depending on the interests of a user any one of these four may be displayed independently or in some combination. If displays of data and reliability are combined it is necessary to know the characteristics of both. The dimension of the field in which these variables are displayed can be two or three (simulated). The choice will be influenced by the observed data dimensions and characteristics of the error analysis result.

Table 1 Error analysis methods and their corresponding dataset and context characteristics.

Error analysis method	*Data status*	*Applicable dimensions*	*Tasks*	*Computational complexity*
Plots	raw	x,y,z,a,t	detection	low
Consistency checks	raw	x,y,z,a,t	detection	low
Ground truth checks	processed	x,y,z,a	detection, evaluation	low
Adjustment computation	raw	x,y,z	detection, evaluation	low–moderate
Cross validation	processed	{x,y,z,a}	evaluation	moderate
Fuzzy classification	processed	{x,y,z,a}	evaluation	moderate
Simulation	processed	x,y,z,a,t	detection, evaluation	high

Table 2 provides the information to link the outcome of the error analysis to a graphic display. It creates the second mapping between characteristics of the error analysis results and graphic display options. The table identifies the level of measurement of the output and the spatial object representation to which the output may attach. These two attributes give sufficient information to display the error information independently of the data. Several previously developed frameworks (Bertin 1983; Robertson 1991) provide the structure for working from this level to assign appropriate visual variables. The more specific objective of Table 2 is to guide the choice of graphic display mode if the data and their reliability are to be displayed together. In a joint display some spatial representation of the error analysis output must be shown in combination with a spatial representation of the data. The key therefore lies in combining the two spatial representations. The graphic modes in the table refer specifically to the graphic techniques for combining data and the reliability representations discussed earlier in the chapter. These include the side-by-side, composite, and sequenced images as described by MacEachren (1994) and the small multiples discussed by Tufte (1983). A composite map is the first choice since it is visually most efficient. The user can focus on one image rather than flick back and forth between two or more as in side-by-side or sequenced images. However the efficiency of the composite image breaks down as the number of variables or the complexity of the spatial representation increases. For example when both the data and the error analysis result are surfaces, a composite image of the two becomes difficult to encode graphically. When this occurs two simple side-by-side images are preferable. Table 2 includes the same error analysis methods which appear in Table 1, except for the plot which is excluded since it already exists as a graphic form.

Table 2 Basis for associating error analysis output with graphic display modes.

Error analysis method	*Level of measurement*	*Applicable spatial object*	*Spatial object evaluated*	*Graphic model*
Consistency check	nominal	point, line, area	point, line, area	composite
Ground control check	real	point, pixel, set of pixels	point, pixel, set of pixels	composite
Adjustment computation	real	point	point	composite
Cross validation	real	point surface	point surface	composite side by side
Fuzzy classification	real	pixel	pixel	small multiple animation
Simulation	nominal real	surface	surface	small multiple animation

To help make sense of this table a few explanations of the table entries may be offered. The row entry for consistency checks indicates that the output of this analysis takes the form of a nominal value (consistent or inconsistent). It is known from Table 1 that this analysis method can apply to any dimension. Column three indicates the spatial representation to which the error analysis result applies. If checking the consistency of the spatial dimension the result applies to the point, line, or area under analysis. If checking the consistency of attributes or times, it is assumed that the outcome can be associated with a point, line, or area. The fourth column refers back to the dataset being evaluated. It indicates the spatial representation (point, line, area) associated with the dimension being evaluated. Column five indicates the graphic mode. In this case a logical choice of graphic model is a composite. An inconsistency detected as the result of the consistency check can be displayed using a visual variable appropriate for nominal valued data such as colour hue or shape.

In the case of adjustment computations the outcomes are real-valued deviations around a point. The spatial object being evaluated is a point and the result of the computation refers to a point. The logical graphical choice is a composite in which for example the data points are shown in combination with their error ellipses. In the case of the fuzzy classifier the spatial object being evaluated is a cell, therefore, the outcome of the analysis is a vector of real numbers (probabilities) which can be associated with a cell. In the case of a simulation the entire dataset, referred to here as a surface, is being evaluated. The results will share the same level of measurement as the input data and take the form of multiple new surfaces, is being evaluated. To illustrate the error or uncertainty it is necessary to display several surfaces, so small multiples or sequenced images are appropriate.

5 FUTURE RESEARCH IN GRAPHICAL METHODS

The constraints on advances in this area are not technological. Technological possibilities have outstripped the ability to understand and model uncertainty and error in spatial data. The most pressing needs still lie in advancing error models for spatial data, the development of error propagation techniques, and enforcement or encouragement of better documentation of datasets (see Heuvelink, Chapter 14). Maps of mean estimates and estimation variance (common in Kriging) present a limited view of error and uncertainty (Hunter, Chapter 45; Cressie 1991; Hunter and Goodchild 1995). In general, graphical presentation of such error descriptors is inappropriate since maps of error descriptions are not possible realisations of stochastic error models. Instead, presentation of a sample of realisations, by animation or simultaneous display, may create the only sound understanding of uncertainty and its implications. Software functions to provide these must be implemented in existing GIS, concurrent with integration of spatial statistical models. Their use and comprehension must be informed by continued empirical testing (not simply of visual variables, but of cognitive use patterns and establishment of user comprehension).

Advantages of feature-oriented approaches to data quality representation have been determined, but more needs to be learned about the computational and data-volume overhead that their inclusion in a database may generate. Layer-based GIS functions must continue to proliferate in part because error modelling remains based upon layer data models, and in part because the layer model is most efficient for raster-based terrain and imagery. It is evident that errors that accrue differentially with specific GIS operations (buffering, overlay, coordinate conversion, etc.) may also depend on the data theme, resolution, and timeframe. Understanding of these differences needs to be formalised. Continued research must reinforce the development of data models and error models that spatially refine information for input to visualisation techniques.

Several error detection and evaluation methods for spatial data are computationally complex. To create these in timeframes acceptable to users will be challenging. The alternative of computing such information in advance and storing it has additional problems. The storage overhead may be quite substantial if the goal is to store several realisations of an error model. This approach also assumes knowledge of what information the users will want. There is, of course, the possibility of being less responsive to users.

Development of error models and more appropriate error analysis will only improve with better data documentation. While efforts are underway to improve metadata documentation,

there is still have a long way to go. Most of the metadata for spatial data archives are being created after the fact. This is an arduous and error-prone process. Metadata collection needs to start prior to data collection and continue parallel with the data lifespan (Beard 1996).

As databases become distributed and shared by multiple users, the need for users to detect and anticipate error information becomes critical. In this context the search for information and assessment of the quality of spatial data is outside a traditional GIS. The implications of these developments are that error analysis and graphical display software must be able to function independently of GIS. Future research should be directed toward interoperable components which could be easily recombined. Users could then select tools that would apply to a specific analysis context rather than having to support a large package which tried to incorporate all error analysis and display functions.

References

Beard M K 1996 A structure for organising metadata collection. *Proceedings, Third International Conference/Workshop on Integrating GIS and Environmental Modeling, Sante Fe.* Santa Barbara, NCGIA. CD and *http://www.ncgia.ucsb.edu*

Becker R A, Cleveland W S, Wilkes A R 1987 Dynamic graphics for data analysis. *Statistical Science*: 355–95

Bertin J 1983 *Semiology of graphics: diagrams, networks, maps*. Madison, University of Wisconsin Press

Bicking B, Beard M K 1995 Toward implementing a formal approach to automate thematic accuracy checking for digital cartographic datasets. *Proceedings AutoCarto* 12: 355–62

Brewer C A 1994 Colour use guidelines for mapping and visualisation. In MacEachren A, Taylor D R F (eds) 1994 *Visualisation in modern cartography*. Oxford, Elsevier Science: 123–48

Burrough P A 1989 Fuzzy mathematical methods for soil survey and land evaluation. *Journal of Soils Science* 40: 477–92

Chambers J M, Cleveland W S, Kleiner B, Tukey P 1983 *Graphical methods for data analysis*. Boston, Duxbury Press

Cleveland W S 1993 *Visualising data*. Murray Hill, AT&T Bell Laboratories

Cleveland W S, McGill R 1984 Graphical perception: theory, experimentation and application to the development of graphical methods. *Journal of the American Statistical Association* 79: 531–53

Cox D R 1978 Some remarks on the role in statistics of graphical methods. *Applied Statistics* 27: 9

Cressie N A C 1993 *Statistics for spatial data,* revised edition. New York, John Wiley & Sons Inc.

Dibiase D, MacEachren A M, Krygier J, Reeves C 1992 Animation and the role of map design in scientific visualisation. *Cartography and Geographic Information Systems* 19: 201–14, 265–6

Englund E 1996 Spatial simulation: environmental applications. In Goodchild M F, Parks B O, Steyart L T (eds) *Environmental Modeling with GIS.* New York, Oxford University Press: 432–7

FGDC (Federal Geographic Data Committee) 1995 *Content standards for digital geospatial metadata* (*June 8*). Washington DC, Department of the Interior and *http://www.fgdc.gov*

Fisher P 1994b Visualisation of the reliability in classified remotely sensed images. *Photogrammetric Engineering and Remote Sensing* 60: 905–10

Fisher P 1994c Visualising the uncertainty of soil maps by animation. *Cartographica* 30: 20–7

Gershon N, Brown J R 1996 The role of computer graphics and visualisation in the global information infrastructure. *IEEE Computer Graphics and Applications* 16: 60

Goodchild M F, Buttenfield B, Wood J 1994 Introduction to visualising data validity. In Hearnshaw H, Unwin D (eds) *Visualisation in geographic information systems*. Chichester, John Wiley & Sons: 141–9

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

Haining R P 1990 *Spatial data analysis in the social and environmental sciences*. Cambridge (UK), Cambridge University Press

Hancock J R 1993 Multivariate regionalisation: an approach using interactive statistical visualisation. *Proceedings AutoCarto 11 Minneapolis*: 218–27

Hunter G J, Goodchild M F 1995 Dealing with error in spatial databases: a simple case study. *Photogrammetric Engineering and Remote Sensing* 61: 529–37

Imhof E 1964 Beiträge zur Geschichte der topographischen Kartographie. *International Year Book of Cartography* 4: 129–54

Leung Y, Goodchild M F, Lin C C 1992 Visualisation of fuzzy scenes and probability fields. *Proceedings, Fifth International Symposium on Spatial Data Handling, Charleston*: 480–90

MacEachren A M 1994a *Some truth with maps: a primer on symbolization and design*. Washington DC, Association of American Geographers

MacEachren A M, Howard D, Wyss M von, Askov D, Taormino T 1993 Visualising the health of Chesapeake Bay: an uncertain endeavor. *Proceedings GIS/LIS 93 Minneapolis*: 449–58

Mackinlay J 1986 Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics* 5: 110–41

Maling D H 1973 *Coordinate systems and map projections*. London, George Philip

McGranaghan M 1993 A cartographic view of data quality. *Cartographica* 30: 8–19

Mitasova H, Mitas L, Brown W, Gerdes D P, Kosinovsky I, Baker T 1995 Modeling spatially and temporally distributed phenomena: new methods and tools for GRASS GIS. *International Journal of Geographical Information Systems* 9: 433–46

Monmonier M 1989 Geographic brushing: enhancing exploratory analysis of the scatterplot matrix. *Geographical Analysis* 21: 81–4

Monmonier M 1991a *How to lie with maps*. Chicago, University of Chicago Press

Morrison J L 1992 Implementing the Spatial Data Transfer Standard – introduction. *Cartography and Geographic Information Systems* 19: 277

Openshaw S, Charlton M, Carver S 1991 Error propagation: a Monte Carlo simulation. In Masser I, Blakemore M (eds) *Handling geographical information*. Harlow, Longman/New York, John Wiley & Sons Inc.: 78–101

Paradis J, Beard M K 1994 Visualisation of data quality for the decision-maker: a data quality filter. *Journal of the Urban and Regional Information Systems Association* 6: 25–34

Robertson P K 1991 A methodology for choosing data representations. *IEEE Computer Graphics and Applications* 11: 56–67

Robinson A H, Sale R D, Morrison J, Muehrcke P 1985 *Elements of cartography,* 5th edition. New York, John Wiley & Sons Inc.

Salgé F, Smith N, Ahonen P 1992 Towards harmonized geographical data for Europe: MEGRIN and the needs for research. *Proceedings, Fifth International Symposium on Spatial Data Handling, Charleston*: 294–302

Tissot A 1881 *Mémoire sur la représentation des surfaces et les projections des cartes geographiques*. Paris, Gauthier Villars

Tufte E R 1983 *The visual display of quantitative information*. Cheshire (USA), Graphics Press

Tukey J W 1977 *Exploratory data analysis*. Reading (USA), Addison-Wesley

Wright J K 1942 Map-makers are human: comments on the subjective in maps. *Geographical Review* 32: 527–54