Title

NCGIA Research Initiative 7 Visualization of Spatial Data Quality: Scientific Report for the Specialist Meeting (91-26)

Authors

Beard, M. Kate
Buttenfield, Barbara P.
Clapham, Sarah B.

Publication Date
1991-10-01

NCGIA

National Center for
Geographic Information
and Analysis

NCGIA Research Initiative 7
Visualization of Spatial Data Quality

Scientific Report for the Specialist Meeting

8-12 June 1991
Castine, Maine

by

M. Kate Beard, Initiative Leader
NCGIA, Department of Surveying Engineering
University of Maine
Orono, ME 04469,

Barbara P. Buttenfield, Initiative Leader
NCGIA, Department of Geography
State University of New York at Buffalo
Buffalo, NY 14261,

and

Sarah B. Clapham, Graduate Research Assistant
NCGIA, Department of Surveying Engineering
University of Maine
Orono, ME 04469

NCGIA Technical Paper 91-26
October 1991

Table of Contents

Executive Summary

This report is a summary of the Specialist meeting for NCGIA Research Initiative 7 entitled "Visualization of Spatial Data Quality" . It presents initial discussions on the role and utility of visualization for understanding and analyzing information about the quality of GIS data. The impetus for the initiative is based on rapid changes in spatial information system technology and a desire to see the technology used more effectively. Technology currently allows us to process and display large volumes of information very quickly. Effective use of this information for analysis and decision making presupposes that the information is correct or reasonably reliable. Information on the quality of data is essential for effective use of GIS data: it affects the fitness of use of data for a particular application, the credibility of data representation and interpretation, and the evaluation of decision alternatives. The credibility of spatial decision support using GIS may indeed depend on the incorporation of quality information within the database and the display.

The Specialist Meeting was held in Castine, Maine, 8-12 June 1991 to discuss issues of Visualizing Data Quality with researchers and representatives from the public sector, private sector, and academia. Participants from North America, Europe, and the United Kingdom worked together for four days to prioritize a research agenda. The highest priorities for initiative research are to develop tools which allow spatial data handling systems to be sensitive to error propagation, to design and implement tools to encourage use of metadata in spatial analysis and spatial decision-support, to facilitate understanding of data quality variations in digital data, and to sensitize the GIS user community to accuracy issues.

Visualization should be explored as a method for capturing, interpreting, and communicating quality information to users of GIS. Clearly, the quality of information varies spatially, and visual tools for display of data quality will improve and facilitate use of GIS. At present, those tools are either unavailable (in existing GIS packages), not-well developed (error models) or only recently developed as a research area (visualization).

Discussion at the specialist meeting by public sector participants indicated that the quality of spatial data and databases is a major concern for producers of digital data. Private sector participants indicated a strong interest in enhancing system support for data quality management. Academics expressed a need to manage quality within spatial analytical tasks as well as a research area in its own right. Jointly the representatives of the various groups entered into the spirit of moving this research area forward. Initial projects and products proposed for the initiative include a working bibliography, a periodic newsletter, presentation of research by external and NCGIA researchers at national and international conferences, options for refereed research published as journal special issues, and generation of a public domain Compendium showing examples of data quality displays that have been or can be implemented in GIS packages.

Acknowledgements

The initiative leaders wish to thank all the Specialist Meeting participants for their energy and enthusiasm in helping to focus and generate research priorities for this initiative. We also extend a hearty thanks to the graduate student assistants for their hard work and patience in recording the discussion and otherwise helping to make the meeting run smoothly. For careful and cheerful preparations behind the scenes, we thank Kathleen Hornsby, Connie Holloman, and Blane Shaw. We also thank the Maine Maritime Academy and other proprietors of Castine, Maine for their hospitality and for supporting a constructive and enjoyable working environment.

The Specialist Meeting and this report are contributions to Research Initiative #7- "Visualization of Spatial Data Quality" of the National Center for Geographic Information and Analysis. We acknowledge support from a grant by the National Science Foundation (SES-88-10917).

1 Description and Scope of the Initiative

The primary goal of this initiative is to develop a research agenda to explore, adapt, and evaluate visualization techniques for the representation and communication of data quality. Several recent developments in GIS as well as projections on future GIS use are the rationale for this enterprise.

Technology currently allows us to process and display large volumes of information quickly. National spatial databases are becoming available, and we are beginning to accumulate regional and local databases as well. These databases are becoming accessible to large numbers of people and being called on to support a wide range of applications. In the course of these developments, the users of spatial data become more and more removed from the details of data collection and processing which have historically been the basis for an awareness and understanding of data quality. As databases become distributed and shared by multiple users, the need to accumulate, store, and communicate data quality information to this large and growing pool of users becomes an important aspect of geographic information system development.

Information on the quality of spatial data and databases is a major concern for both developers and users of GIS (Chrisman 1983). Producers are concerned about the utility and credibility of their products, and users need to be concerned about the reliability of interpretations and decisions which can be made from such products.

The quality of spatial information and spatial information products is multidimensional and complex. The quality of information varies spatially and temporally and the need for such information will vary by application. If we assume that data have been processed and checked sufficiently that gross errors have been removed, we still face the problem of presenting to users the appropriateness of data for their needs. The volume of information required to adequately describe spatial data quality is thus potentially quite large.

Communicating this potentially large and complex pool of information to users presents a challenge for GIS development. Visualization has recently been proposed as a technique for making complex information more comprehensible. It has been variously described as: the organization of abstract concepts into meaningful pictures; the transformation of numerical data into understandable images; and the manipulation of geometry, color, and motion. At present, visualization tools are either unavailable in existing GIS packages or not-well developed. The goal of this initiative is to explore these techniques as tools for communicating the many dimensions of geographic data quality to users in meaningful ways and as a consequence improve and facilitate use of GIS.

This initiative does not stand in isolation but continues and expands work begun by other initiatives. The initiative builds on work from Initiative 1: The Accuracy of Spatial Databases. It has connections to Initiative 3: Multiple Representations since the representation of data quality can be seen as another view of the data. Access to data quality information through visual tools influences the Use and Value of Spatial Information - Initiative 4. It supports work on Initiative 6: Spatial Decision Support Systems by considering the impact of data quality displays on spatial decision-making. It plays a role in Initiative 12: The Integration of GIS and Remote Sensing with respect to enhancing the visualization of data quality through images. Initiative 13: Development of User Interfaces for GIS has direct implications for developing effective tools for user access to data quality information.

Initiative 7 focuses on impediments and research priorities within and across four categories. These include: the components of data quality, representational issues, the development and maintenance of data models and databases that support data quality information, and evaluation of visualization solutions in the context of user needs and perceptual and cognitive skills.

This document reports on the specialist meeting which was organized for initial discussions of the role and utility of visualization for understanding and analyzing information about the quality of GIS data and on research ideas which were generated during these discussions.

2 The Specialist Meeting

The Specialist meeting provided an opportunity to bring together representatives from academia, federal agencies, and industry to discuss their understanding and knowledge about the role of data quality and the potential for visually communicating this information to people. The overall goal was to consider a variety of perspectives on the general topic of Visualization of Spatial Data Quality which could then be formulated as a research agenda available to the general GIS community

2.1 Organization & Preparation

The Specialist Meeting was preceded by a research panel organized at the Baltimore AUTO-CARTO 10 conference in March 1991 with the purpose of garnering maximum input from the general GIS community. The panel was chaired by Barbara Buttenfield, and panelists included:

Alan Saalfeld (US Census);
Robin Fegeas (US Geological Survey);
Ferko Csillag (Syracuse University);
Alan MacEachren (Penn State University); and
Nick Chrisman (University of Washington).

Panelists presented position statements on themes including, Components of Data Quality, Database and Data Modeling, Graphical Representation, and

Assessment of Data Quality. The panel was well-attended, and audience participation was very good. Ideas stemming from panel discussions were brought up during the specialist meeting as partial guidance for prioritizing a research agenda.

2.2 Specialist Meeting Participants

The participants were invited from a diversity of disciplines representing academia, federal agencies, and industry. A number of participants had been involved in previous initiative efforts. However, an effort was made to involve individuals with no previous association with NCGIA and outside the domain of geographic information systems. The participant list is included below. For complete address and affiliation information see Appendix B.

2.2.1 Academics

GIS and Cartography:
Nick Chrisman, U. Washington
Alan MacEachren, Penn State University
Matt McGranaghan, U. Hawaii
Mark Monmonier, Syracuse University

Computer Science and Mathematics:
Virginia Hetrick, UCLA
Gerard Heuvelink, The Netherlands

Cognitive and Perceptual Psychology:
Geoff Loftus, U. Washington
Peter Stringer, Belfast, Northern Ireland

Soils and Environmental Resources:
Ferenc Csillag, Syracuse University
Peter Fisher, Kent State University
Jim Palmer, SUNY-ESF, Syracuse

Spatial Statistics:
Carl Amrhein, U. Toronto
Noel Cressie, Iowa State University
Dan Griffith, Syracuse University
David M. Mount, U. Maryland

2.2.2 Private Sector

Geoff Dutton, Spatial Effects, Watertown, Massachusetts
Bob Maki, ESRI, Redlands California
Mark Litteken, IBM Corp.

2.2.3 Public Sector

Jay Feuquay, NASA-EROS Data Center, South Dakota
John Kick, SCS, Syracuse NY
Alan Saalfeld, Statistics Division, US Census
Denis White, EPA, Corvallis, Oregon
Joel Yan, Statistics Canada, Ottawa

2.2.4 NCGIA

Michael Batty, NCGIA - Buffalo
Kate Beard, NCGIA-Maine
Barbara Buttenfield, NCGIA - Buffalo
Helen Couclelis, NCGIA-Santa Barbara
Andrew Frank, NCGIA-Maine
Michael Goodchild, NCGIA-Santa Barbara

2.3 Meeting Format

The specialist meeting was held in Castine, Maine at the Maine Maritime Academy and covered three and a half days from June 8-12 1991. The general format of the meeting along with other special activities are described in the following sections.

2.3.1 General Format

The general format of the meeting was a series of group discussions, alternating between small and large groups and focussing on a single theme each time. Small groups were selected by the initiative leaders with the goal of mixing participants from different disciplines and affiliations. Small groups were assigned a theme and a series of questions as a basis for discussion. The questions, developed by the initiative leaders, were intended to provoke discussion and not meant to restrict the potential scope of the session. In some cases the questions were refined or reworded by the group participants. Each group was required to select a spokesperson for the group. The dynamics of each group varied from unstructured brainstorming to logical progression through the questions. After small group deliberations, participants reconvened as a large group, and spokespersons presented summaries of their small group discussions. Graduate student rapporteurs were assigned to each small group to record discussions and assist the spokespersons in compiling summaries. Two rapporteurs also recorded discussions during the large group sessions.

2.3.2 Presentations and Software Demonstrations

In addition to the group discussions a two hour period was reserved for five presentations of visualization or visualization related projects by both participants and graduate students. Summaries of these presentations follow:

1
Evaluating the Effectiveness of Graphic Designs for Data Quality
Sarah B. Clapham

A prototype tool was developed to investigate the effectiveness of certain graphic designs at communicating information on positional uncertainty. The prototype was designed to evaluate subject response to variations in graphic design portraying positional precision for pre-described point features. Subjects were shown five variations in graphic symbols. Subject response was recorded as reaction time and symbol selection. While rigorous quantitative analysis has not been attempted with these results, observation were made in two general areas: (1) the performance of the prototype as an evaluation tool in user perception of 'quality' information, and (2) the complex interpretation of 'quality' information in association (or conflict) with feature attribute value.

2
The Data Quality Notebook
Kate Beard

The Data Quality Notebook is a Hypercard stack developed as a graduate student class project. The stack provides definitions of quality, spatial data quality, and spatial data quality components. The main purpose of the stack is to track specific data collection methods to understand how collection and processing activities effect spatial data quality. Two case studies in the notebook, remote sensing and soil mapping, document the steps of data collection and compilation into a product which could be incorporated in a GIS. A matrix is used to show how each compilation step effects positional accuracy, attribute accuracy, completeness, and logical consistency. As there is currently little or no quality information associated with GIS data, the case studies were intended to provide generic information on quality aspects of these GIS data sources. The stack uses the metaphor of a small spiral notebook in which the users can flip through the pages to find the information they need.

3
Experiments in Representing the Uncertainty in the Location of a Point
Diane Schweizer

This project displayed several techniques for representing the positional uncertainty of the same data. Symbols were developed not necessarily to automatically trigger "positional uncertainty" in the users mind but to lend themselves to the idea that something is unclear or uncertain about the representation. Techniques developed included the use of a random distribution of points simulating a cloud, variations of color saturation, broken lines and a clearly outlined symbol displayed within a cloud of points. The symbologies using a random distribution of points or variations of color saturation are easily adapted to a Gaussian distribution of uncertainty. The remaining symbols are useful when the distribution of uncertainty is unknown.

4
Probability Bands for Line and Areal Features
Geoffrey Dutton

Hypercard provided a testbed for generation and visualization of probability contours for line and areal features. The stack illustrates that at any vertex along a segment one may compute the width of a band representing some number of normalized units of error. The area of the band can be computed by integration or approximation. The stack used animation to show that error bands grow dramatically at low levels of probability, but as the probability approaches 100% little growth in area occurs.

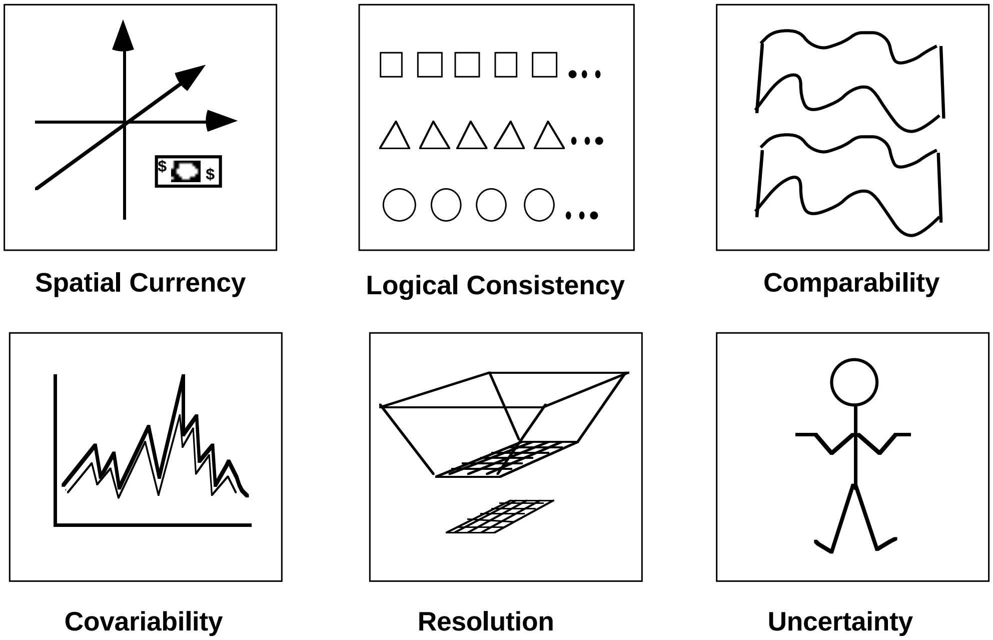
5
An Animated Display of Soil Mapping Unit Inclusions
Peter Fisher

Some spatial data have known levels of precision and known and documented error or noise terms. Many soil maps, for example, are supported with an extensive report which includes for each soil mapping unit, information on other soils that are included within the mapping unit. Peter presented a PC demonstration program developed by himself which uses the soil inclusion information to create an animated display of map-unit noise. The program holds a raster soil map in memory. Cells are continuously selected by random number generation and through a random process they are re-displayed as either the original map unit or as noise. The color displayed in a cell is frequently changed, and particularly the included map units are altered. The intent of the display is to convey the existence of error but the fact that the location of the error is unknown. By linking this display to interpretative tables of, for example, soil suitability of wetland habitat, it is possible to display a map where interpretative units have inclusions and included areas with alternative interpretations to the primary mapping unit are evident. This animated interpretive display is believed to have considerable applicability in land use planning, but no attempt has yet been made to demonstrate that the display performs the desired function of conveying the error information to a user.

2.3.3 Other Activities

2.3.3.1 Pictionary

The first evening's discussion was terminated with a game similar to the well-known parlor game called Pictionary. The group was broken into four small groups, and easels were set up in the middle of the room in such a way that each easel was visible to only one small group. Concepts of data quality (uncertainty, reliability, covariance, spatial currency, lineage, and so on) were drawn from co-leaders' notes of the Theme 1 discussion on Components of Data Quality, and passed one at a time to individuals standing at the easels, who then drew icons and pictures to visualize the concept. Other members in the small groups tried to guess what the concept was, and the first group to determine the concept 'won' the round.   The game continued for 8 or 9 rounds. Selections of winning pictionary pictures are illustrated in the figure below.



Spatial Currency

Logical Consistency

Comparability

Covariability

Resolution

Uncertainty

Figure 1. Winning Pictionary examples originally compiled by Virginia Hetrick.

3   Initiative Themes

This section describes the four themes used as the framework for this initiative. These themes included data quality components, representational issues, data modeling and database issues, and evaluation of solutions. Under each theme we include the questions posed to each of the small groups, the participants comprising each group, and a summary of each group's discussion.

3.1   Data Quality Components

When a person purchases a car they may construct a mental list of quality attributes they desire in a car such as reliability, low maintenance, fuel efficiency, or other factors. Spatial data can be subjected to a similar appraisal, but we need to know the appropriate characteristics by which spatial data should be evaluated.

One of the most commonly cited components of data quality is error. Commonly recognized errors include those associated with data collection (source error) and the processing of data (process error). Process errors have proven difficult to analyze in many cases, for example in studies of digitizing error, or in modeling error associated with soil mapping (Fisher 1991). In statistics, the concept of Least Squares Error has been applied to determine reliability (or what is called 'confidence') in hypothesis testing. A third error component (use error) defined by Beard (1989) is associated with the appropriate application of data or data products.

The Proposed Standard for Digital Cartographic Data Quality (Moellering, 1988) introduced a broader framework for evaluating data quality. The standard included measures of accuracy (positional and attribute accuracy), consistency, completeness, and lineage as important data quality components.

Discussion groups were encouraged to not be bound by these existing notions of data quality, and groups were not expected to arrive at a standard definition or consensus on data quality and its components. This theme was intended as a starting point to investigate what quality components are important to users, how different quality components are related, and methods for quantifying or qualifying quality components in ways that would make them amenable to visualization. Small groups were given the following questions as a basis for discussion.

1. *What are the relevant components of spatial data quality?*

2. *What conceptual frameworks should be used to integrate data quality components with spatial data?*

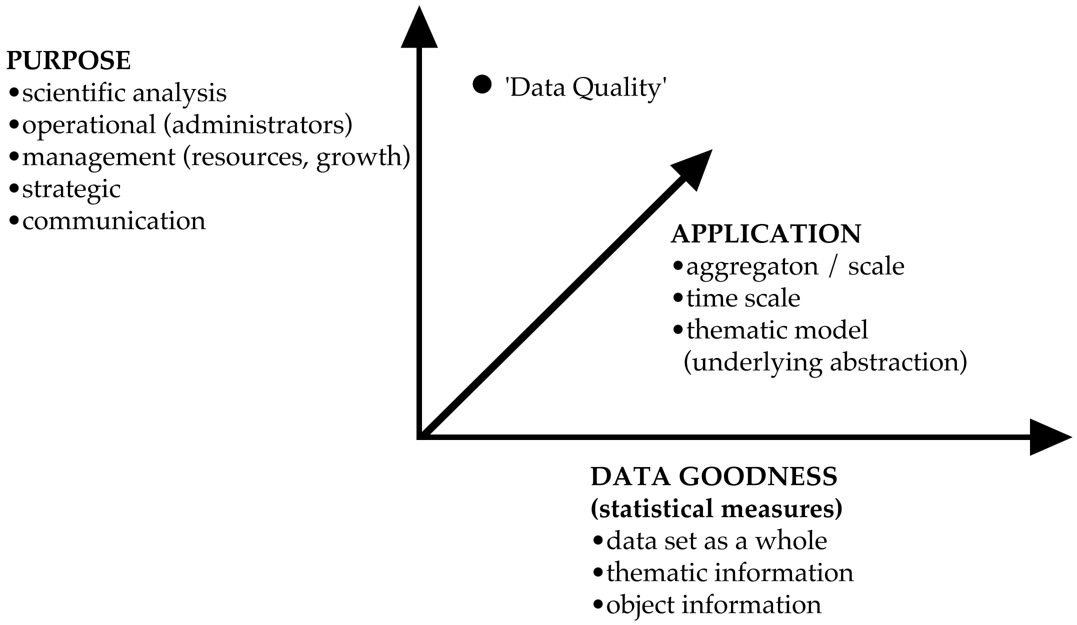
3. *Why are they important for improving our understanding and use of spatial data?*

Group compositions and discussions are summarized below.

Group A (Couclelis, group leader)
Participants: Dutton, Feuquay, Griffith, Hetrick, Saalfeld
Notetakers: Moreno, Soltyka

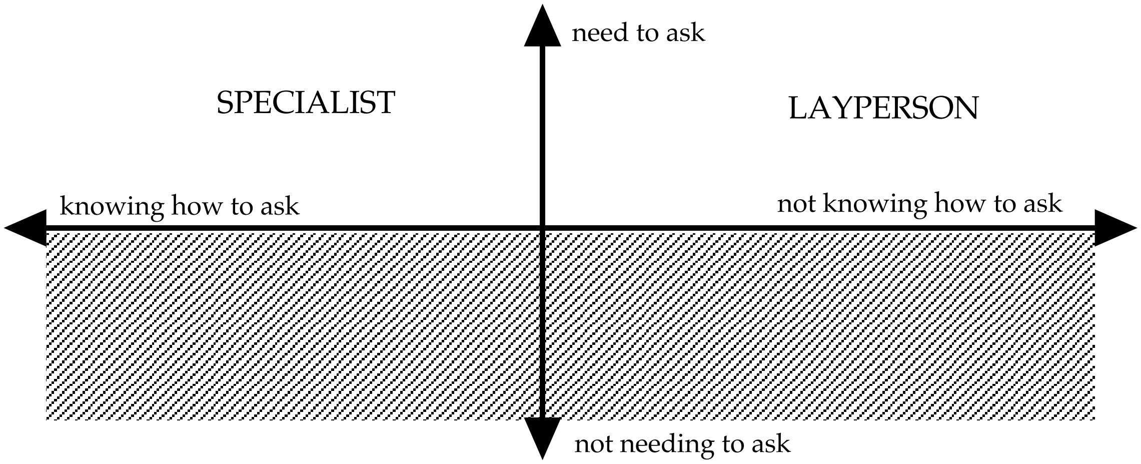
This group felt that "quality" was not the appropriate word to describe the "goodness" of data, and offered alternatives including: data qualities, data properties, data characteristics, and data attributes.

The group stressed that data quality means different things to different people and should be provided for different levels of users, making a distinction between the lay person and the professional, since their needs and level of understanding of data quality are quite different. They proposed a description of data quality as a point in a three dimensional space with axes of purpose, application, and goodness. This was not intended as a continuous space.



In large group discussions it was pointed out that this three dimensional representation of data quality was good in that it suggests data quality depends on the application, but poor in that it suggests goodness and purpose are independent of each other.

The group discussed data quality problems as being difficult to communicate to lay users, since some measures of data quality assume a technical understanding which the lay public does not necessarily possess. The problem may arise that some users do not know the questions they should ask about data quality. The group indicated users need to be educated on appropriate questions to ask and provided with the tools to ask these questions. The group presented another graphic to indicate the dichotomy between the professional and novice user with respect to communication of data quality.



The notion of fitness of data for a particular purpose was deemed important, but assessment of this is difficult. What is reasonable or unreasonable cannot be determined without a model which expresses the behavior of the data. The error in raw data and error which occurs during the manipulation of data cannot always be controlled. For most data there is no way to obtain error measures, and a particular problem with spatial data is that we typically do not have replications. One assessment of quality can be made through knowledge of expertise (of the data collector for example), but this is not easily measured or documented.

Aggregate measures of quality may not be very meaningful, and measures which say nothing about the distribution or geographic pattern of the error are not helpful. The group emphasized the need to describe the distribution of error and its spatial structure. However, simply assigning a number to error and mapping it may not be appropriate. Maps of residuals can leave you asking whether the data value is an interesting phenomenon or an erroneous data point.

The group felt there was a need to avoid overloading the system with quality information. They also indicated that users should be advised about data quality, but what they do after that should not be regulated.

Group B (Yan, group leader)
Participants: Batty, Beard, Heuvelink, MacEachren, Palmer
Notetakers: Schweizer, Rowell

This group focused on the data quality components as specified in the Spatial Data Transfer Standard (SDTS) with a discussion of advantages and disadvantages of this as a framework for data quality. Questions included: Is the framework set up with specific data types in mind? Is a global framework a good starting point and should there be a model that applies to all or several applications?

Lineage was discussed at some length as a key component.The role of lineage was described as tracking and providing information about quality but not serving as a quality measure itself. Objectives, goals, and limitations of the data belong under lineage and the lineage report should state or reference all assumptions, definitions, and procedures applied in data collection.

The location, theme, and time axes were discussed as another useful framework. It was felt that the time component was not comprehensively covered by the proposed standard. Update information was considered essential. Information on lifespan of data may also be important, in other words a projection on the duration of the utility of the information. Users should have the information to determine whether data are relevant to specific points in time, and to construct a longitudinal record for comparability across time. The group also discussed whether enough information should be provided to reconstruct history.

The danger in GIS is that a person is often at the mercy of someone else's data collection. The data may have been good for the purpose for which they were collected or similar applications but not for others. Users need to be aware of intrinsic limits of the data. Whatever classifications or categories are used will have limitations. For example there will never be an optimal classification of landuse or employment from everyone's point of view. These may be reasons why data have fixed limits on utility.

Quality reports should include as much detail as possible about the data, but the question is how much of this information do users really need and do they prefer aggregate measures over disaggregate ones. Different quality indices are appropriate for different purposes, and in some cases global indices are more appropriate than micro indices. The obvious problem with aggregate indices is understanding how measures of quality aggregate. For example is it possible or useful to aggregate indices of vertical and horizontal accuracy?

Users are often interested to know what purposes data can be used for. The group discussed whether quality information should be offered in the form of a list of recommended uses and limitations. It was pointed out that this approach has problems since it is difficult to generate an exhaustive list of uses.

Group C (Cressie, group leader)
Participants: Fisher, Frank, Kick, Monmonier, Stringer, White
Notetakers: Weber, Clapham

This group felt that “quality” was a rather broad word, and that uncertainty was a better word.

The group liked the idea of putting quality into a number, but wanted more than mean square error predication. They felt it was important to quantify the error structure of data models.

The analogy of visualization as a conversation metaphor presented by Helen Couclelis was well received. White suggested that people's expression of quality are inseparable from social and aesthetic values used to determine quality. We should look for ways of capturing these values as it may be too constraining to limit quality to accuracy, completeness, etc.

Frank suggested that assertive, directive views of reality are found in maps though the maps may claim to be declarative. For example, maps of the boundaries of South American countries whose borders are disputed and unclear, determine the activities of land ownership and military actions. They also express the political and economic ideals and needs of the map creators. The issues are complex and using models and formalism may be a constructive way to proceed.

Data quality must be model based to measure uncertainty. There will always be uncertainty and error at some level of resolution, and users need to be aware of such limitations. For example soils data are collected as an attribute at a point and transformed to polygon maps. These are not intended to reflect the actual soil type at any one point, but for purposes of management we consider points within the same polygon to be the same type. Frank suggested this reflects a difference between categorical and mathematical set theories. On cognitive levels the prototypical hierarchy of people's thinking does not fit the mathematical models. In soil mapping we use prototypes but people view the data as pure sets. In this context scale is an important issue that must be addressed. In moving across scales we often must make inferences about block data based on uncertainty of point data.

Stringer asked the group to make reference to the idea that data presuppose a theory and make assumptions about the real world. Levels of validity and reliability may be implicit in the theory. We thus need a theory to quantify data quality. Data models must be developed for different types of data and data inferences, but the models can only be as good as the theories which support them. Current theory is not well honed to quantification of data quality, especially categorical discrete analysis. Consider for example the unreliability of categorical statistics or discrete statistics.

The group suggested relevance is very important in the quality equation, and that we must look at relevance in terms of what people want to do with the data. The depiction of the data quality may interfere with the user's intentions either initially or entirely. It might be useful to distinguish between quality characteristics which are part of the data themselves and those which are a result of the user's use of the data. A conceptual framework should thus establish that quality comes before, with the data, and after you work with the data.

The group also pointed out that it is important to recognize that the need for quality information changes. It changes with time, particularly time of exposure to data, with level of expertise, and with the social, political, and institutional environment. A conceptual framework should thus incorporate social and economic values and people's concepts of reality.

Data quality details are not published as much as possible or necessary though attempts have and are being made. Computers have definitely contributed to the quality management problem because of their ability to produce more maps and more specialized maps, and because of the impact of data which are propagated for uses beyond their original intent. Data may be high quality within one context but poor when employed for another use simply because they pre-exist.

We need to inform users about data quality, but Kick noted that quantifying uncertainty from an institutional view point is not always desirable. Superiors and clientele may not want to be made aware of inadequacies in the data. We should therefore consider the consequences of not recognizing the lack of quality or data certainty. Ultimately caveat emptor applies., but users should not infer or overlook the intentions of the data collector or compiler. Education is one answer to this problem.

Group D (Chrisman, group leader)
Participants: Amrhein, Csillag, Goodchild, Litteken, Maki, McGranaghan.
Notetaker: Volta

The components of quality can consist of the five pieces - position and, attribute accuracy, completeness, consistency and lineage. Chrisman stated that these components were the consequence of a larger framework and derived from the objective of creating an implementable scheme for data quality. The five components of data quality should be linked to model, separate from a map based discussion. Chrisman suggested we not use maps as the basis for a framework since some maps look the same but are different and other maps look different but are the same. A model for data quality however, must consider the relationships among these components.

Sinton (1978) models theme, location and time, and we cannot treat these as independent components. Sinton observes that in data collection one component is typically fixed, one is controlled, and the other is measured. An important point to keep in mind is that for example in choropleth maps, error in the controlled component is not the same as error in the measured component.

There are two kinds of data quality:

1) inherent quality
2) task specific

In the second case, data quality has a direct relationship to use, In either case one number describing data quality is not sufficient. Quality is a multi-dimensional index which implies more than one view. Two views for assessing quality were discussed by the group.

A. Measurement View
assess attributes at x,y against some real value

B. Process View (process history):
assess against process/rules used to create a representation

Do we attempt to measure error directly or by describing the process (by examining the lineage). In terms of the measurement view the group discussed whether it was possible to test against the truth? This creates the problem that we must deal with ambiguity in the truth and the fact that reality is a social construct. One option is to test a database against a replicate (db') rather than truth.

db <- reality
difference
db' <- reality

One outcome of the discussion is that there are perhaps two realities:

db <- reality |
difference need to model this variability
db' <- reality |

The duality is evident in the conceptual framework, with lineage on one side and the testable components on the other. There is a need for both a lineage (history) and testable attributes, and we should not think just in terms of measurements. The question of testing uncertainty vs. lineage is very fundamental.

3.2 Representational Issues

The ease with which visualization tools may be integrated within GIS packages varies considerably depending on at least three issues, including the domain of the phenomena to be studied, and the purpose or intent of the user, and the format of the GIS software . This presents a substantial challenge to the system designer. Buttenfield and Ganter suggest in a proceedings paper for the Zurich Spatial Data Handling Conference (1991) that GIS requirements for visualization include conceptual, technological, and evaluatory solutions, which may be seen to vary over three broad domains: inference, illustration, and decision-making. Each presents a challenge to the integration of appropriate visualization tools, and each domain will enter in the discussions below.

Maps are a primary tool in GIS for analysis, interpretation, and decision-making. Current GIS software includes functions to create cartographic output automatically or interactively. None of the current turnkey systems include mechanisms to ensure the correct use of graphics functions. This may lead to poor use of graphics by untrained users. Poorly designed maps may convey false ideas about the facts represented by the data, and bias the decision-making process. Ways to improve the quality of GIS map products, and increase effectiveness of information transfer based on graphics should be explored. Existing guidelines provide only a rudimentary implementation for visualizing data quality. Research priorities that come immediately to mind under the theme of Representational Issues involve both system benchmarking and cognitive evaluations, as seen for example by the following questions:

1. *What are the range of visualization tools including internal tools (both cognitive and perceptual tools) and external tools (eg., graphical defaults in GIS packages)?*

2. *What design tools are available for depiction?*

3. *Which of these tools are most appropriate for the components and frameworks discussed under Theme 1?*

Group compositions and discussions are summarized below.

Group A (Palmer, group leader)
Participants: Batty, Chrisman, Couclelis, Cressie, Maki, Saalfeld
Notetaker: Soltyka

Visualization tools result from three things: design elements, system components, and constraints. Design elements may be drawn from conceptual frameworks such as presented by Jacques Bertin (1983) in his text Graphical Semiology, and used in many cartographic design classes in Europe and North America. System components may include software modules, graphical defaults, and both of these will impact upon the types of visualization tools at one's command. Constraints of hardware are readily documented, including speed for computation and access of data. Another type of constraint may be discovered in the perceptual and cognitive limitations of the human visual processing system, and will impact upon the use of displayed information. One group member posed the question of whether users should learn to interpret fuzzy data or whether map makers should learn to depict data 'fuzzily'.

Examples of visualization tools that should be explored in the context of data quality displays includes but are not limited to hypermedia, animation, and virtual reality. One idea tied to all of these is that the data should be proactive, in the anthropomorphic sense (data embedded with quality information should alert the user "hey, look behind me!") Non-map representations of spatial data (charts and graphs, for example) should be explored as an extension to developing techniques in Exploratory Data Analysis (Tukey 1977). Creation of spatial error displays implies that the models to describe it are already in place, which in many cases is not true.

The group discussed options for implementing visualization tools into data quality displays. The idea that a system could retain user-specified preferences from one session to the next has already been implemented for word-processing packages, and spreadsheets, and could be implemented in GIS packages as well. Retraining users to acclimate them to use data quality displays will likely be necessary. The concept of rendering a 'faithful representation' as opposed to generating flashy displays to enhance a package's market potential was also discussed. IT would appear that some retraining of vendors may also be called for as data quality displays become more commonly provided in existing GIS software. Finally, the question of creating displays for purposes of analysis versus illustration becomes important in terms of the types of design tools that are provided.

The group felt that users are the most important visualization tool. System implementation ought to proceed with the goal of empowering users, rather than constraining them. Interactivity is paramount for successful processing and understanding of data and data quality, and users must be allowed to operate on data to achieve effects of their choosing, regardless of the intent of the software designer. This mandates a good deal of flexibility in the representational tools provided in GIS packages.

Group B (Dutton, group leader)
Participants: Goodchild, Heuvelink, Litteken, MacEachren, Mount, Stringer
Notetakers: Rowell, Schweizer

The tools for visualization of data quality include easily identified external tools, for example mice and graphics tablets and other peripherals. Internal tools may be more difficult to itemize, as their relationship may be quite indirect. The example presented to the group was "an intuitive understanding of metaphors": the system metaphor framework may facilitate or inhibit users' access to and manipulation of data quality representations. Tools may show transformations from one form of data to another (from data to data quality, for example) as well as transformations from internal to external representations (graphical displays to mental imagery, for example). The example of an internal tool provided by Buttenfield during this discussion is to give users the task 'close your eyes and imagine an actor walking onto an empty stage' and then ask the users which side of the stage did the actor enter from -- an internal visualization operation is required to accomplish the task and answer the question. This example is also cited in MacEachren, Buttenfield, Campbell, DiBiase, and Monmonier (1991), although it is not referred to as a tool.

The group considered whether the issue of representation includes non-visual representations, for example using auditory signals (varying pitch or introducing sonic noise) to represent erroneous or 'noisy' data. Classifying tools overall for representing data quality, the group considered graphical metaphors, such as Bertin's visual variables, and this framework is consistent with discussions in Group A above. That is, the cartographic tools now in use may be sufficient, even though they are not currently implemented explicitly in many GIS systems. Two examples presented included varying saturation to show varying levels of certainty, and regionalizing system color palettes to assist the user in color selection for a particular representation (eg., using red to indicate stress in the data).

Nonconventional cartographic tools should also be explored, for example motion in the positional sense, as in animating motion, or motion in the radiometric sense, as in motion through a color palette (what is called color register animation). Another possibility is to route statistical pattern through the audio channels, for example, to play residuals from regression through speaker channels, listening to the degree of randomness. Blurring, fog, and de-focusing are also good visual indicators of uncertainty, although these will likely require user training prior to common implementation.

Group C (Monmonier, group leader)
Participants: Amrhein, Feuquay, Frank, Loftus, White
Notetakers: Clapham, Weber

The range of visualization tools that are available led to discussing Lakoff's (1987) theory of linguistic tools in his book **Women, Fire and Dangerous Things**, to relate aspects of cognition, categorization, and language. On a lower level, the range of tools can be discussed in terms of physiological variables, such as brightness, contrast, etc. The problem can be described in multidimensional terms, as the type of tools available will depend upon the data domain, the problem or research domain, and the level of user expertise. A toolbox for representing data quality was discussed, to include visualization tools such as brushing scatterplots, standardizing symbols, etc. Organization of the toolbox would include some mapping between the tool and its appropriate application, and this raised training issues associated with using the tools correctly. Training figured prominently in this group's discussion, and their conclusion was that much of the research agenda on visualizing data quality should involve training and evaluation.

Internal tools are not clearly understood, but seem to be mostly physiological. The schema of Lakoff and Johnson (1980) might provide a useful framework within which to pursue an inventory. The list of external tools is long, perhaps not finite for inventory purposes which would require a logical scheme for narrowing down and sequencing. Tools should be targeted to a particular audience. The reliability of tools must be evaluated. Finally, other sensory channels should be considered for implementation tools. The boundary between internal and external tools is likely fuzzy, as interactivity plays an important role and it is often difficult to distinguish.

Design tools available for depiction were discussed within the Bertin visual variable framework. Measures of data quality are intensity measures, said one group member, and thus demand a visual variable such value (progressive shades of gray from white to black), which is effective for intensity scaling. Saturation progressions may also be effective, so long as hue differentiation does not interfere with the progression of saturation. Interactive exploration of data was seen to be of primary importance.

As an afterthought to the small group discussion, Geoff Loftus pointed out to the general group that visualizing data quality might be viewed with a goal of optimizing representation and communication of information. If data quality displays improve understanding about the validity and reliability of data, then communication is improved, and closer to some optimum. However, it is difficult to articulate what exactly should be optimized. If researchers can formalize what is the optimal goal, and if the optimization strategy can be tied to perceptual theory, then the research agenda that is followed will make good progress in developing and evaluating representational tools for visualizing data quality. The process of articulating why it is that visualizing data quality is important should form a part of the research agenda.

Group D (McGranaghan, group leader)
Participants: Csillag, Fisher, Griffith, Hetrick, Kick, Yan
Notetakers: Volta, Moreno

The range of visualization tools were broken into conceptual tools, such as cartographic conventions and models of spatial error. Cartographic conventions can be uncovered from data models (discrete, categorical and continuous) graphical models (point, line and areal depictions) and measurement models (nominal, ordinal and interval scales). The manifestation of cartographic conventions commonly follows guidelines proposed by Bertin, whose visual variables (size, value, texture, color, orientation and shape) can be applied in either traditional displays or in new display media such as animation. Special effects such as blurring of focus can be made interactive, so the user controls the display transformation from data to data quality proactively.

The user must be informed by these conventions; and this calls for training in visualization skills. Users should be informed by associative meaning; and this calls for training in a systematic domain. Users should be informed by constraints on what the group referred to as bandwidth, which implies an understanding of one's perceptual and cognitive limitations.

Software tools range from exploratory tools to tools for presentation, and many products can be found along a continuum from illustration to analysis. Mathematica, Scientists' Workbench, APE, HDF, interactive image processing software, drawing packages, and graphics libraries such as xlib, DI-3000, PostScript, Plot10, etc., continue to find a marketable niche for scientific visualization. the range of hardware tools is dependent on technological developments, that continue to outstrip the capabilities of implemented software. The range of tools outlined by the group in this category of tools included virtual reality, 3D modeling, and animation.

Limitations to development of optimal representational tools include first and foremost the impediment of a lack of error models to describe spatial error and error propagation. A second limitation that the group cited was a lack of research on uses of new technology (virtual reality, for example) and its effects on user comprehension. Implementation of representational tools cannot entirely precede an understanding of how it is that users think about and comprehend the concept of error and certainty in a spatial context.

3.3 Data Models & Data Quality Management Issues

Management of data quality within a GIS database requires attention during manipulation and update, and will likely impact upon the future architecture of such databases. Information about the information within a database is referred to as metadata, and has recently become a research issue in its own right (see for example Lanter and Veregin, 1990). The representation of data quality components in a data structure should not only facilitate their visual display, but also facilitate update operations. Analysis of error propagation might also be facilitated by visual display, and the design of these graphic tools may not be closely aligned with the design of conventional GIS graphics. Small groups were asked to respond to this theme by addressing the following questions.

1. *At what level of aggregation (primitive, object, object class, layer, tile, database) should the data quality information be stored and/or linked to the data?*

2. *What forms of data models lend themselves to what types of visualization methods?*

3. *Where in the data management process is visualization most effective?*

Group A (Csillag, group leader)
Participants: Amrhein, Couclelis, Egenhoffer, Litteken, Loftus, Palmer,
Notetaker: Moreno

This group agreed that data quality information should be stored and linked at all levels. The level at which quality information is stored may depend on the type of measurement. The notion of data quality depends on the primitives, but there is a need to model how quality information aggregates. A simple example of an aggregation issue is the following: A surveyor measures points or lines which have error. The question is then how to infer the error in the area of a polygon from errors in the point or line measurements. The group suggested that once a model of aggregation is known, it could be performed by the system to allow the quality information to be more easily maintained. The group also pointed out that aggregation will be application specific.

Data quality information should be supplied at the lowest levels to ensure that data are appropriate for certain applications, but very often information is not available on the precision or accuracy of the lowest level of measurement. Most data are now collected by someone else. In such cases we should try to predict the types and levels of quality information other users will need.

The methods by which information is collected make a difference. If biases are known to exist for certain methods, then it should be possible to model these. In the census, people are known to lie about their income, but the possible error is difficult to document. In addition to bias in data collection methods, a model for tracking or recording quality information should consider knowledge of the purpose, what information is missing, and what processing was applied.

A structure similar to the quadtree data structure may be useful for some quality indices. If the quality of information is variable over an area, it should be possible to decompose the area into quadrants until each quadrant has a uniform level of quality. Another way to store quality information is as alternate views of the database. Different users could derive different views depending on different needs for quality information.

The group discussed at what levels the computer could/should supply warnings about inadequacies in the data. Using the quadtree example, if you know the smallest tile is a certain size and you know a user wants information on an area smaller than this tile, the computer could signal that this is not possible.

Very seldom does a person make a decision based primarily on the map.
For example a very aggregate soil map is typically not used to determine the location of a house. The map simply provides clues on places to check and where not to check. The group discussed how to deal with the aggregate response of suitable, not suitable from a computer.

In applying a linear optimization model for selection of a hazardous waste site, the model may not designate an area along a geologic fault line as unsuitable although the area is optimal for all other variables. If the position of the fault line is in error then an optimal location may be missed. This is a situation in which visual tools could help. Displays could be used to show not just the results, but the interaction of variables and how they change over an area. The visualization approach could allow users to work with the data, to identify outliers, to discover things they would otherwise not notice by examining only the optimization model results.

In summary the group suggested that we need quality information at all levels, but currently do not have it, and do not have the data structures and models to support it.

Group B (Maki, group leader)
Participants: Chrisman, Fisher, Frank, Griffith, Heuvelink, MacEachren
Notetakers: Soltyka, Rowell

One opinion in the group was that data quality information should be stored at the highest level possible. Arguments for this position were that this would reduce the storage load and that aggregate indices in some cases are more useful. The other issue is that aggregation may be required by fixed constraints, particularly those associated with socio-economic data, such as confidentiality and reliability. For example, the Census Bureau must aggregate some information due to confidentiality requirements.

The alternate view was that data quality information should be stored as specific (or as low as possible), so the location of errors can be known. For example when working with raster data, it was suggested you not throw information away by processing (classification). By retaining the raw imagery, for example, one could recreate probability vectors.

In most cases the level of aggregation of quality information will depend on what information is available and on particular needs. When you store metadata at higher aggregation levels, you always lose information. The group suggested we consider how much loss of information we are willing to sacrifice.

Adding data quality information or metadata directly to data can become unmanageable. Storing quality data for each feature or each mapping unit generates a lot of information. The example of storing probability vectors per pixel, while useful, could become troublesome and may need to be provided in a more aggregate form. In some cases we can store metadata at high levels without lose of information, assuming that lower levels (in the metadata hierarchy) are the same. In terms of storage efficiency it was suggested that the aggregation level be as high as we can accept a value as being representative and as deep as necessary.

It was argued by some in the group that we should not always be worried about storage, particularly in academia. The problem is more interesting (and important) to industry.

Processing efficiency was discussed in terms of query type. The nature of a query (where is the error in the map occurring), could determine at what level quality information should be stored. It was also discussed in terms of data transfer. It was felt that a lot of information could be lost during transfers if quality information were stored separately. Lineage information was suggested as being most prone to separation from the data. Lineage is an interesting component which can be discrete or narrative and which can attach to very different sized spatial objects. For example socio-economic data can come in big blocks, while hospital records are on an individual, per event basis. There was some discussion on whether lineage data should be stored as attributes or treated as a narrative.

Metadata can be stored with the data or it can be linked to the data in a number of ways. One way to link it would be to make data quality and data density interdependent, such that metadata density would vary with data density. Polygon overlay could also be used to associate quality information with the data itself.

The group also suggested we should be cautious about quality information becoming easier to ignore than to process. Do we in fact or should we force users to keep quality information? One proposed solution was to concentrate all metadata in one place, which would simplify retrieval. If metadata are isolated, however, the linkage with data may be lost during transfer.

The counter argument is that you cannot work with metadata without data. Data and quality information should be stored together, or a close association maintained between data and metadata. This is a data structure issue and a good argument for object orientation. Some members of the group pointed to the natural characteristics of object orientated models which are able to capture variability in data quality. An object orientated model captures not just the data but the *method,* which can be embedded with an object. These types of associations are more difficult to construct with the relational model.

In a number of cases, it may not be possible or desirable to explicitly tie data to metadata. (Quality diagram is only implicitly linked to the data.) One reason for working with metadata without working with data is that one should be able to examine the metadata before buying the data. In other words one should have quality information first to decide if the information is appropriate. Visual diagrams (separate from the data) may be useful for such data browsing.

There is a need for software designed to examine quality information. For example query processing could take quality data into account. You could build quality information into a query (ie. I want data that is no more than 5 years old, or show me areas on the map which I have not visited in the past 5 years). If the data set is older than 5 years, then say “look out, this is old data”.

Data quality is not merely a data collection issue. It is part of the whole processing scheme. For example, selecting the wrong map scheme or wrong statistics will effect quality. Data quality should be managed through processing to the generation of final products. The idea with respect to quality information is to try to come up with a model of quality that resembles how quality behaves during processing and product generation.

We can assume that there are some models that do not lend themselves to visualization. There are situations which cannot be easily visualized (eg. high dimensionality situations, multivariate diagnostics) due to an enormous amount of quality information. The complexity of data can determine how well overlays of metadata will work. The level of complexity may depend on the level of abstraction. Simple models will be the easiest to visualize since high dimensionality can cause problems.

We need to test whether humans can cope with complex displays.
There are limitations to any visual variables and the bottom line is we need to consider the limits of human visual perception. We may assume that some people are good at understanding quality measures and that some displays will be obvious, but others will require training. We should explore use of metaphors. Metaphors are typically based on something you can experience, such as fog for example. The ellipse is another example. Many people recognize an ellipse as representing uncertainty, but we must be cautious in assuming that visualizations are obvious to everyone. Error ellipses are common to certain professions (cartography and surveying engineering) which have been trained to use and understand them, but uncommon elsewhere. No visualization will be intuitively obvious to everyone.

Group C (Saalfeld, group leader)
Participants: Batty, Goodchild, Hetrick, Kick, Monmonier, White,
Notetakers: Clapham, Weber

This group felt that quality information must be connected to the level of collection, understanding that data are collected at different levels of detail. The important question then is how do we aggregate from this level. Aggregation needs to be based on a model of how the error behaves. The group generated the following diagram to illustrate the role of models.

Uncertainty in source -----> uncertainty in products

error model
(process)

vs.

description of sources ------> impact on product
model

A stochastic error model offers one possibility. A reliability diagram can also provide a spatial representation of error, but it is incomplete. It shows spatial error but not the size of local error. The reliability diagram tells where the data are reliable and where the user needs to be careful using the data. One issue is whether a reliability diagram can be applied to more than one product produced from the data.

Using the temperature of a room as an analogy, an error model could state that there is a Gaussian standard distribution with an error of 2 degrees. This would allow you to generate a map of temperatures within the room differing only in error. The source of the error differs. The thermometers could differ in reliability, in systematic error (which could be adjusted for), or in the way in which the thermometers were used.

In most GIS applications a product is derived from data, so to assess the reliability of the product, we must know the reliability of the data. For example a DEM has a documented accuracy of +- 7 m. for each elevation, but this information is not useful for determining the reliability of a product. We need to demand sufficient information on uncertainty of data in order to propagate error to products.

The term "error model" should be described by an adjective as there are both qualitative and quantitative types. We have cartographic error models,"process error model", categorical data models. An interesting point is that we consider error in geographic models yet there is no analysis of error in the statistical packages in use. Should we consider the sub-error due to the statistical process? In geography, we know we are only dealing with a representation of truth so we are more sensitive to 'error'. Sensitivity to error varies with the type of data and use. We tend to look for error less in census data because it is the best data we can get versus a soil map which can be always be improved with more money and resources.

Another source of error relates to manipulations of the data which can make an enormous difference in the models being used. We can only make progress if we have a systematic procedure for showing how those types of changes make a difference in the model. For example we know that changing the level of aggregation generates different distributions of errors.

For the room temperature example, the Gaussian assumption restricts the level of complexity in the model. If there is no knowledge of error, the best assumption is a statistical Gaussian distribution. The Gaussian model describes the probability but gives no information about clustering of data points or sub-areas within a given polygon. As new information is gained it can be build into the model.

The group suggested that the process of finding systematic errors is anecdotal and that we may not be able to incorporate those experiences even though we know they exist. The assumptions in data collection should be part and parcel of the data display. We often talk about organizing data by pre-existing theory. We assume a confidence in the theory, but at a later stage may discover basic errors.

We could ask whether these occurred because of data massage or collection (ie. the inherent measurement model). With anecdotal information there is a need to transfer the expertise that is not codified in mathematical formula.

With respect to question two, the group felt visualization techniques were appropriate for all stages. The group agreed that visualization techniques would be useful whenever data move from the control of one level to another, for example from soil scientist to analyst. Visualization techniques could also be effectively used at the data collection level by the individual most familiar with the data. Such a system could support visual clues, visual choice of models which could help soil scientists, for example, to capture transition bands. Differing rates of change could be represented by a menu of differing plausible values. The system might also be used to point out the fuzzy areas or areas where additional samples are needed. The analyst at the next level could use the display of data quality to understand the information known by the collector. It is was therefore thought to be worthwhile to distinguish between visualization for professional users and other types of users.

The group felt user education was essential. In large group discussions it was pointed out that when you buy a VCR you get a video tape of instructions on how to use it. It was suggested that a similar approach apply to spatial data, and that the assumptions and history of data processing be provided in examples or tutorials.

Group D (Stringer, group leader)
Participants: Cressie, Dutton, Feuquay, Kuhn, McGranaghan, Yan
Notetaker: Schweizer

Data quality information in most cases should be stored at the lowest level. However, we would want information at each level to respond to user needs. The group focussed primarily on aggregation of attribute error. Aggregate measures are complex and this is where statistical models can help. For example, using variance or mean square error as a prediction or error, each attribute has associated with it a variance. Ideally we would like to have a variance with each megalevel. The variance of a megalevel is the sum of each of the variances plus the cross term of variances.

The group considered the example of 100 counties in North Carolina with mean square prediction as a measure of how variable the data are. The objective was to map the sudden infant death rate in the northeast section and predict the measure of quality. If you start at a low level and work your way up, you need knowledge of covariances between regions, so not only do you need 100 variances, but you also need a 100 x 100 matrix assuming a 100 county region. The statistical approach provides guidelines, but may not be supported by current storage capabilities. The quality expression however, could be reduced to an analytical, numeric expression.

The problem with taking simply variances is that the variances do not tell you anything about the quality of the aggregation since each is computed individually. In terms of the example, the variances will tell you nothing about the quality of data in the northeast section of North Carolina. It is important to note that this applies to the attribute space not locational space since the aggregation problem is different for positional error than for attribute error.

If you aggregate the megacounties to a highly similar level of aggregation you can use the sample covariance of the megacounties. The other possibility is bootstrapping. Bootstrapping is highly non-parametric and difficult when you have spatial correlation.

In GIS we may be concerned with classes of features, themes, or mixtures of points, lines, and polygons. We should consider the notion of interaction between the mode of representation and the level of abstraction. Measurement or understanding of error becomes very complex in the context of whole layers and overlays of these layers, but if you can estimate the error it is better than having none at all.

Data quality information can be attached to any layer of aggregation. Each feature may have different amounts of error. The group considered the idea proposed by Chrisman in which we have a set of tiles and lineage information is tied to each tile. As you update information, new polygons are created with macro-level quality information which relates to the update. In terms of visualization, the overall information for each tile could lend itself to color coding or some other form of visualization.

For question number three, it was suggested that the most effective links are associated with transformations. Transformations can become very complex. The transformation stage can include real world, to graphic, to mental model, to output of another graphic model, or digital output, to another mental model, to output, to the mental model of the user. At each stage there is room for error. When transformations are involved, such as projections, or overlays, the data are changed. In these cases some sort of report should describe the transformation.

Considering the attribute problem and choropleth mapping, the group discussed the possibility of displaying measures of variation on the same map as the attribute information. It was suggested that this would require a second variable from which you could create a bivariate map or use a graduated symbol inside each polygon. The question, however, is do people need to see each variance or is an overall measure for the map sufficient?

3.4 Evaluating the Solutions

The small group format was working well enough to generate good discussion and mixing of perspectives by the end of discussions on Theme 3. The advantage of the small group format was to generated lots of interaction and debate; the obvious disadvantage is that the entire group heard only summaries of other groups, and many points were not shared with the whole group. The dynamic of the Specialist Meeting was very positive and interaction was high. Co-leaders Beard and Buttenfield decided to cover Theme 4 without breaking into small groups, and the ensuing discussion was both lively and informative. Participants did not appear to be intimidated by opening the forum of discussion to the general group. This session provided a natural transition from identifying impediments to preparing initial proposals for the research agenda, in the session on Wednesday morning. During the Theme 4 session, the group met for close to 3 hours to pursue lines of thought related to the following questions.

Question 1. *What are your expectations about what visualization should accomplish / provide?*
Detection
Notice
Identification
Quantification

The initial response to this question focussed upon definitions for detection, notice, identification and quantification. Additional connotations for the terms were provided, for example, the psychological connotation of “notice” is based on existence, as for example, in noticing an item in an image. In a legal context, though, “notice” is proscriptive, and often intended to imply a warning, as in a “notice to mariners”. In the end, other terms were added to the list, including discrimination, identification, qualification, magnitude estimation, elaboration, selection, and so forth. The basic expectation of the group seemed to be summarized in the statement of Matt McGranaghan that visualization should provide and easier way to identify pattern.

The advantages of visualization mentioned in the discussion included speed of pattern recognition, motion detection, and change detection. Additional advantages cited were to provide mechanisms to see the intangible, as in remote sensing of vegetative vigor, or to access the inaccessible, as with modeling molecular structure. Visualization should provide capabilities for all of the items on the list above, as well as to facilitate the additional operations cited above. For representation of data quality, other sensory input channels must not be ignored; Geoff Loftus reiterated the idea heard often in the small groups, that auditory channels should not be ignored for opportunities to communicate metadata information.

Question 2. Under what circumstances is visualization the appropriate tool for analysis and interpretation of data quality?

Visualization was seen to be particularly appropriate in situations when the quality description varies spatially. For example, varying dot sizes in a display of spatial autocorrelation would embed both density (the pattern description) with size (the quality description) in the same display. The goal here would be efficiency.

Another circumstance appropriate for visualization of data quality is to communicate very large or very small numbers, as in tolerance thresholds. Fisheye displays provide appropriate visualization for changing resolution, which was perceived by the general group as one of the components of data quality (see the discussion of Theme 1, above). Visualization was also deemed appropriate for exploratory data description, for studying a conceptual framework using a real-world metaphor, and for search for errors of consistency in a data structure.

Comments were made that visualization of metadata might transcend the barriers of language, reduce cultural bias, and recover latent structure in the pattern of data quality, just as with the pattern of data. The realization pursuant to this discussion was that data quality displays by and large will not be designed or displayed as finished products but for ephemeral or 'working document' purposes. As working documents, their design requirements are more ephemeral, and the need for flexibility and personal preference is much higher than for an illustrative archival quality product. What this implies is a demand for tools to generate display possibilities, as opposed to a demand for rules to standardize or constrain the process (or the display).

Situations when visualization is not appropriate were also solicited, precipitating group discussion on the need for training. Several group members stated that poorly designed visualization was always inappropriate, as for example in using nonstandardized data for choropleth mapping. Training users to be alert for graphical abuse can alleviate some of these problems, however system designers should also attend to good principles of design (within the demands cited above).

Other situations when visualization is inappropriate relate to data requirements. For example, when precision requirements are very high, tabular presentation will preserve the resolution of information better than a visual display. This may also be true for data sets having a small number of observations, or when the outcome of a particular analysis is simple and obvious. Policy situations may not require visualization, in every case.

It was concluded that visualization acts in a certain sense as its own constraint. A comparison was made to reading a book versus seeing a movie. In the former situation, the reader is free to imagine the appearance of a room, of a facial expression, and this freedom expands upon interpretations of the story line and implications. The movie takes away this freedom, and leaves the viewer with a nonambiguous view of these things. Similarly with visualization of data and data quality, the provided image may in fact tend to preclude other unrealized interpretations and images that might have proved beneficial.

Question 3. How can visualization of data quality impact the reliability and credibility of spatial decision-making?

The question assumes that data quality has an impact on decision-making, which is most likely base upon the premise that people tend to believe what they see, and attach credibility to visual displays. Thus displays of variations in data quality may be perceived as confusing, and perhaps ignored by decision-makers. One might argue for embedding data quality into data displays, although this type of graphic may become quite complex and require some training for appropriate use.

There was some aversion within the group to adopting a 'high-priest' attitude to graphical design, wherein a few individuals are assigned responsibility to decipher graphics as one might decipher tea-leaves. The removes decision-making from many levels of management and policy-making, and will not likely be accepted by the GIS community. Training in use of graphics and in the need for understanding graphical depiction becomes a general mandate, in this case. There was a good deal of discussion about whether decision-makers use information on data quality, and speculation on both sides of the issue of whether such information would be used were it provided.

The level at which a decision is made will affect its impact, and this has implications for credibility and reliability regardless of whether one implies a bureaucratic or conceptual level. The group felt that at higher levels, particularly in policy-making, graphical designs become more slick and polished, and carry less informative detail. By itself, the reduction of included information can affect reliability, stated one of the private sector participants (Litteken), and this may have a positive or negative impact, depending on several factors. What this means is that the impact of reducing information on data quality is not a simple factor to predict.

The group realized that our limited vocabulary about data quality and its variations extends beyond the domain of mathematical or statistical expression of models currently used to describe spatial error and its propagation. At the level of spatial decision support, there is also a limited vocabulary. The GIS community does not easily articulate its needs for data quality information, nor easily articulate the potential impact that knowledge about data quality could provide to analysts and policy-makers. The credibility of one's advice is not easy to assess in the simplest of situations, and assumptions that better data leads consistently to better decisions is likely naive. The group did agree however that visualization will change the decision-making process by virtue of its appearance, and training in the application of GIS for decision support should incorporate training in use and design of graphical displays, as well as in the use and interpretation of statistical information.

> *Question 4. How should displays of data quality be evaluated? How should their impact be evaluated?*

The group discussed at some length the prospect that labs could be identified in the GIS community where evaluation of data quality and data quality displays form a focus of research and production. There are currently academic programs in GIS with high level of expertise in empirical research design, most notably the University of South Carolina in Columbia, S.C. For the most part, however, academic programs in GIS do not emphasize training in perceptual and cognitive evaluation skills. In the private sector (eg. TIME magazine, USA Today, etc., evaluation of graphics is thought to follow a marketing evaluation approach, although none of the meeting participants had direct experience with the methods utilized by these publishers.

This raised the question of what determines a 'proper' graphical display in the first place. Many graphical displays in the private outlets are merely decorative. Should the display of metadata be considered in a similar context? Geoff Loftus (a cognitive psychologist) commented that the goal of the display should drive the evaluatory mechanism. If the point is that the viewer should remember the information, evaluation should be different, as opposed to the evaluation of a graphic whose purpose is to draw attention to a particular item. Peter Stringer once again raised the proposal for developing a laboratory where GIS applications could be pursued in parallel with studies of how those applications and associated displays are believed, comprehended, interpreted, etc.

Several case studies evaluating data quality were offered from the personal experience of participants, including a study of the quality of a Santa Monica data set (Virginia Hetrick), a displacement of time slices in a hologram (Geoff Dutton), a data quality study of urban settlements in the British Green Belt (Mike Batty), and a data quality assessment of 4 quadrangles in Pennsylvania (Denis White).

The discussion concluded with proposals for a concerted effort in the GIS community to pursue evaluations of data quality. In addition to the lab proposal discussed above, gaming and simulations were proposed, to help teach users about data quality and its impacts on spatial data. Joel Yan proposed setting up a compendium of examples of data quality, to encourage the community to generate examples and share these with other GIS users. The benefit would be to learn from each others successes and avoid making the same mistakes. NCGIA was cited as a logical coordinator of this effort, and the proposed compendium was readily accepted as a product to be delivered from the initiative.

4 Research Agenda

After the sessions discussing each of the four themes, the meeting participants were directed to consider researchable topics on the Visualization of Data Quality that might be addressed within an 18 - 24 month period, given the state of knowledge, the state of technology, and the specific expertise and interests of the current GIS community. Participants spent a full day in informal discussions keeping this mind, and reconvened Tuesday evening to articulate proposals for a research agenda.

Two aspects of the agenda are notable in reading through the recap of researchable topics presented below. Most if not all ideas generated at this meeting were proposed as topics of personal interest and commitment, that is, the participants were not tossing around ideas in passive fashion ("this needs to be addressed"), but instead the proposals were personal and proactive ("this is what I intend to do"). Secondly, the level of activity proposed by researchers external to the NCGIA was as high as the level of internal research activity. Many of the ideas expressed at the Tuesday night meeting have been pursued throughout the summer, and participants will be presenting their work at several national and international conferences itemized in the Deliverables section of this report.

At the beginning of the Tuesday evening session, threads from the discussions on the four themes were re-presented to the group, to guide formation of the research agenda. After some discussion by participants, and some modification of the original list, the final grouping within which to categorize researchable topics was agreed upon. Relationships between the original four themes and these categories are evident, although not a one-to-one match. Topics proposed by researchers are presented under relevant categories, including any illustrations they used in proposing their topics to the group. It should be understood that in many cases, a research topic might fit just as well under more than one of the headings below; the categories of research are not necessarily discrete. Proposals are listed alphabetically by participant.

4.1 Conceptual Frameworks and Prototype Displays

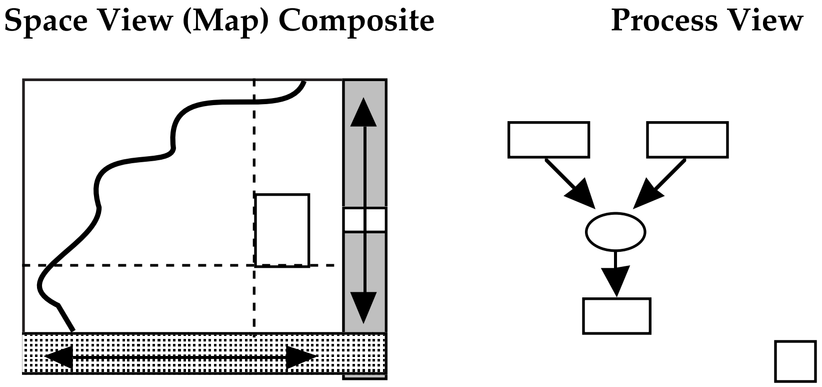
This category incorporates research proposals relating to formalization of knowledge about data quality, conceptual frameworks, paradigms and models useful for describing and monitoring data quality during GIS operations. Representational issues are partially included under this rubric. The specific focus lies with matching data quality components with a specific graphical symbolization method, that is, 'choosing the right tool for the job'.

Carl Amrhein
Clarify definitional ambiguities identified during the meeting. This is especially evident in references to data quality (vs error models, quality data, etc)

Mike Batty
I am also interested in a more specific problem of displaying networks, single planar directed graphs such as those involved in transport-interaction modelling. The size and complexity of such networks poses enormous problems of visualization in that data must be suppressed and displayed sequentially thus limiting the overall comprehension of the phenomena new types of representations are needed.

Kate Beard
Changes in data quality components produced by transformations of the data are often not apparent to users. A reference grid (mesh) may be one method of making such transformations visible. The reference grid could be displayed as a backdrop to the data of interest. A regular grid (squares, rectangles, or possibly other partitions) would signal an undistorted, unbiased state. Distortions of this grid could then indicate types of distortions in the data. For example zooming beyond an appropriate resolution of the data could be indicated by the fabric of the grid pulling apart or disintegrating.

Nick Chrisman
Develop and test a presentation method for mixed lineage, modeled in three windows: one displays the spatial partitions of areas with different sources or subsequent alterations, the other display a time line and a source and process flow chart. In the extreme of unitary sources, the spatial display is simple and the time line is a diagram of map algebra and other transformations. Complex spatial lineage can be represented using the spatiotemporal composite described by Langran and Chrisman (1988). The attributes of the composite will be maintained at the appropriate hierarchical level of spatial application. The composite detailed units will have pointers to those attributes.

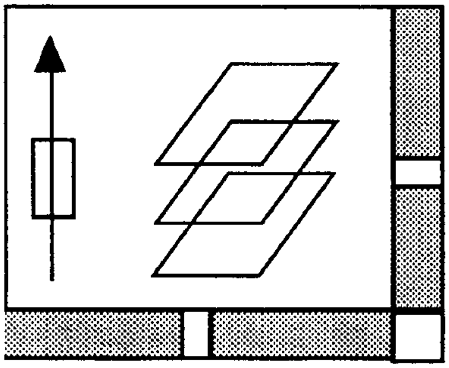


Space View (Map) Composite

Process View

The windows would be linked through dynamic behavior. Touching a temporal box would select the spatial unit. The composite map would select a path through the time model. To see an event, a space-time display would be required.

Time - Space Display



To make this worthwhile, a reasonable test case will be needed. The nautical chart is probably not the right one. Forestry may offer the spatial aggregation required.

Helen Couclelis

Visualization of product quality has as its goal to provide different types of users the kind of data quality information they need - no more, and no less. As a first step, construct a rough typology (4 - 5 categories) of spatial models (in the broad sense) by purposes of use, and then clarify data quality needs within each category. This may be approached through examination of kinds of queries and sequences of queries characterizing each category of model/purpose.

Noel Cressie

Develop a realistic model for locational errors.

Geoff Dutton

Using coverages (separate layers/ themes) for a test region, perform 1 to n stages of polygon overlay, keeping track of each feature, boundary and node’s positional uncertainty. Also perform attribute allocation using standard assumptions for categorical and numerical attribute data.

Geoff Dutton

Error in the derived features and attributes can be modelled and hopefully compared to some independent source of ’truth’. Statistical tests, as appropriate can be applied.

Geoff Dutton

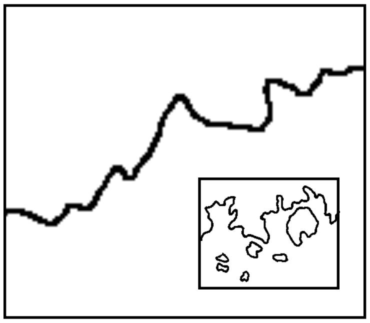
Generating, analyzing and visualizing lineage data concerning positional and attribute error.

Geoff Dutton

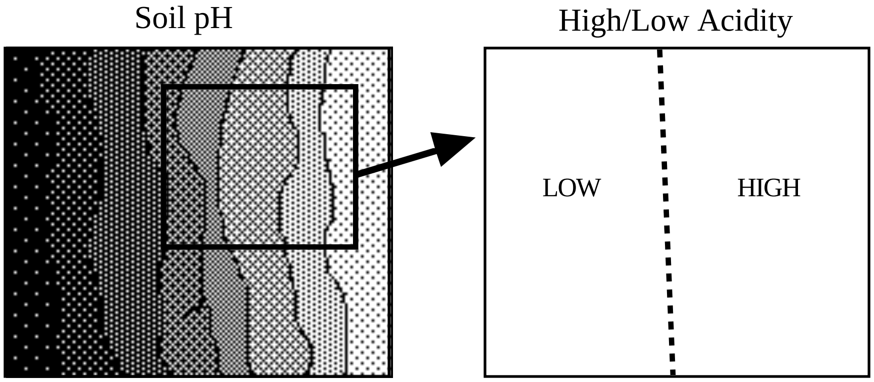
Line and point buffering tools can be used to visualize positional error, and standard choropleth techniques can be used to show attribute reliability (residual maps etc). New techniques to show how data quality degrades (or improves) using some form of lineage display such as a tree display or a 'multilayer lineage diagram. (This could build on work by Openshaw, Heuvelink, Chrisman, and others...)

Ferko Csillag
Adjusting resolution to classify uncertainty.
Intuitively, when you think something is wrong, you try to have a closer look. The idea in a spatial data base would correspond to something like zooming. Imagine a system that has two kinds of special functions: one that works like a video-memory, storing the most detailed zoom a user found necessary, and another which displays (A) fields according to some function of spatial autocorrelation, (B) objects with varying resolution.

Coast of Maine with
Special Interest in Castine



Note: Topology is NOT
maintained on the display



Note: When you pick an area, to form classes, high and low acidity can be easily identified at extremes. First, you don't classifiy the zone where classification accuracy is lower than a threshold...

Jay Feuquay
Investigate the use of transparency and translucency as indicators of quality

measures. Investigate the degree of obscurity/clarity related to these techniques.

Peter Fisher
I would like to further explore the effective display of fuzzy images derived from LANDSAT data or viewshed analysis (etc.).

Andrew Frank
What is the impact of scale change on data quality?

Andrew Frank
Specification of uncertainty and visualization tool to see what matches (fog, specula.., de-focusing an image). What are the different parameters, how are they used (effectively), how should they be aggregated during data aggregation processes.

Mike Goodchild
I'm interested in exploring several possible approaches to visual depiction of uncertainty on surfaces (e.g. elevation). One is to exploit rendering techniques - highly accurate areas could be glossy, and uncertainty could be matte. Another is to use the color space. We could color a surface using the standard color range (green, yellow, brown etc) and then depict uncertainty by reducing saturation, i.e. migrating toward [sic] the axis of an HLS space. Another is to color a surface, superimpose contours, and then broaden the contours in areas of uncertainty. Many of these ideas could be used to depict uncertainty in choropleth maps also. Finally, I am interested in methods for depicting the spatial dependence of errors. For example, autocorrelated errors might be shown using a series of realizations of an autoregressive process. By rapidly switching through the sequence the user would be given the impression of spatial dependence, but the precise positions of large errors in any one realization would not dominate the view.

John Kick
Use animation to show variation of output of natural resource numerical models, test the developers of the model to see if graphic depictions agree with the intent of the models.

Mark Litteken
Develop a system for describing, in some relatively compact format, how appropriate it is to use a particular map for various purposes. Then measure how helpful this is for choosing maps that are appropriate for a particular task and rejecting inappropriate maps.

Alan MacEachren
Is information about uncertainty more easily understood if visualization uses sequenced images, side-by-side images, or merged/overlayed images?

Matt McGranaghan
Develop matrix of data quality classes and mappable types.
Look for measures of data quality that cannot be mapped.
Consider typology of and communication about data quality to build mapping to visual variables. Redundancy may be lost. Look at time to learn codes, and big blunders in code interpretation (as in Cuff's (1973) study on misinterpretation of hypsometric symbols).

Jim Palmer
I'm interested in the use of photographs/video as realistic portrayals of spatial information. What is the fidelity of such stimuli? How can we enhance the understanding of their quality as data. The obvious extension of this work is to create realistic simulations.

Gerard Heuvelink
Before visualizing uncertainty, one must first define uncertainty (an error model is required). More research on error modelling is therefore needed.

Mark Monmonier
What theories can assist in the automated development of MEANINGFUL sequences (e.g. cognitive theories of effective writing style)? How can the sequencing most effectively address a user's information-seeking goals, background, level of expertise, etc.?

Peter Stringer
Examination of mental models of spatial uncertainty in for e.g. advanced geography students (e.g. Kahnemann and Trersky). Research might be accomplished in many ways, eg. by eliciting scenarios of how uncertainty data might be used in an interactive graphics environment. I am also interested in exploring the use of the metadata in data interpretation.

Denis White
Interested in the effects of scale on quality of sample data. How should regional estimates and sample data from (environmental) sampling be displayed through scale changes, considering constraints of sample size and possible confidentiality restrictions? By contrast with the perhaps more common notion that clarity should increase as scale enlarges, displays of sample data from probability samples need to convey less precision as sample size decreases. Even model-based estimates would presumably be similarly affected.

Joel Yan
Take a statement of data quality that is already done (ex. StatsCan or someone else) and consult with experts from NCGIA on graphics or visualization techniques. Then consult with data producers, who could test the data using data quality without vs. data quality with visualization tools embedded.

4.2 Software Development and Evaluation

Under this heading are the remainder of representational issues discussed under Theme 2. It should be emphasized that a conclusion made by several participants recognized that the community needs tools to help users create their own displays of data quality, as in a toolbox approach. This is not the same as deciding upon pre-determined symbolization schemes to represent particular data quality variables. The topics here emphasize using new technologies, hypermedia and animation, or modifying existing paradigms, most notably gaming and simulation routines, in order to provide users with an array of representational capabilities to explore and display data quality information.

Carl Amrhein
Link a large spatial micro-simulation model to some sort of real-time visualization 'machine'

Barbara Buttenfield
Render a multi-media document that has text discussing Spatial Data Quality and Spatial Data Transfer Standard concepts. Some items in the document would be "hot", i.e. clicking the mouse on a word or item would activate graphical displays (statistical chart, animation loop, tabular information) to further explain the concept. A follow-up project would utilize empirical testing to determine what impact such multi-media documents have on user comprehension.

Nick Chrisman
The virtual world interface offers a chance to avoid many of the classical constraints on traditional mapping. A virtual world has two basic forms: a transparent view projected onto a view of the real world and an immersed world totally created by computer. In both cases, data quality information can be the subject of the displays and interactions. Sound information offers a new range of possibilities to cartographic design, particularly with immersed stereo.

The technical possibilities are large, as a research project would have to be focussed fairly carefully. A first project would be a comprehensive review of the technology interpreted for its potential for GIS. One particular case would be developed as a prototype. A transparent field worker station would be one possibility.It may be easier to start with an immersed environment as an exploration of existing data.

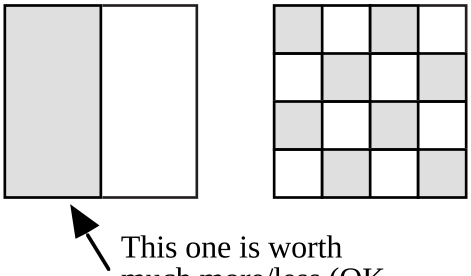
Helen Couclelis
Develop a graphic language for providing data quality information interactively in the context of typical sequences (this effort might be merged with other research on interfaces).

Jay Feuquay
Use particle systems to portray fuzzy boundaries, uncertainty, probability.

Ferko Csillag

1. Multipoly
   (Question: Can you play monopoly so that your real estate value is multiplied by the spatial autocorrelation of them?)

   - extension to 2d(3d), thresholds, change in value
   - merge function for size



This one is worth
much more/less (OK
given value?!)

Peter Fisher
I plan to enhance my existing animation program for simulating quality on soil maps, and to examine its effectiveness on communicating the data quality, and further, to expand its use to other data types - remotely sensed classified images.

Andrew Frank
Use games to test the quality of visualization (to have a near realistic problem - spatial decision making); then improve to include uncertainty and its visualization -- for example an urban dynamics simulation, like SIMCITY.

Mark Monmonier
How can an ordered sequence of graphics (i.e. maps, statistical graphs, data (term definitions, tables, juxtaposed views, etc.) promote an understanding of a“concern” such as

a) geographic correlation
b) risk associated with an environmental hazard
c) data quality, scale effects?

Which components of the “concern” are most suitable for this treatment?

Gerard Heuvelink
What are appropriate ways of visualizing quantitative attribute inaccuracy? How useful are dynamic representation (animations) of uncertainty such as that demonstrated by Peter Fisher?

Alan Saalfeld:
New tools and measures for image matching and pattern recognition for vector databases.

Goals: Automated feature matching
Measures of map difference

A tool to depict and zoom on large non-hierarchical geographic lattices (ordered by inclusion)

4.3. Database Management of Quality Information

This topic is a reduction of the original Theme 3, as the data model issues are incorporated under section 3.3. Monitoring data quality during GIS operations and during database update remains a primary challenge to the research community.

Carl Amrhein
Distribute and maintain benchmark or standardized data sets as discussed in NCGIA Initiative 1.

Barbara Buttenfield
Generalize a public domain data set of information that has quality information embedded in multiple layers. This could be distributed by NCGIA on a cost-recovery basis, with data provided by federal agencies. A similar project for multi-scale data worked quite successfully during a previous research initiative.

Noel Cressie
Suppose the elements of the spatial data base have measures of quality associated with them. The spatial data are often modified, smoothed , or transformed. What quality measure can we associate with the transformed spatial data? When the transformation are linear (eg. aggregation) the quality measure is typically a simple function of the original quality measures. For nonlinear transformation, the problem is very difficult. Possible approaches I would take are:

1. Develop a spatial bootstrap
2. 2D spectral methods

Bob Maki
What are the most effective techniques for storing, accessing, and displaying textual metadata within the context of data structures and associated data bases? Commercial industry's immediate concern is in data quality management and access. Once this is understood, the issue of visual communication could, perhaps, be addressed more effectively.

Jim Palmer and Ferko Csillag
It would be nice if USGS (and other data producers) would produce a metadata file at the front end of their distribution media. Such a file could be used by GIS vendors to create the data frame into which the data will be put. It would include information like:

1. resolution, scale
2. precision
3. variability (at least min,max, s.d., covariance, number of cases)
4. spatial autocorrelation
5. lineage, both a textual file and a graphic file similar to the update sketch on quad sheets
6. location and values of control points
7. reliability / ground truth results (eg. table of classification purity)

An analogy would be that vendors of GIS applications need to supply documentation / lineage of their algorithms and precision information.

Joel Yan
Exploit lineage information and data directories to better educate users of what is archived in a lineage report and how to use it. Perhaps do this using multi-media. Provide users with information on what lineage is, how to use it. If this can be embedded with the data products, include information on future lineage, eg. dates expected for future versions, updates of the data.

4.4. User Needs

User expectations about data quality surfaced over and over during the Specialist Meeting as requiring immediate research attention. Without a clear sense of what GIS users expect to discover in a data quality display, or even how users comprehend data quality concepts and implications, software and database solutions cannot be optimized. The role of visualization of data quality in decision-making and dispute resolution should also be explored; herein lie ties to NCGIA Research Initiatives on Spatial Decision Support and on Legal and Cadastral Issues. Finally, the development of educational tools and documentation should improve user awareness about the incorporation of data quality information into spatial databases.

Carl Amrhein
Compile and distribute a bibliography of examples/efforts involved with visualization problems, especially in a dynamic environment.

Mike Batty
We need a more definitive study or even catalog of design principles for visualization in general, visualization of data quality in particular. This would draw together principles of 'best' cartographic design and other 'best' forms of

graphic design. Such a study would also anticipate new principles of design posed by the new media, particularly those relating to the juxtapositions of graphics, text and numeric data in various windows on the visual display unit.

Andrew Frank
How is uncertainty data effectively used? - measures are more a reflectance of what we can measure then how these measures are used. What are effects on the decision maker?

Dan Griffith
What visualization tools do researchers currently use to evaluate data quality?

Project 1. survey users acquiring TIGER files

Project 2.

Step 1 compile benchmark data sets (suggested in I1)
Step 2 deliberately introduce certain errors
Step 3 distribute data to selected users for analysis
Step 4 survey users to see what they did

Mark Litteken
A compendium of error/quality classes, and examples of visualization techniques for each class. Test each technique with real subjects and determine which techniques best visualize each error/quality class.

Geoff Loftus
General Question: How do people “understand” a geographical area, any point of which is characterized by multidimensional “goodness values”?

General Idea: Split Screen with a) a map of some area assuming you indicate a particular point and b) a plot relating normalized “goodness values” as a function of dimension. Note that the pattern of values represents the overall quality corresponding to the indicated point. As you move a cursor around the map the pattern changes. Would people internalize the dimensional structure of the map space by mentally correlating the changing goodness value pattern with change in geographical location? In the empirical test to evaluate, you allow people to solve some problem (e.g., where should one optimally place a waste dump), then assess transfer of user comprehension to other problems. Is transfer mediated by learning the dimensional structure of the map space?

Alan MacEachren
Does visual presentation of information about uncertainty alter spatial decision making (or hypothesis formulation)? If so, how?

Matt McGranaghan
Determine whether (and which) people have any idea how to use quality information if it is displayed. Find the notions of data quality held by people.

Matt McGranaghan
See whether, given data quality codes for map objects (rather than redundant identification of the objects) people can use the display to assess quality (rather than only confound object identification). How long does it take to get users to correctly interpret the displays?

Jim Palmer
One of the several areas I research and teach about is dispute resolution, particularly environmental conflicts. Data issues and developing a common data base is one of the standard hurdles in environmental dispute negotiation. There is always disagreement on what constitutes appropriate or acceptable data, measurement, uncertainty, and the like. In the real world, decisions must be made in the face of ignorance and data uncertainty. I would like to look at the influence of data quality visualizations on negotiations like:

a) locating solid waste incinerators and landfills
b) allocating/obtaining water for a large region like
Southern California from Colorado
c) preparing an air quality management plan
d) addressing toxic waste hazards such as rehabitation of Love Canal

Jim Palmer
I'd like to study how people (professionals, students and lay people) use spatial data/GIS to solve problems and how the inclusion of data quality information influences their thinking. One basic objective would be to bring people's confidence in analysis into appropriate/realistic context.

Alan Saalfeld
A "challenge" along lines of DIMACS Network Flow Challenge. (Dimacs is Discrete Math & Compiler Science center at Rutgers/Princeton, funded by NSF). This challenge involved an open call to find and evaluate efficient and robust implementations of algorithms for the minimum flow problem, the maximum flow problem, the assignment problem, and nonbipartite matching. A conference was held to present submitted solutions, which could involve algorithm implementation, creation of a test data set, or testing an existing algorithm on an unusual architecture. Translating this into the context of visualizing data quality, one might envision a challenge to design and evaluate models of error propagation, design and evaluate displays of certainty and quality information, and so on.

Peter Stringer
A programmatic evaluation of data quality visualization in naturalistic settings (i.e. naturalistic with respect to users, stimulus sets, task structure, response mode) with a view to informing design, interfaces, documentation and training/education.

Joel Yan
Collect a compendium of standard and good examples. Collect a wide set or compendium of statistical graphics / visualization examples for data quality that are good and not good

- 2a. perhaps offer recognition or awards for submissions
- 2b. comment on submissions in the Compendium - provide a good taxonomy of techniques
- 2c. procedures and publish a compendium of successful or unsuccessful techniques (solicit usage notes along with submissions)
- 2d. encourage the community to test compendium examples to improve and refine the better techniques in a later phase
- 2e. extract general principles and propose some general guidelines

Joel Yan
Understand better how users perceive data quality and why they have troubles understanding data quality.

5 Proposal for Initiative Deliverables

5.1 Newsletter

A newsletter is being organized to provide notice of announcements, publications, and events related to the research topics. The newsletter may also provide a forum for updates on recent research activities as well as on the planned compendium of visualization products. Two issues per year (May and November) are anticipated to be available through electronic mail and/or page layout form with graphics. Geoff Dutton has agreed to act as editor. Items for the newsletter will include progress reports on I-7 projects and research, notification of upcoming conferences related to I-7 topics, announcement of relevant publications, newsworthy items and requests for collaboration on new projects that may develop. The newsletter will be sent out in hardcopy form which may included images from time to time; and an ASCII version will be distributed on GIS-L List server and announced on other electronic bulletin boards.

5.2 Bibliography

An annotated bibliography is being compiled jointly by graduate students at Maine and Buffalo. Sarah Clapham at Maine and Victor Wu at Buffalo are the graduate students assigned to coordinate this effort. The list explores literature across disciplines considering notions of data quality and visualization concepts and techniques from statistics, scientific computing, medical imaging, automated manufacturing, dynamic process analysis, and graphic design. The list currently contains over 100 entries. In addition to the hardcopy version, an electronic version will eventually be compiled using the bibliographic utility EndNote. It is anticipated that the electronic bibliography may grow through publication notices posted in the above newsletter.

5.3 Conferences

In an effort to support research presentations, NCGIA Associate Directors at Maine and Buffalo have agreed to cover conference registration costs for paper and special sessions acknowledging I-7 research efforts. Special paper sessions are being arranged at several upcoming conferences to report research generated during the I-7 Specialist Meeting.

A conference on "Spatial Issues in Statistics" will be hosted by Statistics Canada in Ottawa, Ontario, Canada, November 12-14, 1991. The conference will include workshops on record linkage and spatial auto-correlation. Joel Yan from Statistics Canada has organized a paper session focusing on I-7 research including papers to be presented by Mike Goodchild (error and accuracy issues) and Barbara Buttenfield (discussing the research agenda formulated at the Specialist Meeting). A third paper will be presented by a researcher from Statistics Canada.

Special sessions are being organized for the April 1992 meeting of the American Association of Geographers. Barbara Buttenfield has organized a Cartography Specialty Group special paper session on Cartographic Issues. Participants will include Mark Monmonier (Syracuse), Alan MacEachren (Penn State), and Matt McGranaghan (Hawaii). The papers will be submitted for review and possible publication as special issue of Cartography and GIS or Cartographica.

A special session is also being organized by Dan Griffith (Syracuse) on "statistical issues and visualization" as a GIS Specialty Group paper session. He is currently arranging details of the session.

Kate Beard is coordinating presentations for the "5th International Conference on Spatial Data Handling" to be held in Charleston, South Carolina in August, 1992.

A joint symposium on NCGIA Initiative 7: Visualization of Spatial Data Quality and NCGIA Initiative 13: User Interfaces for GIS is proposed to take place in summer, 1992. The focus will be on user access to data quality information. The symposium is anticipated to involve 15-20 participants and include both research papers and demonstrations.

5.4 Compendium for Data Quality Examples

A compendium of representative examples for the visualization of data quality was suggested by Joel Yan. In addition to exemplary products, the collection potentially may include documentation on procedures used to generate the visualization as well as evaluations and expert commentary on the designs. The Compendium will include images displaying 'good' and 'bad' examples of data quality displays, experimental graphics, and GIS display screens illustrating techniques of displaying metadata and data quality information. Recognition for submittal was suggested as a means to encourage contributions. The compendium will be coordinated by a post doctoral researcher (William Mackaness) at Maine.

The proposal is to collect and periodically disseminate collections for the compendium in a format such as VCR videotape. Video format will accommodate electronic images in a variety of formats (Postscript, TIFF, PICT, Bitmap) from a variety of platforms and will also be easy to mail to participants. Video additionally can display static and dynamic graphics (eg., animation, broadcast graphics). Distribution to participants and other interested people will be covered on a cost-recovery basis, and announced in the I-7 newsletter and the NCGIA newsletter. Final decisions about formats, image resolution and data file volume need to be worked out before the project commences. Detailed information on the format of compendium and instructions for submitting graphic material will be specified at a later date and posted in the newsletter mentioned above.

5.5 Journal Special Issues

Publications are planned for journal special issues to include selected position papers, and a summary report of the specialist meeting. Helen Couclelis has tentatively agreed to act as special editor for a special issue of Environment and Planning B. Peter Fisher, Editorial Board Member of Computers and Geography, has suggested developing a special issue for that journal. In addition, papers presented at the upcoming AAG special sessions will be submitted to a refereed journal in the cartographic discipline for consideration as a special issue.

6 Summary of Research Priorities

Topics discussed at the Specialist Meeting can be summarized under two broad categories. First, one may address the issue of visualization tools that may be applied to display data quality information. Second, one must consider the impact these tools may have upon decision-making, analysis, and comprehension of data quality components.

6.1 Visualization Tools

The range of visualization tools available for representing data quality is very broad, and incorporates cognitive and perceptual aspects as well as technological (hardware and software) solutions. Frameworks for itemizing this range were cited from a variety of domains, including cognitive linguistic frameworks, perceptual theory, and inventories of existing libraries of computer graphics plotting commands. There is some question as to the discretization of internal and external tools.

The importance of interactivity was emphasized, and impact of external tools upon the internal representation and subsequent improvement of understanding must not be discounted. Hardware and software tools should be developed that facilitate user manipulation of data quality, and these may be implemented in toolbox fashion. Users must also be expected to bring with them some expertise in the domain of the problem or data set they wish to explore, and the focus in all group discussions was on exploration in lieu of confirmation. That is, the representation of data quality was discussed as a process of data exploration, rather than as a type of hypothesis testing and confirmation of *a priori* assumptions.

Design tools available for graphical depiction were consistently cited using the Bertin framework of visual variables, and groups emphasized the importance of training users and vendors to implement sound principles of design in their displays and GIS display software. The notion of protecting users from themselves (constraining users against making poor graphical depictions) seemed to take a lesser priority in relation to the need for providing flexible and interactive representational tools to explore spatial metadata.

The need for tools to represent error models led to a general group discussion about what constitutes an error model. Mike Goodchild offered a definition that an error model is a (sometimes stochastic) process capable of generating many versions of a map that differ only in terms of their error. He qualified this definition that the error differentiation is intended in an inferential sense. Noel Cressie commented that it is important to accept that one cannot always invert the process, that is, given a map one can generate many manifestations of its associated error surface; however, a given map of an error surface cannot always be used to reconstruct its associated (map) data source. This topic re-appeared throughout the Specialist Meeting, and it became apparent that the many disciplines represented had arrived with somewhat discrepant views about the nature of an error model. This particular general group discussion about what constitutes an error model provides a fairly succinct summary of the debate, which was not completely resolved at the end of the four days. However, the group reached quick consensus on the point that all error models are mappable, and for this reason the Goodchild definition took a strong position during the ensuing day's discussions.

6.2. Impacts on the User

Visualization and its utility for data quality may be viewed from a primarily cognitive as opposed to graphical perspective. From this point of view, one should accept that discovery and innovation have traditionally involved visual thinking. This argues for the importance of understanding information processing capabilities which rely on visual tools and skills. Scientists and engineers define problems and design solutions through observation, imagination, and logic.   Evaluation of user demands for data quality information requires attention to the internal (perceptual and cognitive) mechanisms by which spatial and temporal patterns are interpreted.

Equally important is the need for sensitivity to the domain of the GIS application. For example, reliability associated with a routing of emergency dispatch vehicles will likely vary with each link of the route; this information must be presented with high precision and in a short timeframe. Reliability variations associated with the environmental impact of a timber clear-cut operation will vary with the polygons in a categorical coverage, as opposed to varying between discrete links in a routed network. As the domain varies, so must the presentation of data quality information, and so must its evaluation.

The acceptance of visual displays as true may complicate acceptance of graphics depicting variations in accuracy or certainty. Many decision-makers may not utilize metadata displays, and this attitude will likely become more prevalent at higher bureaucratic levels, where the amount and detail of provided information will decrease even as the impact of decisions based on that information will increase. The importance of training and education about data quality, its variations, and its impacts was emphasized throughout the discussion. Other proposals for evaluating data quality included development of GIS labs focusing on evaluatory research, and the development of a compendium of good and bad examples of data quality displays to be shared in the general GIS community.

7 References

Beard M. K. 1989. Use error: The neglected error component. **Proceedings**, AUTO-CARTO 9, Baltimore, Maryland, March 1989: 808-817.

Bertin, J. 1983. **Semiology of Graphics.** Madison: University of Wisconsin Press.

Buttenfield, B . P. and Ganter, J. H. 1990. Visualization and GIS: what should we see? what might we miss? **Proceedings**, 4th International Symposium on Spatial Data Handling, Zurich Switzerland, July 1990, vol.1: 307-316.

Buttenfield, B. P. and Mackaness, W.A. 1991. Visualisation. Chapter II.a.4 in Maguire, D. Goodchild, M.F. and Rhind, D (Eds.) **GIS: Principles and Applications**. London: Longman Publishers Ltd. (forthcoming)

Chrisman, N. R . 1983. The role of quality information in the longterm functioning of a geographic information system. **Cartographica** 21 (2/3): 79-87.

Cuff, D. 1973. Shading on Choropleth Maps: Some Suspicions Confirmed, **Proceedings of the Association of American Geographers.** vol. 5: 50-54.

Ganter, J. H. 1988. Interactive graphics: linking the human to the model. **Proceedings**, GIS/LIS '88, 230-239.

Lakoff, G. 1987. **Women, Fire and Dangerous Things, What Categories Reveal About the Mind**. Chicago: The University of Chicago Press.

Lakoff, G. and Johnson, M. 1980. **Metaphors We Live By.** Chicago: The University of Chicago Press.

Langran, G. and Chrisman, N. 1988. A Framework for Temporal Geographic Information. **Cartographica.** Vol. 25/3: 1-14.

Lanter, D. P. and Veregin, H. 1990. A lineage meta-database program for propagating error in geographic information systems. **Proceedings** GIS/LIS '90, Anaheim, California, November 1990, Vol.1: 144-153.

MacEachren, A. E., Buttenfield, B .P., Campbell, J. C. and Monmonier, M. S. 1991. Visualization. In Abler, R. A., Olson, J. M. and Marcus, N. G. (Eds.) **Geography's Inner World**. Washington, D. C.: AAG (forthcoming).

Moellering, H. 1988. The proposed standard for digital cartographic data: report of the digital cartographic data standards task force. **The American Cartographer,** 15(1) (entire issue).

Sinton, D. 1978. The Inherent Structure of Information as a Constraint to Analysis. **Harvard Papers on Geographic Information Systems** (ed. G. Dutton. Reading: Addison-Wesley.

Tukey, J. 1977. **Exploratory Data Analysis**. Reading: Addison-Wesley.

Weibel, W. R. and Buttenfield, B. P. 1988 Map design for geographic information systems. **Proceedings,** GIS/LIS 88, November 1988, San Antonio, Texas vol.1: 350-359.

Appendix A Specialist Meeting Agenda

Saturday June 8

2:00, 5:00, 8:00 pm Vans depart Bangor International Airport for Castine
6:30 Dinner at Maine Maritme Academy

Sunday June 9

8:45	Introduction - Initiative objectives
10:00	Break
10:15	Small group discussions Topic 1 Data Quality Components
11:30 - 12:30	Lunch
12:45	Review of small group discussions
2:15	Break
2:30	Small group discussions Topic 2: Representation Issues
5:30 - 6:30	Dinner
8:00	Review of small group discussions

Monday June 10

8:45	Small group discussions Topic 3: Error Models, Database Issues
10:00	Break
10:15	Review of small group discussions
11:30	Lunch
1:00	Visualization Presentations
2:30	Break
2:45	Small Group discussions of Topic 4: Evaluation and User Needs
4:00	Review of small group discussions
6:30	Dinner at Castine Inn

Tuesday June 11

8:45	Boat trip to island in Penobscot Bay
5:30	Dinner
8:00	Large Group discussion of research agenda

Wednesday June 12

9:00	Recap and Discussion of research agenda
11:00	Break
12:00	Box Lunches - Vans return to Bangor Airport

Appendix B Participants, Visitors, and Rapporteurs

Participants

Carl Amrhein
Dept. of Geography
University of Toronto
Toronto Ontario M5S 1A1
CANADA
AMRHEIN@UTORONTO

Michael Batty
NCGIA/Dept. of Geography
SUNY-Buffalo
Buffalo NY 14261
716-636-2545
GEOBATTY@UBVMS.BITNET

Kate Beard
NCGIA/Dept. Surveying Engineering
University of Maine
Orono ME 04469
207-581-2147
BEARD@MECAN1.BITNET

Barbara Buttenfield
NCGIA/Dept. of Geography
SUNY-Buffalo
Buffalo NY 14261
716-636-3834
GEOBABS@UBVMS.BITNET

Nick Chrisman
Dept. of Geography, DP-10
University of Washington
Seattle WA 98195
CHRISMAN@MAX.U.WASHINGTON
.EDU

Helen Couclelis
NCGIA-Santa Barbara
UCSB-3510 Phelps Hall
Santa Barbara CA 93106
805-893-8224
COOK@POLLUX.GEOG.UCSB.EDU

Noel Cressie
Iowa State University
Department of Statistics
Ames IA 50011
S1.NAC@ISUMVS.BITNET

Ferenc Csillag
Department of Geography
Syracuse University
Syracuse, NY 13244
FCSILLAG@SUNRISE.BITNET

Geoff Dutton
150 Irving Street
Watertown MA 02172
QTM@CUP.PORTAL.COM

Jay Feuquay
EROS Data Center
Sioux Falls SD 57198
FEUQUAY@SUNJ.CR.USGS.GOV

Peter Fisher
Department of Geography
Kent State University
Kent OH 44242
PFISHER1@KENTVM.BITNET

Andrew Frank
NCGIA/Dept. Surveying Engineering
University of Maine
Orono ME 04469
207-581-2174
FRANK@MECAN1.BITNET

Michael Goodchild
NCGIA/Dept. Geography
UCSB-3510 Phelps Hall
Santa Barbara CA 93106
805-893-8224
GOOD%POLLUX.NCGIA.UCSB.EDU

Dan Griffith
Department of Geography
Syracuse University
Syracuse NY 13244
GRIFFITH@SUNRISE.BITNET

Virginia Hetrick
Office of Academic Computing
5628 Math Sciences Addition
UCLA
Los Angeles CA 90024
CUSGRAF@UCLAVMXA.BITNET

Gerard Heuvelink
Dept. of Physical Geography
University of Utrecht
P.O. Box 80.115
3508 T.C. Utrecht
The Netherlands
IAAHEUV@CC.RUU.NL

John Kick
John M. Hanley Fed. Build.
100 S. Clinton St., Rm. 771
Syracuse NY 13260
V069E9VW@UBVMS.BITNET

Mark Litteken
IBM Corporation
44V/975, Building 970
Kingston NY 12401
VOICE MAIL: 914-385-3314

Geoff Loftus
Department of Psychology
University of Washington
Seattle WA 98195
GLOFTUS@
MILTON.V.WASHINGTON.EDU

Alan MacEachren
Department of Geography
Penn State University
University Park PA 16802
NYB@PSUVM.BITNET

Bob Maki
ESRI
380 New York St
Redlands CA 92373
will acquire email address soon

Matt McGranaghan
NCGIA/Dept. Surveying Engineering
University of Maine
Orono ME 04469
MATT@UHUNIX.UHCC.HAWAII.EDU

Mark Monmonier
Department of Geography
Syracuse University
Syracuse NY 13244
MON2IER@SUNRISE.BITNET

Jim Palmer
Coll. of Env. Sci. & Forest.
SUNY-ESF
Syracuse NY 13210
ZOOEY@SUVM.BITNET

Alan Saalfeld
Bureau of the Census
Statistical Research Division
U.S. Department of Commerce
Washington DC 20233
SAALFELD@TOVE.CS.UMD.EDU

Peter Stringer
Policy Research Institute
105 Botanic Avenue
Belfast BT7 1NN
NORTHERN IRELAND
UAPH87@UK.AC.ULSTER.UCVAX

Denis White
Env. Research Lab, U.S. EPA
200 SW 35th St
Corvallis OR 97333
DENIS@ORSTATE.BITNET

Joel Yan
Statistics Canada
FAX: 613-951-0569

Visitors

Max Egenhofer
NCGIA/Dept. Surveying Engineering
University of Maine
Orono ME 04469
MAX@MECAN1.bitnet

Werner Kuhn
NCGIA/Dept. Surveying Engineering
University of Maine
Orono ME 04469
KUHN@MECAN1.bitnet

David M. Mount
Department of Computer Science
University of Maryland

Rapporteurs

Sarah Clapham
NCGIA/Dept. Surveying Engineering
University of Maine
Orono ME 04469
SarahB@MECAN1.bitnet

Ricardo Moreno
NCGIA/Dept. Surveying Engineering
University of Maine
Orono ME 04469
Ricardo@MECAN1.bitnet

Todd Rowell
NCGIA/Dept. Surveying Engineering
University of Maine
Orono ME 04469
Rowell@MECAN1.bitnet

Diane Schweizer
NCGIA/Dept. Geography
UCSB-3510 Phelps Hall
Santa Barbara CA 93106
Swizzla@VOODOO.bitnet

Nicole Soltyka
NCGIA/Dept. of Geography
SUNY-Buffalo
Buffalo NY 14261
V113LU75.UBVMS.bitnet

Gary Volta
NCGIA/Dept. Surveying Engineering
University of Maine
Orono ME 04469
GaryV@MECAN1.bitnet

Chris Weber
NCGIA/Dept. of Geography
SUNY-Buffalo
Buffalo NY 14261
ESRISTO8@UBVMS.bitnet

Comments on Visualising Data Quality

Carl G. Amrhein
Department of Geography and
Institute for Land Information Management
University of Toronto
Toronto, Ontario M5S 1A1

Introduction

There is no longer any serious debate about the ability of computer visualisation to improve the transfer of information produced by models, to the modeller. With respect to the four themes mentioned in the letter of invitation, the discussion that follows focuses on the visual representation of modelling results.

The availability of ever-increasing amounts of inexpensive computing power have permitted a rapid increase in the size and sophistication of simulation models of various spatial processes. These simulation models range from physical systems to labour markets and economic systems, and seek to replicate in an idealized (or controlled) context what are thought to be the key generating mechanisms that underlie the observed patterns. However, while the available computing power has encouraged modellers to create "better" models, the modeller is still faced with limited means of analyzing the output. For example, imagine a micro-simulation model of a labour market that includes 10,000 workers and 10,000 jobs; each worker has a unique location and characteristics, and jobs are clustered in 1,000 locations. The volume of data generated with each iteration of the model is enormous. The analyst, using summary statistics and perhaps simple graphical output, is simply unable to effectively access all of the information present in individual output values, or in various aggregations of the output values. There is simply an overload of information.

Computer visualisation techniques provide a solution. The combination of the colour, size, shape, and location of symbols creates images that can be repeated in quick succession. Such a sequence permits the rapid and effective representation of large volumes of information.

The range of display issues that hinder the effective communication between the modeller and the model has obvious overlap with cartographic communication. These issues will not be discussed here, However, it is clear that there is a need to arrive at some common language that analysts from a variety of backgrounds can use to communicate findings. Of interest is exploring the opportunity to Join the capability of computer visualisation with the potential of evolutionary economic models to further our understanding of the operation of dynamic spatial systems.

Analyzing simulation results

Conducting effective experiments using simulation models requires an examination of the model's behaviour under a variety of conditions. These idealized conditions are created by varying the values of key parameters, thus generating a distribution of results. This distribution of results effectively describes the response of the entire system to perturbations in one (or more) value. Visualisation offers great potential in the examination of the accumulated effect of small changes in values, propagating through the system.

Problems

1. Clearly the greatest technical problem facing the implementation of visualisation techniques in the context of simulation models is one of integrating the graphics software into what is often "homegrown" code. In the ideal situation, results are presented to the screen as they are generated, rather than being stored for later display. Given the often large amounts of data that are involved, the file management can become quite cumbersome. However, the realtime display of results requires fast display along with fast data transfer.

2. A real-time visualisation problem deals with all of those issues of data display that have been around for a very long time, but in the context of potentially enormous amounts of data that must be displayed very quickly. While cartographers may have solved some of the problems, those unfamiliar with the cartographic literature may be uninformed of these advances. Many issues remain: the use of colour, the rapid update of images that occurs in real-time display, the appropriateness of certain symbols, rotating images to achieve maximum effect, the interaction between vision and machine, and others. If these issues have been resolved, then a bibliography might easily be constructed and distributed. If not, then more basic research can be undertaken.

3. Finally, a very routine issue concerns the dissemination of research results. Journals cannot distribute VHS tapes and most conferences do not provide televisions and tape players without additional cost. Of what use are sophisticated movies if they can be

seen by very few people? Of what use are the results of visualisation if they are reduced to two dimensional pictures, usually in black and white, for publication in journals?

Analyzing Data Quality

The framework described for the sensitivity analysis in the context of simulation models is similar to the methodology one might employ in assessing the quality of data in a data base. A simulation process that explores the nature, magnitude, and distribution of error in relation to spatial process models can add greatly to our understanding of the impacts of variable data quality on the outcome of spatial analysis. The classic statistical framework in such a study requires a large number of replications with key values selected from some distribution. The outcomes of the analyses (in the form of some distribution) are then analyzed for significant variation.

The idea underlying this framework is simple: since we have no reliable information about the distribution of error in the secondary data (or worse) that usually comprises spatial data bases, there is great difficulty in establishing a level of confidence in our results. For example, a map is digitized from some paper document whose lineage is not well known. This digital map is combined with data derived from census data. While the aggregate level of quality in the census results is known, the variation in quality over space (and scales of resolution) is not well known. A potential surface of some sort is calculated after some form of interpolation occurs. How do we establish confidence in such a result? Of what use are subsequent analyses in which this trend surface is one of many inputs?

Attacking the problem

There is an enormous range of issues that can be included in this section. What is needed however, is some agreement on a methodology with which issues relating to data quality can be addressed. Before much progress toward this goal can be made however, a precise get of operational definitions relating to various aspects of data quality is needed. Such a task is itself an important research undertaking. For my purposes, data quality is equivalent to error in the data susceptible to analysis using standard (and perhaps not so standard) statistical tools. This definition provides a context within which data problems can be discussed (not all "datasituations" are problems in all applications).

The area of interest here relates to the use of data in spatial process models; for example, optimal network models, forecasting models, and location allocation models. Ibis style of analysis requires that data from a variety of sources are combined in a multi-stage analysis. Two problems are present.

The first problem relates to the issue of propagation of the error through the various stages of the analysis. For example, in a location allocation problem, an error in the digitization of boundary points produces errors in centroids which produce errors throughout subsequent analyses.

The second problem is linked to the first and relates to a remedial process. While fieldchecking may adequately produce insights into the nature and degree of error in files of points related to surface features, it is not reasonable to attempt to verify every piece of data from secondary sources (for example, census files). Rather, a statistical filter is needed that permits some form of "correction" to be introduced into the analysis. Such a process does not eliminate, the error, but rather manages it.

Ultimately, the creation of such filters will require experimentation. Initially, however, it is important to agree on problem statements, and operational definitions before spending time and money on experiments. This need was very evident in the responses Dan Griffith and I received to some preliminary ideas we circulated recently. Finally, in order to facilitate progress in research related to the quality of spatial data, a set of data sets, disseminated and maintained by the Center, would provided a common basis on which to compare results. A similar proposal for standardized data sets came from the specialists meeting for Initiative One. I do not know whether such sets have, in fact, been collected and/or are available for use.

NCGIA Initiative 7
VISUALIZATION OF DATA QUALITY

Position Paper

MODELLING SPATIAL DATA THROUGH VISUALIZAT71ON

Michael Batty

NCGIA, SUNY-Buffalo,
301F Wilkeson Quad, Buffalo, New York 14261
Tel: 716-636-2545: Email geobatty@ubvnis

There are two broad questions which I think should be at the forefront of research into new methods of visualizing data in science. The first relates to the extent to which the current theoretical framework which conditions the science we are involved in, influences the definition and collection of data, and its interpretation as information, while the second deals with the nature of theory itself and the extent to which we might validate or falsify it through the data we have available. In this short note, I will begin by sketching my thoughts on these questions and then continue by relating these to my own interests in geographical and spatial data and the extent to which contemporary methods of visualization might lead to new insights which constitute scientific progress. This in turn enables me to outline some of the ideas which I am exploring at present with respect to how computer graphics can be used to enrich the modelling and simulation of urban systems which is my domain of interest and research.

I regard it as a truism that whenever we define data we bring to bear some theory which we have already accepted as a means by which to interpret such data. Data is inevitably and necessarily theory-laden and there are few if any circumstances in which we are able to approach data with no theory in mind, at least implicitly. This is easily illustrated by the notion that as soon as new data becomes available we immediately seek to set it in some confirmatory context, either validating or falsifying the often implicit theory which we use in interpretation. Although we all accept the distinction between inductive and deductive scientific method, the extent to which we adopt one or the other in the development of better explanation is always a matter of degree, involving a mix of induction and deduction which characterizes the to-ing and fro-ing of searching for more and better data from which to confirm, falsify and hence change the hypotheses in question.

The second issue relates to the types of theory which categorize our fields of interest. In my view, most contemporary theory certainly in the social sciences, and possibly in the physical sciences too, is characterized by ideas which are many orders of magnitude richer than the data which we are ever able to identify and measure. Thus empirical validation of theory usually consists of simplifying the theory through a hierarchy of ’levels defining models whose predictions embody some but never all the elements of the base theory. Such models enable theory to be tested against data but only ever at a restricted number of important points. In short, most theory can only be confronted with data where the theory manifests itself is terms of the available data and thus science is forever conditional in that no theory can ever said to ’proven’, only that the theory is consistent, so far, with data describing the domain of application. Much of this can be easily demonstrated from classical and contemporary physics. From the Seventeenth century on, physical theory came to be regarded as of universal applicability but this was and always remains an act of faith for the observations against which such theory can be tested always form an infinitesimally small number in contrast to potential observations. In fact this view of science is now fairly widely accepted. It is the basis of Karl Popper’s positivism in which the quest to develop theory must be seen as a continuing process of conjecture and refutation (Batty, 1980). Thus, theory in the physical sciences is always orders of magnitude richer than the data which can be found to confirm it; and perhaps it is of even greater import in the social sciences where it is often more difficult to identify and collect appropriate data and where it is usually impossible to establish the sorts of controls on experimentation which characterize the physical sciences.

What has all this got to do with visualization in general and the development of visualization in our own domain which seeks to develop theory and applications involving geographic space ? There are several issues. First, the line between theory and data is inevitably blurred in that data is always some manifestation of some theory and vice versa. Envisaging the quality of information is thus a process of evaluating the essence of some theory in terms of the data available. Second, if the theory and data are so inextricably intertwined, it does not make sense to restrict discussion to data per se. Methods of developing insights from such data thus correspond to insights through theory and the extent to which the quality of the data makes this possible is part and parcel of the to-ing and for-ing in the search for better data.

Theory and data represent different sides of the same coin. We may be able to approach visualization from the viewpoint of one or the other - data or theory - but the types of visualization we may develop depend intimately on pattern and structure in the theory for such pattern only becomes meaningful in the context of theory. Finding other patterns may be of interest but only if they can be interpreted in terms of the theory in mind. Rarely has science progressed through serendipitous events involving truly accidental discovery for as Poincare so clearly articulated 'Discovery favors the prepared mind!' An obvious consequence of this is that I do not believe there are any situations we are searching for patterns in data simply for their own sake without any theory in mind and certainly without any idea or expectation of what we are likely to find. In fact, I would hazard an opinion that it is in those areas where large quantities of data have been available and researchers have resorted to indiscriminate searching for correlations in data that least progress has been made, as for example in the development of a social ecology of the city which provided one of the spurs to quantitative geography in the 1950s and 1960s.

A third issue in exploiting the power of visualization in evaluating the quality of data concerns the various types of visualization which might be appropriate. At present because visualization is such a new art form, most discussion is dominated by simply demonstrating that visualization is possible and that hitherto unmanageable data sets can become manageable when organized in the visual media. Simply enabling large data sets to be comprehended through graphical portrayal has been the dominant mode so far, and there has been little discussion of systematic procedures for visualization consistent with our field, its methodology and its potential for visualization in the first place. We are still at the stage where visualization is a novelty and interesting for its own sake, but to make some progress we need to structure the process and inquire into what the best ways forward are. I feel that if the Initiative is able to agree on some such directions for research along these lines, then the meeting will have been successful. One obvious way of visualizing data is to exploit the structure of the theory which is implicit in its organization and collection but without straight-jacketing the analyst into such strict categorizations of the data through the theory, that nothing but the theory becomes admissible. The process however requires some structure if it is not simply to spinoff onto all sorts of tangents which have become possible because of the sudden availability of cheap computer memory and the emergence of better graphics software.

A fourth issue relates to using visualization to give working scientists some of the power of visual thinking which is available to but a few. Most people have extremely poorly developed or at least poorly accessible visual imaginations and there seems no distinction in this between the public-at-large and more narrowly defined scientific peer groups. There are obvious applications such as visualizing phenomena which is often portrayed only in verbal or numerical terms in 2- and 3-d which scientists find hard to imagine, and quickly presenting data which traditionally has been slow to be visually examined. Moreover, the speed at which data can be recast into different visual representations enables some sense of the extent to which theory can be modified to accommodate to the data, a prospect which has only become available since large quantities of data are able to be rapidly displayed and visualized.

There is however a largely undefined research agenda in terms of the way scientists might best develop the process of visualization. We need to be concerned with the idea of 'optimal visualization', a process in which a scientist might make identifiable progress in examining data from many different perspectives while at the same time ensuring that promising avenues are exploited and scientific dead-ends avoided. I hope that the meeting gets to grips with such strategies for our own field because I think that if such issues are not explored, scientific visualization will just become another of those 'gee-whiz' fields spawned by the increasingly low cost of computer memory which adds little to the quest to do better science. We also need to explore parallels between this process of computer visualization, the use of visual imagination in more general terms particularly in the creation of hypotheses and the extent to which these visual clues can be made explicit in the process of hypothesis testing. We all have some idea of the power that visualization has to mystify and obfuscate, and the widespread use of these techniques are likely to throw up as many problems as opportunities.

Returning to the issue of data quality in its narrower sense, techniques of visualization can be used to identify the extent to which the patterns we expect to see in data are distorted by inaccuracies and errors due to the processes of measurement/ collection /organization. This is in contrast to those which might be associated with structural features of the data which show themselves as irregularities and are unlikely to be due to measurement. So far the way such problems have been identified is largely through a brute force approach based on visual exploration. By visualizing the data from every perspective in sight and by subjecting it to a wide battery of statistical tests and measures, it has been possible to identify areas of inaccuracy. But when it comes to a more considered approach to data analysis, there are few guidelines as yet on how to separate structural from more superficial problems in the data, the signal from the noise where both signal and noise confound one another and where the implicit theory explaining the data may also be weak. I would like to see the Initiative get to grips with some of these problems, to agree on promising approaches which will form a good basis for a best practice of visualization in geographical analysis, and the extent to which information systems should be informed and linked to techniques of visualization. For example, there are an increasing number of ideas for linking conventional spatial statistics to map patterns using 2-d and 3-d visualization and using the power of contemporary window systems to display more than one perspective on pattern in the data simultaneously. Questions of geometry and topology can now be linked to spatial regularities in data such as homogeneity and heterogeneity, boundary effects, multicollinearity, autocorrelation and such paraphernalia of spatial statistics developed over the last two decades. At present, there are several computer programs which develop these types of ideas and quite rapid progress is being made at this basic level. I hope the meeting will define the state of the art in this domain so that we can establish the most promising directions and the most urgent questions to be addressed.

However I feel that the most problematic issue involving visualization is the extent to which we can divorce data and data quality from the wider issues of theory and model which underpin any interpretation of data. I consider that it is essential to structure the process of visualization providing the analyst with a well-worked out chart of the way data should be explored. Such charts are perhaps the sorts of product which we might aim for in the NCGIA Initiative but if we agree on this, it would seem to me to be essential to broach the question of whether we can restrict visualization to simply data analysis. I think it is possible to do this in a limited sense but I consider the real potential of visualization to be in demonstrating how theory and models can be improved in terms of their data, and perhaps more important in the applied domain, how their use and operation can be communicated. To conclude this brief note on what I consider the key problems of this Initiative to be, I will discuss one of my own research projects and how computer graphics is being used to enable theory and models to be improved and continually matched against available data. The project involves the construction of a large-scale urban simulation model for the Buffalo-Niagara region. The model structure is based on several land use activity sectors such as population /demographic, industrial /commercial, educational, retail, recreational and so on. These sectors form part of a system of economic relations which are organized along lines similar to multi-regional inputoutput models through economic linkages. However more significant are spatial linkages simulated through interactions such as commodity flows, work and shopping trips and so on. The model is static in conception at least for the moment so spatial structure is the prime focus. The model theory is described in more detail elsewhere (Batty, 1983, 1986).

The various sectors of the model can be viewed as windows through which the workings of the model can be understood and displayed and there are well-worked out procedures for this in the literature. However, the way the user interacts with the model is through four related but distinct model-building processes: through data **exploration**, through model estimation or **calibration**, through using the model for **prediction** and by using the model for prescription or **optimization**. These four processes can also be viewed as windows through which to see the model operating. In the project so far, the idea of viewing the model through sectoral and through model-building windows are associated with the physical windows on the display device on which the model is run. So far we have developed a two-sector, 8-zone residential-employment model using data on the city of Melbourne. The model can be run in terms of data exploration, calibration, and prediction but not as yet optimization although the basic model theory exists. We are also considering other graphical windows through which we can view the model: for example through conventional map analysis functions which are to be found in GIS systems, and through spatial aggregation, both of which will involve the user going into greater depth with respect to the process, sector, aggregation and functional analysis in question.

In essence, in this project it is hard to see data analysis per se as the guiding light for visualization. Data exploration is just one element of model-building and the accuracy and quality of spatial data cannot be viewed in isolation. For example it is only when the model or its sectors are calibrated that the problems of data quality might become evident. Moreover, there exists the prospect of using the model predictively or prescriptively to change the data. The model can be operated in inverse fashion to get a sense of what the spatial system might appear as if certain outputs from the model are to occur. In short, the structure of the model and the various modelling processes condition the process of visualization as much as the implied functionality of spatial analysis which is also not simply the prerogative of the data exploration stage. Standard spatial analysis can be applied in any of the four modelling processes.

In summary then, I hope the meeting gets to grips with what we mean by visualization, the best ways of using it for a spectrum of purposes, not just exploring the quality of data analysis but the quality of theories or models associated with the data. I also feel that we need to classify various types of visualization appropriate to our own spatial problems and context which exist across a variety of levels. We need to define optimal processes of visualization. To generate such processes, we must tie visualization to explicit theories and models, exploiting their structure as a powerful means to effective visualization which produces fertile theoretical and applied insights.

References

M. Batty (1980) Limits to Prediction in Science and Design Science, Design Studies, 1, 153-159.

M. Batty (1983) Linear Urban Models, Papers of the Regional Science Association, 53,5-25.

M. Batty (1986) Technical Issues in Urban Model Development: A Review of Linear and Non-Linear Model Structures, in B. G. Hutchinson and M. Batty (Editors) Advances in Urban Systems Modelling, North Holland Publishing Company, Amsterdam, pp. 185-208.

Position Statement on Visualization of Data Quality

Kate Beard
National Center for Geographic Information and Analysis
University of Maine
Orono. ME 04469

Introduction

The historic meaning of quality is associated with fitness for use. This implies some direct interaction on the part of a particular user in the specification of quality or the institutionalization of a common definition or specification of quality by some group of users and/or producers. This position paper will briefly cover characteristics of spatial data which could be considered as contributing to fitness for use. There is both some general agreement about what constitutes spatial data's fitness for use as well as some individual specifications or requirements. Another important consideration for spatial data is its life span and the ways in which fitness for use can change over this life span. The life span and multiple use of data within a geographic information system will generally require that quality is assessed repeatedly and from different perspectives. Given this context, the overall objective is to determine where visualization methods might play a constructive role.

The Life Span of Spatial Data

The life span of spatial data covers collection, processing, storage, analysis, presentation, and maintenance functions. In some cases the entire cycle of quality activities over the life span of the data can be performed by the same individual. If a user identifies a problem, collects the data to solve that problem, writes algorithms to process the data, analyses and presents the results, that user has essentially complete control over data quality. Quality activities become more complex as data collection, processing, storage, maintenance, analysis and presentation functions become distributed over several groups. The way GIS activities are evolving, these functions are tending to become more and more dispersed. As end users become more removed from original data collection and processing, the need for quality information becomes more crucial. At each step in the process different groups may have different quality control concerns. For data producers, the quality of their data must be sufficient to continue their enterprise or justify their existence. GIS software vendors must be concerned about the validity of their algorithms for processing spatial data.

GIS users at the end of this sequence are often merging data of mixed quality heritage, perhaps using software of different quality heritage and then trying, in some cases, to say something about the quality of their results. End users of a GIS therefore potentially have the most difficult quality estimation problem, and unless information from the first steps has been tracked in some reasonable way, the quality assessment of final products can be rather meaningless.

Research over the course of this initiative should explore where and how in this sequence of events visualization techniques could be used most effectively to assess or evaluate quality.

Three primary areas of concern can be identified:

- assessing the quality of raw observations or measurements
- assessing the quality of data when processed for incorporation in a GIS
- assessing the quality of GIS products

Considerations include: does the data meet fitness for use requirements as it was originally collected (from a producers perspective? from an individual users perspective?), has the data's fitness for use changed with subsequent processing, has it changed significantly, and if so along what dimensions has it changed. At this point it would be useful to itemize the dimensions by which producers and/or users might evaluate spatial information's fitness for use.

Dimensions of Fitness for Use

As Sinton (1978) suggests all spatial data should be observed with respect to location, theme, and time. Using the three axes, we can itemize several characteristics of quality or fitness for use. Under location we might judge fitness for use on the basis of locational accuracy, spatial dimension, spatial resolution, consistency of spatial relationships, completeness of spatial coverage, and area of spatial coverage.

Locational dimensions of quality
- locational accuracy,
- spatial dimension,

- spatial resolution
- consistency of metric and topological relationships
- completeness of spatial coverage
- area of spatial coverage.

Locational accuracy is the correctness of a positional observation with respect to a true (or accepted as true) location. Spatial dimension could refer to the number of dimensions associated with a measurement, or the number of dimensions supported by a data model or by the system software. This could be one dimensional as in distances and angles, two dimensional as a coordinate location on a plane, or three dimensional - an X, Y, Z coordinate in 3D space. For geologists, geotechnical engineers, archaeologists, oceanographers, and others, the third dimension could be an important quality aspect of the data. Spatial resolution is a measure of the level of spatial detail. It is also dimension dependent. In one dimensional space it is the smallest linear distance which can be observed, resolved, or represented. In two dimensions it is the smallest area which can be observed, resolved or represented and similarly for the third dimension, the smallest volume which can be observed or represented. The consistency of spatial relationships refers to the establishment and preservation of expected or desired relationships among the data or as required by a data model. Completeness of spatial coverage refers to how exhaustively spatial information has been observed and represented for some area. Area of spatial coverage would indicate areas where positional measurements have been made.

Under theme, quality characteristics could include thematic accuracy, thematic resolution, thematic validity, and thematic consistency.

Thematic Dimensions of Quality
- thematic accuracy
- thematic resolution
- thematic validity
- thematic consistency

Thematic accuracy refers to how closely a measurement or value reflects a true value. Thematic resolution refers to the smallest thematic unit which can be observed. For continuous variables this would be the same as spatial resolution, i.e. the smallest thematic unit which can be observed or represented. For discrete measures resolution would refer to the number of classes or ranks, for example the number of land use classes. Thematic validity would be an indicator of how closely a surrogate measure approximates a desired thematic variable. For example electromagnetic reflectance is used as a surrogate measure for several thematic variables and not always appropriately. Thematic consistency would refer to the preservation of expected thematic relationships.

Fitness for use or quality characteristics associated with time include accuracy as well as currency, temporal resolution, and temporal consistency.

Temporal Dimensions of Quality
- temporal accuracy
- currency
- temporal resolution
- temporal consistency

Temporal accuracy refers to how closely a time measurement reflects true (or accepted as true) time. Currency refers to how recently observations were made. Currency will be relative to the dynamics of the phenomena being observed. Thus ten year old data can be current if the phenomena has not changed over that period. Temporal resolution refers to the smallest interval of time separating observations. Temporal resolution can be an important quality variable as an event may be missed if it falls between two observation times. Temporal consistency refers to the preservation of expected temporal relationships.

These characteristics correspond generally with quality components which have been described by others (Mead 1979, Chrisman 1984, Arnoff 1989). These characteristics have been previously described as: positional accuracy, attribute accuracy, logical consistency, completeness, resolution, time, and lineage.

Dynamics of Quality Descriptions

Assuming quality assessments are made on the above dimensions when data are collected, many of these assessments will change as the data are processed and subsequently used and analyzed. Documentation of data quality will be assisted by knowing which processes are applied to the data, which dimensions of quality are effected by which process and to what degree. This sounds relatively straight forward, but the potential problem is that changes in one dimension can create changes in other dimensions. For example, processes which change spatial resolution can change thematic accuracy, changes in thematic resolution can create changes in positional accuracy and changes in currency can change both thematic and positional accuracy. So over time and after several processing steps the quality equation becomes quite complex.

Visualization of Quality Components

Visualization provides an ability to organize abstract concepts into meaningful pictures. It is a tool for seeing the unseen and exploring complex relationships. It has been described as a manipulation of geometry, color and motion. Visualization has recently been popularized by an increase in the tools available to support it. Systems now support multiple windows to display different views of the same data or to compare different data sets. We can now combine graphics, text, and audio in many ways to present information. We can layer information more effectively in a single view, and we can animate displays. An additional benefit is the increased power and immediacy of user interaction. Fast interaction allows the direct visualization of the connection between an action and a result.

Visualization can be most appropriate where complex spatial patterns are involved, and where multiple variables and relationships must be grasped quickly. In this context, it has the potential to be a very effective tool for expressing spatial data quality. The down side of visualization, however, may be ambiguity. A recent study performed at Penn State showed that graphs allowed people to grasp trends among numbers more quickly, but less accurately.

Some dimensions of quality will lend themselves to visualization better than others. Some of the dimensions are simple enough that short statements or numeric codes could be sufficient descriptions. Visualization could become more appropriate as the level of complexity increases or the as the number of quality dimensions to be considered increases. Visualization may allow us to aggregate several dimensions of quality and view them simultaneously. Alternatively we could use visualization to view quality dimensions as separate but linked displays in which relationships between dimensions could be visually monitored as changes are made to each dimension. Following are some thoughts about how visualization might be applied at different stages.

Visual assessment of the quality of original observations

If end users are isolated from the data collection task they should be provided with information to assess the quality of data before proceeding with a particular analysis. This would be an exploratory phase which could borrow techniques from exploratory data analysis (Tukey 1977, Wills et al 1990, Monmonier 1990). SPIDER (Haslett et al 1990) is very interesting interactive exploratory data analysis software which allows multiple linked views of spatial data with statistical graphs. Methods similar to these could be used to browse a database of raw observations to assess their fitness for a particular analysis. Multiple linked views for example could allow users to select and simultaneously view data sets and the quality dimensions of interest.

Visual assessment of processing

As mentioned above several GIS processes change quality characteristics. In some cases it would be useful to see the effects of a process or even to view effects during the processing itself. Visualization might be used effectively in this context to isolate the effects of single processes. One example would be to visualize the change in thematic accuracy as spatial resolution is changed. This could be used as a training exercise for lay users, such that they become familiar with the possible effects of specific processes.

Visual assessment of the quality of GIS products

GIS products are frequently graphic displays with typically no description or reference to the quality or reliability of the results. The results can be the outcome of complex processing and combinations of multiple mixed quality data sets. If quality were more effectively controlled and documented in the previous steps it could be less problematic in presenting final results, but that is not yet the case. One question at this stage is whether visualization should
be (or could be) an aggregate expression of quality - a combined index of all the quality dimensions.

References:

Arnoff, S. 1989. Geographic Information Systems : A Management Perspective. Ottawa: WDL Publications.

Chrisman, N. 1983 "An Interim Proposed Standard for Data Set Quality" in Moellering, H. ed. Issues in Digital Cartographic Data Standards

Haslett, John, G. Wills, A. Unwin: 1990. ’SPIDER- An interactive Statistical Tool for the Analysis of Spatially Distributed Data. International journal of Geographic Information Systems 4:285-296.

Mead, D. 1979 ’Assessment of Data Quality in GIS’

Monmonier, M. 1990. Strategies for the Interactive Exploration of Geographic Correlation’, Proceedings , International Symposium on Spatial Data Handling 1: 512-521.

Sinton, D. 1978, ’The Inherent Structure of Information as a Constraint to Analysis’. Harvard Papers on GIS.

Tukey, J. 1977. Exploratory Data Analysis Reading: Addison-Wesley.

Wills, G, R, Bradley, J. Haslett, and A. Unwin. 1990. ’Statistical Exploration of Spatial Data’ "Proceedings, Fourth International Symposium on Spatial Data. 1:491-500.

VISUALIZING CARTOGRAPHIC METADATA

Barbara P. Buttenfield
National Center for Geographic Information and Analysis
Department of Geography, SUNY - Buffalo
Buffalo, NY 14261
email: GEOBABS@UBVMS.bitnet

ABSTRACT

Data quality is important for effective use of GIS map displays and overlays, and impacts upon reliability and credibility of data representation, decision-making and model-building. The Proposed Standard for Digital Cartographic Data includes specifications for data quality reports, including components of lineage, positional and attribute accuracy, completeness, and logical consistency. The Proposed Standard states that when spatial variation in quality occurs, thematic overlays may be constructed as diagrams or thematic map depictions. No guidelines are provided for symbolization, level of generalization, or other graphic design criteria for these quality reports. This position paper does not dwell directly on the modeling of error, but on its appropriate display. The conceptual framework in this position statement has been refined after its first presentation at the San Antonio GIS/LIS meetings (Buttenfield and Weibel, 1989) and builds on recommendations of the Proposed Standard as well as from conventions of cartographic design. The framework acknowledges distinctions between the five components of quality, incorporating various data formats, types of geographic phenomena, and specific visual variables to present a scaffold for design of visual tools for spatial metadata. Throughout the paper, the terms 'metadata' and 'data quality' will be used interchangeably, as a convenience.

INTRODUCTION

With development of GIS and related technologies, we are able to display and analyze large volumes of data rapidly. Our interpretive capabilities remain to a certain extent rooted in our ability to visualize accurately, and to discern the quality of the patterns that we identify. Information on data quality is important for effective use of GIS data. It impacts the credibility of the representation and the confidence that we attach to our interpretations. It impacts the reliability of interpretations and thus decision-making based on GIS modeling, sensitivity analysis, and data exploration.

The spatial variation in data quality across a map surface is not clearly understood, although efforts to formalize error assessment have increased in recent years (Unwin, 1981, Wrigley 1975, Goodchild and Gopal, 1989). Statistical information has also been studied to determine reliable graphic techniques for creating data displays that do not mislead the reader. William Playfair's efforts provide one of the best known examples of statistical charting efforts (but see also Raisz, 1948, Schmid and Schmid, 1979, Cleveland, 1985). More recent attention has been provided by Tufte (1983; 1990). The development of Exploratory Data Analysis (Tukey, 1977) led to creation of innovative graphical tools, embedding graphical depictions of standard error and confidence intervals with data displays.

Until recently, no consensus existed on how to explore data quality. Recent publication of the Proposed Standard for Digital Cartographic Data (NCDCDS, 1988) included a section categorizing five types of data quality. The Proposed Standard states that when spatial variation in quality occurs, thematic overlays may be constructed as diagrams or thematic map depictions. No guidelines are provided for symbolization, level of generalization, or other graphic design criteria for these quality reports. These issues are introduced in the position paper, and a framework is proposed for selecting graphical solutions. The paper will not dwell on the modeling of data quality, but on its appropriate display.

ACCURACY, ERROR, AND QUALITY

To fully understand the development of visualization of data quality, it is important to distinguish 'quality' by formal definition. In the context of spatial data, several terms are commonly surrogated. for quality. For purposes of discussion, consider the definitions as cited in Webster's Encyclopedic Dictionary (1971):

Accuracy (from the Latin *ad* , to, and *cura* , to care). Extreme precision or exactness; exact conformity to truth, or to a rule or model; correctness.

Error (from the Latin *errare* , to wander). Deviation from accuracy or correctness; (in mathematics) the difference between the observed or approximately determined value and the true value.

Precision (from the Latin *prae* , before, and *caedo* , to cut off). The sharpness of definition or limit; exactness; characterized by extremely exacting measurement.

The interaction of these concepts is obvious. Data accuracy may be viewed in terms of 'conformity to a rule or model'. Geodetic models are applied in surveying and in map projection; data models are used for both generalization and for statistical analysis; and in light of these and other examples that might be cited, data accuracy relates an observation to a model by similarity. Error, on the other hand relates an observed value to the true value, by discrepancy, and differs from accuracy but is not its opposite. Precision refers to measurement and not to the thing being measured; it is often considered in terms of the number of digits available for recording the measurement. According to these definitions, then, one can cite a level of accuracy or of error with varying precision, and the level of precision implies something different in each case.

For spatial data, particularly continuous spatial data, such as terrain, bathymetry, temperature variation, population density, or fiscal transfers, values are sampled, or interpolated from sampled values. The true data surface is rarely determinable, and cartographic representations of the statistical surface are rendered with the implicit assumption that the representation is at best a sound approximation (to a spatial statistician, a representative sample). Thus for much of the data archived in a GIS, discrepancy from 'true values' cannot be computed since there is no 'true value' archive.

In contrast, the concept of data quality is formalized without reference to truth, precision, or to data type (eg., continuous data). Continuing to cite from Webster's Encyclopedic Dictionary (1971):

> **Quality** (from the Latin *qualis*, or such). That which makes or helps to make anything such as it is; a distinguishing property, characteristic, or attribute; (in logic) the negative or affirmative character of a proposition.

Data quality can be defined in terms of either an affirmative or negative character. The connotation of quality implies that data posesses attributes of both accuracy (an affirmative attribute, measured in terms of similarity) and error (a negative attribute, measured in terms of discrepancy). Accuracy is the more readily quantified, by comparing measures with a model (geodetic, statistical, cartometric, or thematic) constructed for a specific purpose. Cartographic representations can be rendered to serve these purposes, as will be demonstrated.

COMPONENTS OF SPATIAL DATA QUALITY

The Proposed Standard for Digital Cartographic Data (NCDCDS, 1988) includes specifications for five measures of data quality, including positional accuracy, attribute accuracy, logical consistency, completeness, and lineage. Note that two of the five are explicit expressions of accuracy; error is not explicitly incorporated, and none of the measures are constrained by precision. In the next sections, types of data quality will be related to different types of cartographic expression.

Positional Accuracy. For the most part, cartographic perspectives on data quality have been limited to a concern with positional accuracy. Positional accuracy has accommodated many definitions over time, but the common thread has related map distance and direction to earth location in either relative or absolute measure.

Geodetic mapping accuracy is probibilistic in its cartographic quality requirement. The definition of horizontal accuracy in the National Map Accuracy Standard (Thompson, 1975, p.--) states that "90% of all well defined features be located within 1/50 inch at map scale of their true position". No stipulation is made that the positional accuracy lie within in a particular spatial distribution across themap, nor what positional deviation the remaining feature positions have, nor even what is meant by 'well-defined features'. Nonetheless, American federal mapping agencies have used this definition to establish quality thresholds for horizontal position on topographic map series for many decades.

Attribute Accuracy. Attribute accuracy concerns the thematic labels applied to positional data. Labels may refer to items hierarchically. For example in defining an underwater hazard, an attribute code may assign the hazardous attribute with additional specification as a natural feature and further define its type (eg., a submerged reef). Statistical attribute accuracy may involve indications of dispersion, or of certainty. For example, magnitude of a confidence interval is readily included with statistical graphs, and Tukey's notched box plots (Tukey, 1977) show levels of confidence in a tiled data display. Climatological maps may embed monthly precipitation graphs to display moisture variability implicitly.

Logical Consistency. In defining logical consistency, the NCDCDS Proposed Standard adopts a connotation of accuracy often applied in surveying, that an accurate measurement will yield similar values when repeated many times. The Proposed Standard defines logical consistency as "the fidelity of relationships described by the data structure". For example, the topological model (Corbett, 1979, White, 1983) used for US Geological Survey's DLG format, for the Census Bureaus' TIGER system, and for Defense Mapping Agency's Standard Linear format (Langran, 1989) sets constraints on definition and cycling of chains and nodes around polygons. In order to preserve logical consistency to this model, a data file must adhere to these constraints. Other data models are applied in various cartographic situations, including relational structures, object orientation, and run-length encoding. In each case, logical consistency is determined in conformity to the particular model after which the data is structured. Problems with logical consistency may be encountered when queries are made to a multi-scale database, for example (Bruegger and Frank, 198-). This type of hybridized data source issue is not covered by the Proposed Standard, but forms a real and pressing issue for the treatment of data quality in general, and cartographic representations in particular.

Completeness. Completeness implies that mapping rules are applied in equal fashion to all data. The model may be developed for selection criteria or geometric thresholds in simplification, for categorical or metric data classifications, for application of weights in a discrete suitability model, and for other GIS operations. The Proposed Standard also refers to completeness as exhaustiveness. Completeness also indicates whether missing values have been encountered, and identifies gaps in the data progression. .

Lineage. The lineage of a data base includes reference to source materials, data collection, and preprocessing including geometric transformations applied to the data. Whereas the first three types of data quality are archived as attributes of stored data, and the fourth type (completeness) is often stored implicitly, lineage archives require access to database pointers. Langran (1989, p. 20) connotes an almost purely temporal model to which lineage must adhere. "All known dates that apply to source, capture and update also comprise lineage. The dates when information was discovered in the physical world are preferred; if lacking, however, lineage dates can describe source publication dates if declared as such." However, lineage in its most comprehensive form requires access to both temporal and spatial information. Buttenfield and DeLotto (1989) have argued that queries as to the scale of a data base involve access to lineage, because most large data bases archive information from many sources compiled at multiple scales. The reliability diagrams included in some US Geological Survey topographic series might also be encompassed by lineage, as they provide graphic depiction of the overlap of photorevision and ground survey source data, along with specific dates.

SPATIAL DATA TYPES

The configuring of data types is specific to the domain of interest, as the operations in a domain will be designed to modify some but not all types of data. For example, mathematical operations operate directly on computational data types including scalars, vectors, arrays, and matrices, etc.. For the most part, it is simpler to think about mathematical data in these forms, although the configuration must be expandable to accommodate special situations, (to continue the example) as in operating on surfaces and manifolds. Other data typing configurations have been applied in cartography, including those based on the nominal ’ordinal, interval and ratio scales of measurement, and those based in linguistic categorizations (Child, 1984).

The traditional configuration of cartographic data types include points, lines, areas, and (sometimes) volumetric data (Robinson and Sale, 1968). The configuration was originally used for describing paper map production and some aspects of thematic symbolization. The model has constrained recent cartographic paradigms, however, as it limits operations to the realm of illustration, and describes geometry to the exclusion of topology. For example, a TIGER file description includes not only the geometry of an item (a street, or a block face, for example) but also its topology, and the limited data types available in the 'point-linearea" configuration cannot describe both aspects of TIGER information.

When dealing with spatial data quality, it seems most direct to apply accuracy measures to discrete, categorical and continuous data types (Wrigley, 1975), for several reasons. First, the assessment of accuracy involves statistical operations (computation of least squares deviations, maximum likelihood, etc.). Second, statistical data types allow easy incorporation of complex digital objects, including hierarchical representations. Finally, statistical data types do not constrain the assessement of data quality to a single map product produced at a single map scale, and this means that GIS map composites and model output may also be encoded within this configuration.

Definition of the three data types in the context of cartographic data quality will help clarify the framework. Discrete data is characterized with the property of finite boundaries, and includes (but is not limited to) cartographic point and line objects. Notice that discrete data may have an area] extent, for example, Buffalo may be represented by a point symbol at one scale, as an areal patch at another scale. Continuous data exists everywhere on the map, defining a surface, and their representation is often generated by some form of interpolation from sampled points. Terrain elevation and volume of migration flows provide geographic and statistical examples of continuous data.

Categorical data is contained within enumeration unit boundaries, and may take two forms. Nonmetric categories such as landuse, vegetation or soil type are enumerated in coverages that are data-determined, that is, the boundaries of the enumeration units are determined by the data. Overlay operations are commonly applied to nonmetic categorical data. Metric categories are enumerated in predetermined areal units that often have little to do with the data itself. For example, population density, crop yield per acre, and mortality rates per thousand population are each tabulated by census zone, county, or other politicallydetermined units. Metric categorical data must be standardized to its enumeration unit size for some types of cartographic depiction (eg. choropleth maps), to compensate for the discrepancy in resolution between data source and data representation. GIS operations applied to define class breaks for metric categorical data include partition and aggregation; and nonclassed categorical data displays are increasingly generated by means of overlay or buffering.

The expression of data quality varies for each data type. For discrete data, positional accuracy is readily determined at the boundaries of a feature. The boundaries of some categorical coverages (for example, soils parcel data) are not so readily determined, and in some mapping situations positional accuracy may not be meaningfully determined. Logical consistency is assessed in conjunction with the data structure model, and clearly this aspect of data quality will vary accordingly. The logic of TIN models will require that links do not connect nodes of the same elevation for example, but the logic of contour models will restrict all linkages to equal-valued points. And as the assessment of data quality varies for different types of data, it can also be shown that graphic representations of data quality vary in appropriate use for both data type and data quality measure.

A FRAMEWORK FOR VIEWING CARTOGRAPHIC METADATA

The Proposed Standard states "When spatial variation in quality occurs, a quality report must record that variation" and later on the same page "... the quality overlays appear as diagrams with text labels or thematic map depictions." (NCDCDS, 1988, p.131) No guidelines are indicated for symbology or level of detail. No specification is provided for particular indices of data quality that may be appropriate to a given quality type or data type. A large body of cartographic research has demonstrated that design strategies affect reader comprehension of data patterns (Dent, 1990; Buttenfield and Mackaness, 1991; MacEachren et al, 1992). Clearly, the design for a quality display may bias reader comprehension of spatial variations in data quality, which in turn may affect subsequent modeling or decision-making. Furthermore, one should anticipate that particular symbology will be appropriate to represent particular combinations of data type and data quality type. Figure 1 crosstabulates these combinations, and cells of the matrix contain design strategies for mapping all possible data quality configurations. These strategies are not comprehensive, but present generic formats with specific examples from the literature where appropriate.

The design strategies compose a graphical and lexical syntax that is proposed as a mechanism for visualizing the quality of cartographic data. Graphical syntax' refers to those designs that are iconic, including one or more of the visual variable system first proposed by Bertin (1983). The visual variables include size, shape, value, orientation, visual texture, color (hue and saturation), and two-dimensional position. Graphical-Lexical syntax is composed of displays that include verbal information, including marginalia, legend text, and other lexical information, such as pointers within a data structure. Three types of matrix cells may be found in the figure. First are those design strategies accepted by cartographic convention, and most of which are tested by use over long periods of time. Second are those cells for which no convention is accepted at present, but for which visually logical design strategies may be proposed. Third are cells in the matrix for which design solutions are problematic, either for graphical limitations or for conceptual constraints (for example, situations where error models have not yet been well-developed, as in soils mapping).

Figure 1
A Framework for Visualizing Cartographic Metadata

DATA QUALITY / DATA TYPE	POSITIONAL ACCURACY	ATTRIBUTE ACCURACY	LOGICAL CONSISTENCY	COMPLETENES	LINEAGE
DISCRETE points and lines	*size* *shape* (error ellipses) (epsilon bands)	*value* *color saturation* (feature code checks)	*value* redundancy by overprintin slivers by solid fi *shape* (topological cleaning)	**Mapping techniqu** density traces **Marginali:** generalization algorithm snapping tolerance buffer size	
CATEGORICAL Aggregation and Overlay (tesselation, tiling, areal coverages)	*texture* *value* (certainty of boundary location)	*color mixing* (attribute code checks) (topographic classifier)	lack error models	**Mapping techniqu** missing values logial adjacency surface **Marginali:** discrete model weights	**Mapping technic** Minimum Bounding Rectangles (reliabilty diagrams
Partitioning and Enumeration (metric class breaks)	not meaningful	*size = height* *value* (blanket of error)	*size = height* (maximum likelihood prism maps)	**Mapping techniqu** missing values misclassification matrix classing scheme OALTAI	**Marginali:** source of data scale / resolution date geometry
CONTINUOUS Interpolation (surfaces and volumes)	no clear distinction between the two *value* *color saturation* (continuous tone vignettes) (continuous tone isopleths)		*size = line weight* *color* *shape = compactness* (TIN links)	not possible by definition **Mapping techniqu** surface of search attenuation **Marginalia** interpolation algorithm	

← GRAPHICAL SYNTAX →

← GRAPHICAL / LEXICAL SYNTAX →

SUMMARY AND CALL FOR RESEARCH

Many research questions come to mind in viewing the proposed syntax for visualizing data quality. First, one might address the matrix cells containing graphical tools that are well-tested, and ask representational questions. For example, should data quality be displayed simultaneously with the data? Or should metatdata displays be in separate windows, or even toggled views displayed in a single window? Second, for cells containing tools that are not well-tested, should data quality be displayed in the same symbology as data? That is, must chorpleth error be displayed choroplethically? This begs questions about the nature of error - discrete data/continuous error-surface, and forces confrontation with the concept of the map as statistical sample. For the third type of matrix cell, research questions can be asked to delineate user assumptions about error, in general. What do our visual displays imply to the GIS user? For example, an interpolated fishnet surface implies terrain of an either geographical or statistical domain. Does the display of metadata in this format imply a particular value of spatial autocorrelation, as we assume for terrain? Other questions must wait for discussion at the Specialist Meeting, but perhaps these ideas give a perspective to issues of interest to at least one participant.

REFERENCES

Bertin, J. (1983) Semiology of Graphics. Madison: University of Wisconsin Press.

Buttenfield, B.P. and DeLotto, J.S. (1989) Specialist Meeting for the Initaitive on Multiple Representations. NCGIA Technical Report 89-3, Santa Barbara California.

Buttenfield, B.P. and Weibel, R. (1989) Visualizing the Quality of Cartographic Data. Paper presented GIS/LIS '89 meetings, San Antonio, Texas, November, 1989.

Buttenfield, B.P. and Mackaness, W.A. (1991) Visualisation. Chapter II.a.4 in GIS: Principles and Applications. (Eds. MacGuire, Goodchild, and Rhind) London: Longman. (in press)

Child, J.C. (1984) Cartographic Structure and Function: A New Perspective on Map Language. Unpublished PhD thesis, Dept. Geography, Univeristy of Washing ton-Sea ttle.

Cleveland, W,S. (1985) The Elements of Graphing Data. Monterey, CA: Wadsworth Books.

Corbett, J. (1979) Topological Principles in Cartography. US Census, Washington, DC.

Dent, B.D. (1990) Cartography: Thematic Map Design. 2nd Ed. Dubuque, Iowa: Wm. C. Brown.

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

Langran, G. (1989) Time in Geographic Information Systems. Unpublished PhD Dissertation, Department of Geography, University of Washington, Seattle.

MacEachren, A. E., Buttenfield, B. P., Campbell, J. C. and Monmonier, M. S. (1992) Visualisation. In: Abler, R. A., Olson, J. M. and Marcus, N. G. Geography's Inner World. New Brunswick, NJ: Rutgers University Press (in press).

NCDCDS (National Committee for Digital Cartographic Data Standards) (1988) Proposed Standard for Digital Cartographic Data. The American Cartographer, vol. 15 (1).

Schmid, C.F. and Schmid, S.E. (1979) Handbook of Graphic Presentation. New York: McGraw-Hill (2nd edition)

Raisz, E.J. (1948) General Cartography. New York: McGraw-FEII.

Robinson, A.H. and Sale, R. D. (1968) Elements of Cartography. New York: John Wiley.

Thompson, M. (1975) Maps for America. US Geological Survey.

Tufte, E.R_ (1983) The Visual Display of Quantitative Information. Chesire, CT: Graphics Press.

Tufte, E.R (1990) Visualizing Information. Chesire, Connecticut: Graphics Press.

Tukey, J.W. (1977) Exploratory Data Analysis. Reading, MA: Addison-Wesley.

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

White, M. (1983) Tribulations of Automated Cartography, and How Mathematics Helps. Proceedings AUTO-CARTO 6, vol. 1, Ottawa, Canada, p. 408-418.

Wrigley, N (1975) Categorical Data Analysis. London: Methuen.

Position paper for Visualization of Data Quality Initiative

Nick Chrisman
Department of Geography
University of Washington

At AUTO-CARTO 10, I joined a panel organized to discuss this topic. I talked and debated. Whatever I said then, I did not write down. I will reconstruct what I now think are key points.

I am not primarily concerned at the moment with the visualization component of this initiative. My concern is the nature of data quality. I do not think that Veregin's "Taxonomy" of error is sufficient or useful as a result of I1. 1 do support the general structure of the Data Quality Standard now imbedded inside SDTS. Certain colleagues do not find it sufficiently inclusive, but I tried to demonstrate that the standard does include the various components required.

Thus, research on visualization should follow some general structure for the information. A taxonomy of error can be developed from Sinton's framework for the components of geographic information: space, time and attribute are one fixed, one controlled, and one measured. Error in a measurement is easiest to handle and model and portray. Error in control can lead to wider problems. Error in the fixed element will appear to be error in one or more of the other elements. This framework has been elaborated in my chapter to the Maguire, Goodchild and Rhind book, but nobody will have a copy because it is so expensive.

Data quality contains lineage and testable components of accuracy and consistency. Lineage, the time dimension of a developing GIS, may offer special interest for visualization. Showing the time dimension may provide an important connection between the processes which generate information(permits, surveys, maps) and the user's awareness. Cartographers used to hide such details in the illusion of an authorative snapshot. Some areas ARE less well known than others. Some are changing more rapidly and old data is less useful. On top of lineage, data quality involves testing, not just making assumptions. Tests involve reproducing the information from some other source (as independent as possible). Such tests will take time to develop as a technique. I disagree with some researchers who question all categorizations and assert that the world is infinitely complicated. Maps are deliberate simplifications, purposeful abstractions, and must be tested within that expectation.

On a broader topic, I find data quality a topic that will not go away. I visualize data quality as the subtext to our data, the unspoken story which contains the interesting stuff. Error analysis need not be approached entirely as a clinically positivist framework. Some components can be tested against a model of the "real" world, but much of this world is socially constructed. Error has to be understood inside a framework of cultural and social and institutional values which provide the source of meaning.

Position paper prepared for the NCGIA Specialist Meeting on "Visualizing the Quality of Spatial Data", Castine, Maine, June 1991.

Visions of Quality: Visualizing product quality in GIS

by
Helen Couclelis
University of California
Santa Barbara

The subtle contradiction in the title of this workshop, *Visualizing the Quality of Spatial Data*, provides a welcome, opening to my remarks. Clearly, the choice of visualization as a theme indicates a focus on the user end of things; yet, the GIS user intent on getting a good product for his or her purposes is not interested in data quality per se any more than the taster of a chocolate cake is interested in the taste of the flour or the baking soda that went into it. My point is simply the following: *data* quality is not directly an issue for the user of a GIS application; *product* quality is. I use the term 'product' here with some reluctance, for want of a better word, aware that the static finality it suggests is at odds with some of what is best about GIS technology: the potential for involved, dynamic, flowing interaction between user and system. I thus see the topic of this workshop breaking up into two separate questions:

1. The relation between data quality and product quality;
2. the question of visualizing product quality (at least that part of product quality affected by the quality of the data).

The issue of data quality is difficult enough in itself. It is obviously-associated with that of data error, although the two are not synonymous: whereas data error can be discussed in absolute terms, data quality is a function of purpose. Tell me what you intend to do with them and I will tell you how good your data are. Gross errors in the third dimension are irrelevant if all you want is a two-dimensional map. Bad data may look good if your other data are even worse. Data that are unacceptable at some scale may be just what is needed at a coarser level of resolution, and longitudinal data too noisy for year-by-year analysis may be fine for longer intervals. Slivers and gaps between what should have been adjacent polygons could have homicidal consequences on a cadaster map, but may be welcome on a planner's sketch where hard boundaries are barriers to the imagination. And one of the best known, most widely used, and most useful maps in history, the London subway map, reflects spatial data that would be terrible for most purposes, though obviously right for that particular one. These facts are well known. It is also well known that things can get even worse when you consider what happens to source data that must undergo a complex sequence of manipulations. Years of work on error propagation in models, culminating in the flurry of spectacular graphical outputs from models of chaotic systems, have yielded sobering insights: while in some cases errors may cancel out, in other cases they are so quickly amplified through non-linear dynamics that even the best of data can produce garbage after a while. Putting a spatial data base through a GIS system is also like running a complex spatial model of sorts, albeit one that is very largely unexamined and ad hoc in terms of its underlying structure. The vast literature on data error, and on error in spatial data bases in particular (as reviewed by Veregin, 1989), obviously has a direct bearing on the issue of product quality in GIS: still, by itself, it does not even get close to answering the question.

The difficulty of defining product quality for GIS has much to do with its hybrid nature as part electronic cartography, part technological artifact, and part expression of new spatial knowledge. The most characteristic product of the GIS application, the map-like display or hard copy, is on the surface similar to the product of cartographic design. Yet the question of product quality for maps does not have anywhere near the same urgency. The cartographic map is the product of a design process which, while still evolving, has been codified and refined over many centuries. Part science and part traditional art, making a good map is like making good wine, produced by a few experts for the benefit of all. The GIS product by contrast is more like what comes out of a kit for do-it-yourself chemistry experiments: it can be tailored to one's desires, it is endlessly varied, often surprising, frequently hard to understand, sometimes insiduously lethal, and the (amateur) maker and (naive) user are often one and the same. Safeguards and warnings are built into the kit to help avoid the most likely mistakes and disasters, but as for every open-ended process, there can be no guarantee as to the quality or safety of the final product, and accidents do happen.

Qua artifact produced by advanced technology, the GIS product is controlled by a large number of safeguards built into the system to prevent manipulations that are illegal in some well defined sense, safeguards deriving from requirements of logic, software design, or the underlying geographic data model. This ensures up to a point the technical adequacy of the product, but-does little for those aspects of product quality relating to its function and purpose. In his classic monograph on the theory of design, Herbert Simon (1969) distinguishes between the *internal* and *external* environments of a design product. The internal environment is made up of the physical and logical constraints that a design must meet to be technically viable, while the external environment consists of the set of requirements that it must satisfy in order to be able to fulfill successfully its intended purpose. A good design reflects both what the artifact is made of, and what it is made for. Because the products of GIS are members of an open set of possible designs and purposes, the quality controls relating to the external environment cannot be built in. This clearly includes the aspects of product quality relating to the quality of the data.

Another source of difficulty in trying to define product quality for-GIS derives from its third aspect as a means for producing new spatial knowledge. If the cake has too much' salt, or if the chemistry experiment blows up in your face, you know it at once; if the model airplane you just built cannot fly, you soon find out. But a bad GIS product does not necessarily carry with it the signs of its own inadequacy any more than a bad theory of education does. Bad knowledge is not intrinsically different from good. Information technology differs from ordinary, material technology in that the most critical aspects of the quality of its products are only testable through their indirect consequences.

A key aspect of GIS as information technology is its interactive nature. In a GIS application, new knowledge about some aspect of the geographic world is produced cooperatively between a human of varying ability and sophistication, and an electronic partner who is incredibly smart in some respects, and unfathomably stupid in others. Although there is no direct analog of this kind of partnership in the realm of human interactions, it is a case of interactive knowledge production involving two parties with complementary forms of expertise.

A persistent theme in the literature on human-computer interface design concerns the utility of models of humanto-human communication. I agree with those weary of anthropomorphizing the computer too much, and have little patience for over-friendly systems that call me "Helen". At the same time, I believe in the usefulness of the conversational metaphor for GIS, and see init a clue to some of the answers that this workshop is seeking.

Viewing the GIS application as an instance of cooperative production of applied spatial knowledge makes questions about the quality of the GIS product analogous to questions about the quality of knowledge produced in conversation. A prerequisite for meaningful conversation to take place among humans is that the partners share a *schema* of the situation at hand, that is, a cognitive model of the issue discussed along with its background of assumptions, tacit presuppositions, appropriate dispositions for action, and other relevant forms of understanding. Failure to share such a background in human interactions results in the partners talking past each other and using the same words but meaning different things, giving rise to *quid pro quos* and other predicaments celebrated in farce and drama. The schemas that GIS systems embody are highly specialized, focussed and controlled (though not necessarily by their users), but can still differ considerably from those of their human partners, or even from those of other systems sitting in the same room. Dutton (1984) argues that the question of data quality in GIS often relates to the "spatial ontology" embodied in the system. This is usually understood to mean a field or object view of geographic space, which boils down to the technical raster-vector distinction. In actual fact, there are several more ontologies of space reflected in geographic theory and applications. In a recent paper (Couclelis, in press) I distinguished between geometrical, physical, socioeconomic, behavioral and experiential space, few of which can be accommodated comfortably within either the field or the object perspective. We can expect that the quality of a GIS product (as well as the quality of the data used in it) will be to some extent a function of the degree of agreement or discrepancy between the model of space embodied in the system and that presupposed by the application itself.

Even assuming compatible spatial ontologies, the interactive production of spatial knowledge using a GIS is subject to the same difficulties as the production of any kind of applied knowledge through conversation. Well-meaning conversation partners can arrive at bad conclusions for many different kinds of reasons, having to do with the quality of their beliefs relating to the area under discussion, the quality of the inference processes used, or the quality of the conversational interaction itself. In close analogy, the GIS product may suffer from insufficient (for the particular application) quality in the source data, from problematic data manipulations due either to the operation of the system itself or to a poor model of the spatial phenomena analyzed, and from a misleading, inefficient, or inefficiently used interface.

Human knowledge structures are characterized by several kinds of imperfections, of which those discussed in the quantitative literature on error are only a small part. Thrift (1985) discusses five forms of "unknowing": that which is unknown or unsuspected; that which is not understood; that which is undiscussed or undisputed; that which is deliberately hidden; and that which is distorted. All of these can occur in the kinds of spatial knowledge represented and produced through GIS, although only the last kind, distorted knowledge, is much discussed in the literature on error.

Processes of inference in humans can go wrong because of faulty deductions, because of hasty inductions, because of unwarranted generalizations, because of improper handling of dubious beliefs, because of the use of inappropriate analogies or heuristics, because of loss of focus on the problem at hand, because of having been led up the garden path, because of the use of bad examples, because of drawing the wrong lessons from experience, and so on. Quantitative work on uncertainty and reasoning, and in particular formal logic, has modeled some of these inference processes in ways that can be directly transferred to computer-based systems, although many would argue that such models are more appropriate for computer than for human reasoning anyway.

In conversational interaction, beliefs and inferences are expressed in the kinds of *speech acts* studied by Searle (1969) and others. These are summarized by Winograd and Flores (1987, p.58) as follows:

Assertives commit the speaker to something being the case; *directives*, including questions, requests and commands, attempt to get the hearer to do something; *commissives* commit the speaker to a future course of action; *expressives* express psychological states, such as relief or apology; and *declarations* bring about the correspondence between the propositional content of the speech and reality, as when pronouncing a couple married.

A key point in speech act theory is that conversation creates patterns of commitment between the partners, in the sense that each of them must assume that the other is basically sincere in his or her manifest intentions, and uses speech acts appropriately given the context of the interaction. A similar notion is stated in Grice's (1975) 'cooperative principle' in human communication, and its maxims of quantity, quality, manner, and relation (see Note). Failure to meet any of these in conversation may result to poor quality outcomes, because the partners have talked past each other, because they have used terms in different ways, because they have used the wrong body language, have conveyed the wrong signals to each other, because they have withheld critical information, because they have not been as clear and to the point as they might have been, because they have tried to conceal their ignorance, please, flatter, manipulate or mislead their partner, or because they had incompatible intentions from the start.

By and large though human conversations work because the cooperative principle is followed. This involves, among other things, giving your partner information on an ongoing basis about the degree of validity of your speech acts.

Humans have a variety of means at their disposal for qualifying what they say and for marking uncertainty and error in speech or body language. Precision or vagueness in assertions is expressed by means of adverbs such as exactly, precisely, vaguely, approximately, relatively, roughly, basically, and so on. Phrases such as I am sure, I am positive, think, I believe, it seems to me, on the other hand, let me think, I take this back, are hedges intended to convey to your partner the force of your own confidence in your assertions, while other phrases (is that right? are you sure? how come? come on! ...) invite your partner to give more information about the quality of his or her speech acts. Pauses, hesitations, frowning, head scratching, shrugging, looking away, some kinds of smiling or gesturing, complete this vocabulary, along with a variety of subliminal vocal signs of the kind picked up by lie detectors.

The analogy between conversation and human-computer interaction can never be perfect, because human conversation is governed by the intentions of the partners, something that computers and all other non-conscious things cannot have (Dreyfus, 1979). The day is saved by Dennett's (1978) notion of an *intentional system*, defined as any system that behaves *as if* it had beliefs, desires and intentions. Many systems designed for specific purposes meet that definition, and a computer programmed to help produce some piece of new knowledge is perhaps the best example of one. The computer behaves *as if* it were a friendly, docile, knowledgeable, sincere, benign, but not overly insightful or creative partner, whose intention is solely to help you solve whatever problem you may have. That thorny philosophical problem swept under the rug, most key aspects of conversational interaction have direct analogs in the interaction with a GIS. Beliefs and inferences in humans map rather conveniently into data, data structures and processes in computer systems. Speech acts have their analogs in the graphical displays, icons, menus, windows, messages, commands, warnings, queries and other user inputs supported by the user-system interface. The requirement for compatible background schemas can in principle be met, on the system side, by a database structure embodying a model of geographic space appropriate for the application at hand. The intentions of the user can be matched, on the system side, by the intentionality simulated by the GIS's purposeful, application-directed problem-solving behavior. The only thing missing to make the analogy complete is a counterpart to the human compulsion to provide abundant direct and indirect clues as to the quality of what they say in conversation (see Figure 1).

It is only appropriate that the shortcomings of applied spatial knowledge of the kind best expressed in map form, and derived through interaction with a user interface that is primarily graphic, should also be expressed visually.

The analogy with knowledge production through conversation provides a meaningful typology for the kinds of problems that need to be made explicit, if possible in visual format: problems of imperfectly shared or partially incompatible background schemas (spatial ontologies); problems of inadequate beliefs (including erroneous and missing source data, and inadequately represented spatial relations); problems of faulty inference (inappropriate data transformations due to system operation or implicit model); and problems due to flawed speech acts (unsatisfactory usersystem interaction), resulting in misleading, insufficient, or largely irrelevant results.

Here is, then, how I view a possible agenda for this initiative: to elaborate a coherent system of signs that can help the user visualize the quality of GIS products at any stage of deriving an application. That visual semiology should be analogous to the rich system of qualifiers of speech acts used in human conversation. It should be a visual language for indicating hesitation, doubt, ignorance, uncertainty and error which, as in human conversation, is a function of background model, purpose and context, rather than some extrinsic criterion of "quality of spatial data".

To be done properly, the task should be seen as more than just a mapping of one category of signs into another. It is more a question of deriving a complete visual grammar, syntax and semantics for qualifying the GIS product, not only in its final form but in every step of its creation. Insights should be drawn together from the psychology of perception, cognitive linguistics, conversation analysis, and semiotics, as well as from the more immediately obvious areas of spatial data error analysis and modeling, computer graphics and animation, data base design, cartographic theory and design, and the machine representation of spatial categories and objects. The task seems daunting at first sight, but many of these issues are already being tackled under other NCGIA Initiatives. Since it brings so much together, 17 may soon be recognized as the Mother of All Initiatives. All this might take is a slightly modified title.

Note

Grice's "cooperative principle" of human communication is expressed in the following maxims, which have obvious relevance for the issue of controlling product quality in GIS. (Reported by S.E. Brennan: "Conversation as direct manipulation: an iconoclastic view", in Laurel, 1990).

Maxims of quantity
-make your contribution as informative as required
-do not make your contribution more informative than is required

Maxims of quality
-do not say what you believe to be false
-do not say that for which you lack adequate evidence'

Maxims of manner
-be perspicuous
-avoid obscurity of expression
-avoid ambiguity
-be brief
-be orderly

Maxim of relation
-be relevant

References

Couclelis, H. (in press) "Location, place, region and space" in R. Abler, M. Marcus and J. Olson (eds) Geography's Inner Worlds, Rutqers University Press, NJ.

Dennett, D. C. 1978 Brainstorms: Philosophical essays on mind and Psychology. Bradford Books, Montgomery, VT.

Dreyfus, H. L. 1979 What computers can't do: the limits of artificial intelligence, Harper&Row, NY,

Dutton, G. 1984 "Truth and its consequences in digital cartography", Technical-Papers, 44th Annual Meeting, American Congress on surveying and Mapping, 273-283.

Grice, H. P. 1975 "Logic and conversation" in P. Cole and J Morgan (eds) Syntax and Semantics 3: Speech Acts, Academic Press, NY.

Laurel, B. 1990 The Art of Human-Computer Interface Design
Addison-Wesley, Reading, Mass.

Searle, J. 1969 Speech Acts Cambridge University Press, Cambridge.

Simon, H. 1969 The Sciences of the Artificial The MIT Press, Cambridge, Mass.

Thrift, N. 1985 "of flies and germs: a geography of knowledge" in D. Gregory and J. Urry Social Relation and Spatial Structure Macmillan, London.

Veregin, H. 1989 "A taxonomy of error in spatial databases" NCGIA Technical Pape 89-12, University of California, Santa Barbara.

Winograd, T. and Flores, F. 1987 Understanding Computers and Cognition: a New Foundation for Design Addison-Wesley, Reading, Mass.

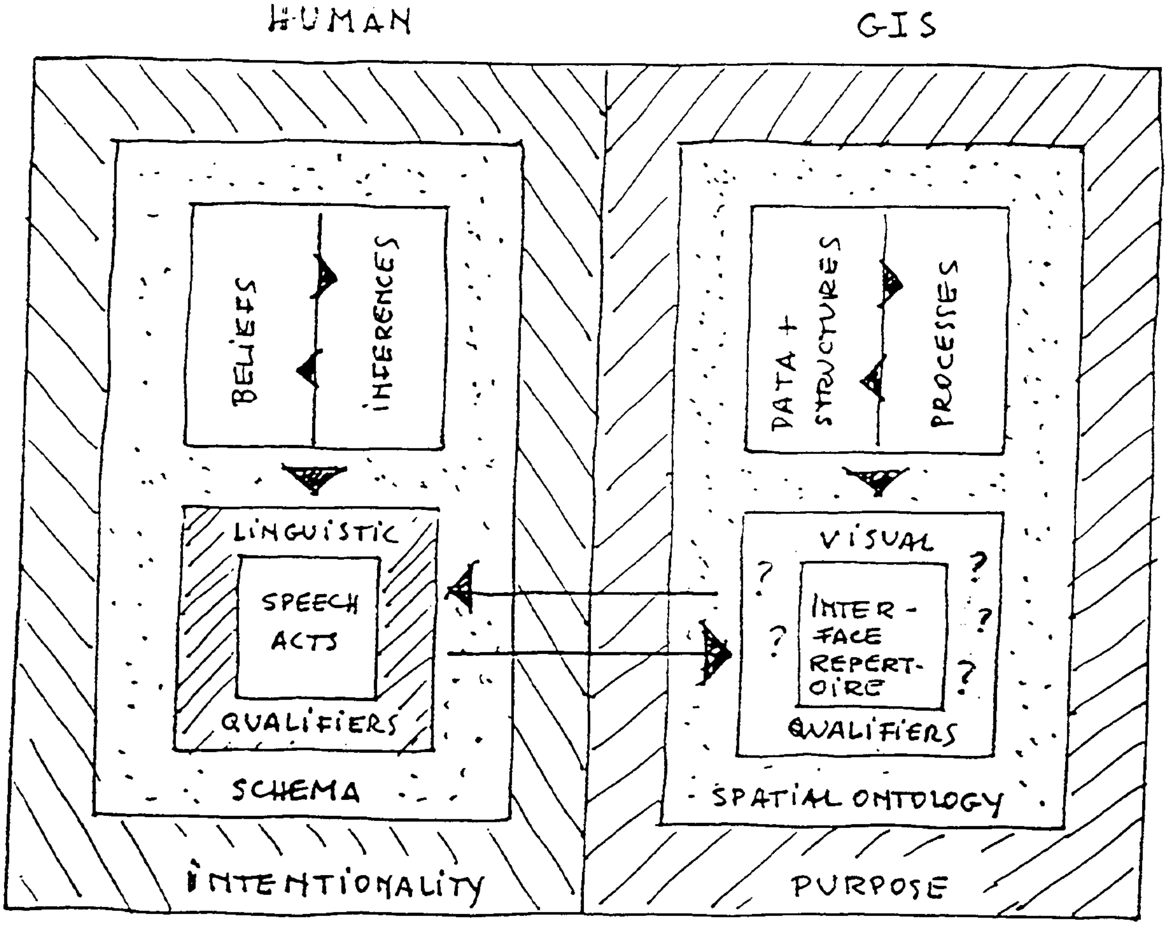


Figure 1. The Conversational metaphor for GIS.

MODEL–BASED SMOOTHING OF SPATIAL DATA

Noel Cressie
Department of Statistics
Iowa State University   Ames IA 50011

A Position Statement for the Specialist Meeting on:
"Visualizing the Quality of Spatial Data"
April 8 1991

1. Introduction

This paper takes a statistical-modeling point of view to the assessment of spatial data quality. Having fit a spatial statistical model, regional variables can be predicted. Then, the quality of that prediction is readily available through the mean squared prediction error.

For the purpose of comparison across regions or through time, variables such as unemployment, crime, or disease incidence should be expressed as a percent or a rate. A large amount of regional data are counts from a base which is itself varying, so that although the standardization in some sense yields comparability of "means," the unequal base from region to region results in unequal "variances". Of course, without a statistical model, there is no notion of mean or variance, but I claim that the very act of shading the rate on a regional map (i.e., choropleth mapping) and then looking for geographical clusters is an attempt to assimilate a more or less vague statistical model.

Unusually high or low regional values are featured by choropleth mapping; "unusually high" and "unusually low" are terms that, in statistical parlance, mean "in the tails of the distribution". However, if the precision of each datum is very different from one to the other, then the data should not be thought of as coming from a single distribution. An unusally high or low rate for a region may be due to very few base counts from year to year, so that chance fluctuations may cause the rate to be high one year and low the next. In order to compare regions spatially, allowance must be made for nonconstant spatial variability of rates.

These issues are raised by Cressie and Read (1989), who analyze SID rates for the counties of North Carolina. Wallin (1984) discusses *inter alia* the problem of mapping ratios and identifies a need for an analysis of their uncertainty and variability. However, he does not pursue this, instead suggesting that mapping of ratios should be avoided. Here, it will be suggested that the problem can be dealt with by spatial modeling and empirical Bayes prediction.

2. Bayesian modeling and empirical Bayes prediction

Often the regional variable being mapped is a ratio of a count to a base count. With sufficient aggregation, the variable has an approximate Gaussian distribution. However, when the regions are small, it may be more sensible to model the counts as Poisson or binomially distributed, or to transform the data so that they are more like Gaussian variables.

2.1 Gaussian model

Suppose that there are n regions in the domain of interest. These are often contiguous but may be separated by water or other regions that are not of interest. Associated with each of the n regions is a variable Z (e.g., percent unemployment, cancer incidence per person years at risk) that is to be mapped. Let

$$\underset{\sim}{Y} \equiv (Y_1, \ldots, Y_n)' , \tag{1}$$

denote the vector of data; Y_i denotes the observation in the i-th region. For various reasons, that will be addressed below, $\underset{\sim}{Y}$ is a noisy version of the variable of interest,

$$\underset{\sim}{Z} \equiv (Z_1, \ldots, Z_n)' . \tag{2}$$

Assume that the conditional distribution of $\underset{\sim}{Y}$ given $\underset{\sim}{Z}$ is given by:

$$\underset{\sim}{Y} \,|\, \underset{\sim}{Z} \sim \mathrm{Gau}(\underset{\sim}{Z}, \Delta) , \tag{3}$$

where "$\sim$" denotes "is distributed as," and $\mathrm{Gau}(\underset{\sim}{\mu}, V)$ denotes an n-dimensional Gaussian distribution with mean $\underset{\sim}{\mu}$ and variance matrix V. In a Bayesian framework, $\underset{\sim}{Z}$ also has a distribution, which is usually called the *prior* distribution. Assume,

$$\underset{\sim}{Z} \sim \mathrm{Gau}(X\underset{\sim}{\beta}, \Gamma)\,, \tag{4}$$

where X is an $n \times p$ matrix of known explanatory variables ($p \leq n$), and $\underset{\sim}{\beta}$ is a $p \times 1$ vector of unknown coefficients. Specifically, (4) assumes that $Z_i = \sum_{j=1}^{p} x_{ij}\,\beta_j + \nu_i;\ i = 1, \ldots, n$, where $\mathrm{var}\{(\nu_1, \ldots, \nu_n)'\} = \Gamma$. For example, $(x_{1j}, \ldots, x_{nj})$ might denote an index of the regions' social composition; then β_j would denote the direction and strength of dependence of the variable Z on this index.

Based on (3) and (4), it is straightforward to show that $(\underset{\sim}{Y}', \underset{\sim}{Z}')'$ has a joint Gaussian distribution with mean $((X\underset{\sim}{\beta})', (X\underset{\sim}{\beta})')'$, and variance matrix,

$$\begin{bmatrix} \Sigma & \Gamma \\ \Gamma & \Gamma \end{bmatrix},$$

where

$$\Sigma \equiv \Delta + \Gamma\,. \tag{5}$$

Thus, the posterior distribution of $\underset{\sim}{Z}$ given $\underset{\sim}{Y}$ is:

$$\underset{\sim}{Z} \mid \underset{\sim}{Y} \sim \mathrm{Gau}(\Gamma\Sigma^{-1}\underset{\sim}{Y} + (I - \Gamma\Sigma^{-1})X\underset{\sim}{\beta}, \Gamma - \Gamma\Sigma^{-1}\Gamma)\,. \tag{6}$$

This last result can be found, e.g., in Lindley and Smith (1972).

The foregoing distribution theory is necessary for the Bayesian decision theory that is to follow. The goal is to predict the random vector $\underset{\sim}{Z}$ from the data vector $\underset{\sim}{Y}$. Let $\underset{\sim}{p}(\underset{\sim}{Y})$ denote any such predictor, and suppose that the loss incurred by using $\underset{\sim}{p}(\underset{\sim}{Y})$, when the true value is $\underset{\sim}{Z}$, is summarized by the matrix,

$$L(\underset{\sim}{Z}, \underset{\sim}{p}) \equiv (\underset{\sim}{Z} - \underset{\sim}{p})(\underset{\sim}{Z} - \underset{\sim}{p})'\,. \tag{7}$$

When only one region is to be predicted, (7) reduces to the usual squared error loss.

Define the risk matrices of the predictors $\underset{\sim}{p}^{(1)}(\underset{\sim}{Y})$ and $\underset{\sim}{p}^{(2)}(\underset{\sim}{Y})$ by,

$$M^{(j)} = E\{(\underset{\sim}{Z} - \underset{\sim}{p}^{(j)}(\underset{\sim}{Y}))(\underset{\sim}{Z} - \underset{\sim}{p}^{(j)}(\underset{\sim}{Y}))'\}\,; \quad j = 1, 2\,,$$

where the expectation is taken jointly over $(\underset{\sim}{Y}', \underset{\sim}{Z}')'$. Then $\underset{\sim}{p}^{(1)}(\underset{\sim}{Y})$ is said to be *better than* $\underset{\sim}{p}^{(2)}(\underset{\sim}{Y})$ if $M^{(2)} - M^{(1)}$ is nonnegative-definite. Now, it is easy to show that the posterior mean $E(\underset{\sim}{Z} \mid \underset{\sim}{Y})$ is better than any other predictor. This result holds for any distribution; from the Gaussian assumptions (3) and (4), it can be seen from (6) that

$$E(\underset{\sim}{Z} \mid \underset{\sim}{Y}) = \underset{\sim}{p}^{*}(\underset{\sim}{Y}) \equiv \Gamma(\Delta + \Gamma)^{-1}\underset{\sim}{Y} + (I - \Gamma(\Delta + \Gamma)^{-1})X\underset{\sim}{\beta}\,, \tag{8}$$

is better than any other predictor. That is, (8) "minimizes" the mean squared prediction error matrix,

$$E\{(\underset{\sim}{Z} - \underset{\sim}{p}(\underset{\sim}{Y}))(\underset{\sim}{Z} - \underset{\sim}{p}(\underset{\sim}{Y}))'\}\,, \tag{9}$$

over $\underset{\sim}{p}$. Hence, $\underset{\sim}{p}^{*}(\underset{\sim}{Y})$ is an attractive smoother of the data $\underset{\sim}{Y}$, since it predicts the vector $\underset{\sim}{Z}$ with smallest mean squared prediction error. In statistical parlance, (9) is called a *Bayes predictor*.

Since $\underset{\sim}{\beta}$ in (4) is unknown, (8) is not yet a statistic (i.e., is not a function only of the data). Moreover, the model variance matrix Δ in (3) is often expressed in terms of parameters, and likewise for the prior variance matrix Γ in (4). The proper Bayesian approach would be to put further priors and hyperpriors on all unknown parameters. This solution to the conundrum of unknown parameters is sometimes called hierarchical Bayes, and demands a prior knowledge of process variability that many scientists and engineers do not feel they have.

An alternative approach, the one suggested in this paper, is to treat all parameters, except $\underset{\sim}{Z}$, as fixed but unknown, and use the data $\underset{\sim}{Y}$ to estimate them. This approach is called *empirical Bayes*.

Henceforth, assume that $\underset{\sim}{\theta}$ is a $k \times 1$ vector of variance matrix parameters; write,

$$\Sigma(\underset{\sim}{\theta}) = \Delta(\underset{\sim}{\theta}) + \Gamma(\underset{\sim}{\theta}), \tag{10}$$

although the parameter set from Δ and that from Γ are assumed disjoint. (For brevity, sometimes the dependence of Σ, Δ, and Γ on $\underset{\sim}{\theta}$ is dropped.)

Assume, for the moment, that $\underset{\sim}{\theta}$ is known, but that $\underset{\sim}{\beta}$ is unknown. Now, the generalized least squares estimator of $\underset{\sim}{\beta}$ is

$$\hat{\underset{\sim}{\beta}} \equiv (X'\Sigma(\underset{\sim}{\theta})^{-1}X)^{-1}X'\Sigma(\underset{\sim}{\theta})^{-1}\underset{\sim}{Y}.$$

Then, it is easily shown that the best linear unbiased predictor (BLUP) is

$$\hat{\underset{\sim}{p}}(\underset{\sim}{Y};\underset{\sim}{\theta}) \equiv \Gamma(\underset{\sim}{\theta})\Sigma(\underset{\sim}{\theta})^{-1}\underset{\sim}{Y} + \{I - \Gamma(\underset{\sim}{\theta})\Sigma(\underset{\sim}{\theta})^{-1}\}X\hat{\underset{\sim}{\beta}} \equiv \{I - \Delta(\underset{\sim}{\theta})\Pi(\underset{\sim}{\theta})\}\underset{\sim}{Y}, \tag{11}$$

where

$$\Pi(\underset{\sim}{\theta}) = \Sigma(\underset{\sim}{\theta})^{-1} - \Sigma(\underset{\sim}{\theta})^{-1}X(X'\Sigma(\underset{\sim}{\theta})^{-1}X)^{-1}X'\Sigma(\underset{\sim}{\theta})^{-1}. \tag{12}$$

The mean-squared-prediction-error matrix of $\hat{\underset{\sim}{p}}(\underset{\sim}{Y};\underset{\sim}{\theta})$, given by (11), is,

$$\begin{aligned} M_1(\underset{\sim}{\theta}) &\equiv E\{(\underset{\sim}{Z} - \hat{\underset{\sim}{p}}(\underset{\sim}{Y};\underset{\sim}{\theta}))(\underset{\sim}{Z} - \hat{\underset{\sim}{p}}(\underset{\sim}{Y};\underset{\sim}{\theta}))'\} \\ &= \{I - \Delta(\underset{\sim}{\theta})\Pi(\underset{\sim}{\theta})\}\Sigma(\underset{\sim}{\theta})\{I - \Delta(\underset{\sim}{\theta})\Pi(\underset{\sim}{\theta})\}'. \end{aligned} \tag{13}$$

The more realistic state of knowledge is where *both* $\underset{\sim}{\beta}$ *and* $\underset{\sim}{\theta}$ are unknown. The empirical Bayes predictor, $\hat{\underset{\sim}{p}}(\underset{\sim}{Y};\underset{\sim}{\theta})$ is no longer a statistic, although it becomes one if an estimator $\hat{\underset{\sim}{\theta}}$ is substituted for $\underset{\sim}{\theta}$.

2.2 Poisson model

Suppose that the data are counts $\{C_i : i = 1, \ldots, n\}$ and the variable of interest is the relative risk $\{Z_i : i = 1, \ldots, n\}$ for each region. Assume that the conditional distribution of C_i given Z_i is given by:

$$C_i \mid Z_i \sim Po(Z_i n_i), \tag{14}$$

where n_i is the total number at risk in the i-th region; $i = 1, \ldots, n$. In (14), $Po(\lambda)$ denotes a Poisson distribution with mean λ, the distributions are assumed independent, and $\{n_i : i = 1, \ldots, n\}$ is a known exogenous variable. Clayton and Kaldor (1987) propose a spatial linear model for

$$W_i = \log Z_i\,; \quad i = 1, \ldots, n. \tag{15}$$

More generally, suppose

$$\underset{\sim}{W} \sim \text{Gau}(X\underset{\sim}{\beta}, \Gamma(\underset{\sim}{\theta})). \tag{16}$$

The prior parameters $\underset{\sim}{\beta}$ and $\underset{\sim}{\theta}$ could either be assumed distributed according to some hyperprior, resulting in an hierarchical Bayesian predictor of $\underset{\sim}{Z}$, or could be assumed fixed but unknown and then estimated from the data. This latter approach yields an empirical Bayes predictor of $\underset{\sim}{Z}$.

3. Estimation of model and prior parameters

In this and subsequent sections, only the Gaussian model of Section 2.1 will be considered. Consider the optimal (Bayes) predictor (8); in reality, both $\underset{\sim}{\beta}$ and $\underset{\sim}{\theta}$ are unknown. Now, under the model (3) and (4), the marginal distribution of $\underset{\sim}{Y}$ is,

$$\underset{\sim}{Y} \sim \text{Gau}(X\underset{\sim}{\beta}, \Sigma(\underset{\sim}{\theta})), \tag{17}$$

where $\Sigma(\underset{\sim}{\theta})$ is given by (11). Then, the negative loglikelihood of $\underset{\sim}{\beta}$ and $\underset{\sim}{\theta}$ is,

$$L(\underset{\sim}{\beta},\underset{\sim}{\theta}) \equiv (n/2)\log(2\pi) + (1/2)\log(|\Sigma(\underset{\sim}{\theta})|) + (1/2)(\underset{\sim}{Y} - X\underset{\sim}{\beta})'\,\Sigma(\underset{\sim}{\theta})^{-1}(\underset{\sim}{Y} - X\underset{\sim}{\beta}), \tag{18}$$

which is to be minimized. The scoring (or Gauss-Newton) algorithm can be used to yield maximum likelihood (m.l.) estimators $\hat{\underset{\sim}{\beta}}$ and $\hat{\underset{\sim}{\theta}}$.

4. Properties of the empirical Bayes predictor: bias and mean squared prediction error

Consider the optimal (Bayes) predictor $\underset{\sim}{p}^*(\underset{\sim}{Y})$ given by (8). Since the Gaussian model is assumed, this is better than any other predictor, and it is clearly unbiased since $\underset{\sim}{p}^*(\underset{\sim}{Y}) = E(\underset{\sim}{Z}\,|\,\underset{\sim}{Y})$.

When $\underset{\sim}{\theta}$ is known but $\underset{\sim}{\beta}$ is unknown, the predictor $\underset{\sim}{\hat{p}}(\underset{\sim}{Y};\underset{\sim}{\theta})$ given by (11) is unbiased and better than any other linear unbiased predictor. However, when $\underset{\sim}{\theta}$ is also unknown and estimated by $\hat{\underset{\sim}{\theta}}$, the empirical Bayes predictor $\underset{\sim}{\hat{p}}(\underset{\sim}{Y};\hat{\underset{\sim}{\theta}})$ is typically no longer linear nor Bayes. Nevertheless, its close ties with the BLUP, $\underset{\sim}{\hat{p}}(\underset{\sim}{Y};\underset{\sim}{\theta})$, makes it the predictor of choice. (Sometimes, the empirical Bayes predictor $\underset{\sim}{\hat{p}}(\underset{\sim}{Y};\hat{\underset{\sim}{\theta}})$ is referred to as the EBLUP or estimated best linear unbiased predictor.)

Using analogous arguments to Zimmerman and Cressie (1991), it is straightforward to show that

$$E(\underset{\sim}{\hat{p}}(\underset{\sim}{Y};\hat{\underset{\sim}{\theta}})) = E(\underset{\sim}{Z}) = X\underset{\sim}{\beta}\,. \tag{19}$$

That is, the EBLUP is unbiased.

Consider now the mean-squared-prediction-error matrix of $\underset{\sim}{\hat{p}}(\underset{\sim}{Y};\hat{\underset{\sim}{\theta}})$, where $\hat{\underset{\sim}{\theta}}$ is even and translation invariant. Define

$$M_2(\underset{\sim}{\theta}) \equiv E\{(\underset{\sim}{Z} - \underset{\sim}{\hat{p}}(\underset{\sim}{Y};\hat{\underset{\sim}{\theta}}))(\underset{\sim}{Z} - \underset{\sim}{\hat{p}}(\underset{\sim}{Y};\hat{\underset{\sim}{\theta}}))'\}\,, \tag{20}$$

which should be compared to $M_1(\underset{\sim}{\theta})$ given by (13). Since $M_2(\underset{\sim}{\theta})$ is defined for a predictor that involves the estimation of $\underset{\sim}{\beta}$ *and* $\underset{\sim}{\theta}$, it is expected that $M_2(\underset{\sim}{\theta}) - M_1(\underset{\sim}{\theta})$ will be nonnegative definite. The results of Harville (1985) can be used to establish the truth of this conjecture. (The Gaussian assumptions are important here.)

But, there is another potential source of bias due to the fact that $M_1(\hat{\underset{\sim}{\theta}})$, not $M_1(\underset{\sim}{\theta})$, is used to estimate mean squared prediction errors. Suppose that $\hat{\underset{\sim}{\theta}}$ is chosen so that $E(\Sigma(\hat{\underset{\sim}{\theta}})) = \Sigma(\underset{\sim}{\theta})$ and $E(\Gamma(\hat{\underset{\sim}{\theta}})) = \Gamma(\underset{\sim}{\theta})$. Then, the results of Eaton (1985) and Zimmerman and Cressie (1991) can be used to establish that $M_1(\underset{\sim}{\theta}) - E(M_1(\hat{\underset{\sim}{\theta}}))$ is nonnegative-definite. (The proof relies on a multivariate version of Jensen's inequality, and the fact that $\underset{\sim}{\hat{p}}(\underset{\sim}{Y};\underset{\sim}{\theta})$ is best among all linear unbiased predictors.)

Even if an expression for $M_2(\underset{\sim}{\theta})$ were known, it is likely that $M_2(\hat{\underset{\sim}{\theta}})$ would be biased, further illustrating the inherent difficulty in estimating mean squared prediction errors. An approximation based on an asymptotic expansion of $M_2(\underset{\sim}{\theta})$ has been suggested by Prasad and Rao (1990); call the resulting estimator $M_2(\underset{\sim}{\theta})^*$. Then, an approximately unbiased estimator of $M_2(\underset{\sim}{\theta})$ is:

$$M_2(\underset{\sim}{\theta})^* \equiv M_1(\hat{\underset{\sim}{\theta}}) + 2C(\hat{\underset{\sim}{\theta}})\,; \tag{21}$$

in (21), $C(\underset{\sim}{\theta})$ has (i,j)-th element,

$$c_{ij}(\underset{\sim}{\theta}) \equiv \mathrm{tr}\{A_{ij}(\underset{\sim}{\theta})\,B(\underset{\sim}{\theta})^{-1}\}\,, \tag{22}$$

where

$$A_{ij}(\underset{\sim}{\theta}) \equiv \mathrm{cov}(\partial\hat{p}_i(\underset{\sim}{Y};\underset{\sim}{\theta})/\partial\underset{\sim}{\theta}, \partial\hat{p}_j(\underset{\sim}{Y};\underset{\sim}{\theta})/\partial\underset{\sim}{\theta})\,, \tag{23}$$

and $B(\underset{\sim}{\theta})^{-1}$ is the (asymptotic) variance matrix of the estimator $\hat{\underset{\sim}{\theta}}$.

Actually, Prasad and Rao's (1990) result allows one to conclude only the equality of diagonal elements of (22). The multivariate result is important if it is decided later to aggregate regions.

Suppose that region i is aggregated with region j, which is denoted $i \cup j$, and that the new variable to be predicted is, $Z_{i \cup j} = \ell_i Z_i + \ell_j Z_j$. Then the empirical Bayes predictor is, $\ell_i \hat{p}_i(\underset{\sim}{Y}; \underset{\sim}{\hat{\theta}}) + \ell_j \hat{p}_j(\underset{\sim}{Y}; \underset{\sim}{\hat{\theta}})$, and its mean squared prediction error is, $\ell_i^2 [M_2(\underset{\sim}{\theta})]_{ii} + \ell_j^2 [M_2(\underset{\sim}{\theta})]_{jj} + 2\ell_i \ell_j [M_2(\underset{\sim}{\theta})]_{ij}$. Thus, in the latter expression, an off-diagonal element of the matrix $M_2(\underset{\sim}{\theta})$ is needed.

5. Discussion

The previous sections give the rationale and the formulas behind smoothing regional maps with an empirical Bayes predictor. Provided the Bayesian model is appropriate (which can be checked using exploratory data analysis and statistical diagnostics; see, e.g., Cressie and Read, 1989), the empirical Bayes predictor possesses statistical optimality properties that make it a very attractive smoother. Moreover, each smoothed value is accompanied by a measure of quality, namely, its mean squared prediction error (whose estimation is discussed in Section 4).

Clearly, empirical Bayes prediction of a region requires a lot more thought and effort than, say, an average over the region and its nearest neighbors. But, the former approach has considerable advantages when one wishes to confirm, statistically, impressions gained from studying a smoothed regional map. Valid inference is relatively straightforward once a statistical model has been fit.

5.1 Quality of spatial data

There are a number of open problems suggested by the approach taken in this paper. If another regional variable $\underset{\sim}{W}$ is observed along with $\underset{\sim}{Y}$, it should be possible to use any correlation between the two variables to improve further the prediction of $\underset{\sim}{Z}$. In the geostatistical literature, this is known as cokriging (e.g., Journel and Huijbregts, 1978, p. 324ff.).

For some purposes, the empirical Bayes predictor $\underset{\sim}{\hat{p}}(\underset{\sim}{Y}; \underset{\sim}{\hat{\theta}})$ of $\underset{\sim}{Z}$ is *too* smooth. For example, to predict the (weighted) proportion of regions whose Z-value is greater than a given $z_0, \underset{\sim}{\hat{p}}(\underset{\sim}{Y}; \underset{\sim}{\hat{\theta}})$ is not appropriate. Since var($\underset{\sim}{Z}$ − var($E(\underset{\sim}{Z} \mid \underset{\sim}{Y})$) is always nonnegative-definite, a "rougher" predictor is needed. The *constrained* empirical Bayes predictor (Louis, 1984; Cressie, 1989) is

$$\underset{\sim}{p}^{\circledast}(\underset{\sim}{Y}) \equiv X\underset{\sim}{\hat{\beta}} + \Gamma(\underset{\sim}{\hat{\theta}})^{1/2}\, \Sigma(\underset{\sim}{\hat{\theta}})^{-1/2}(\underset{\sim}{Y} - X\underset{\sim}{\hat{\beta}}),$$

which is to be compared with,

$$\underset{\sim}{\hat{p}}(\underset{\sim}{Y}; \underset{\sim}{\hat{\theta}}) = X\underset{\sim}{\hat{\beta}} + \Gamma(\underset{\sim}{\hat{\theta}})\, \Sigma(\underset{\sim}{\hat{\theta}})^{-1}(\underset{\sim}{Y} - X\underset{\sim}{\hat{\beta}}).$$

Thus, $\underset{\sim}{p}^{\circledast}(\underset{\sim}{Y})$ involves less "shrinkage," of the data $\underset{\sim}{Y}$ towards the estimated model mean $X\underset{\sim}{\hat{\beta}}$, than does $\underset{\sim}{\hat{p}}(\underset{\sim}{Y}; \underset{\sim}{\hat{\theta}})$. To the leading order of magnitude, the following approximation holds:

$$\text{var}\{\underset{\sim}{p}^{\circledast}(\underset{\sim}{Y})\} \simeq \text{var}\{\Gamma(\underset{\sim}{\theta})^{1/2}\, \Sigma(\underset{\sim}{\theta})^{-1/2}\,(\underset{\sim}{Y} - X\underset{\sim}{\beta})\} = \Gamma(\underset{\sim}{\theta}) = \text{var}(\underset{\sim}{Z}).$$

Then the Gaussian model guarantees that, to the same leading order of magnitude, the quantity $\underset{\sim}{g}(\underset{\sim}{p}^{\circledast}(\underset{\sim}{Y}))$ is an unbiased predictor of $\underset{\sim}{g}(\underset{\sim}{Z})$. One would use this predictor in cases where calculating the optimal predictor $E(\underset{\sim}{g}(\underset{\sim}{Z}) \mid \underset{\sim}{Y})$ is computationally prohibitive.

5.2 Visualizing spatial data

The choropleth maps that are usually used to depict regional maps have the disadvantage that large areas of low population density dominate the map, whereas often most of the interest is in cities, small areas of high population density. One possible alternative is the demographic base map (e.g., Forster, 1972; Kidron and Segal, 1984), where the area of the i-th region is made proportional to the base n_i; $i = 1, \ldots, n$. Contiguity of the geographical boundaries and the relative geographical positions are maintained as far as possible. Although offering interesting possibilities, constructing the required map is difficult and often arbitrary.

Cleveland and McGill (1984) and Cleveland (1985, p. 208) suggest that the traditional choropleth map be replaced with a regional map that has framed rectangles of *equal* size superimposed on, or attached in some way to, the corresponding regions. The framed rectangle is like a chemist's measuring beaker with differing amounts of black liquid in it; its degree of "fullness" is proportional to the observed (transformed) incidence rate. Dunn (1988) describes an experiment in graphical perception that demonstrates the superiority of the framed-rectangle method of mapping over choropleth mapping. One possible enhancement is to make the width of the i-th framed rectangle proportional to $n_i^{1/2}$ (i.e., wider framed rectangles correspond to rates with smaller standard deviations).

Having spent considerable effort in obtaining accurate estimators of the quality of predicted data (i.e., estimated mean squared prediction errors), one should also visualize that data quality. At the very least, another choropleth map depicting *root* mean squared prediction errors, superimposed on the regions, should be presented side-by-side with the choropleth map of predicted data. It is an open problem to make just one map with *both* the predicted values and their root mean squared prediction errors presented effectively.

References

Clayton, D., and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, **43**, 671–681.

Cleveland, W. S. (1985). *The Elements of Graphing Data.* Wadsworth, Monterey, CA.

Cleveland, W. S., and McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, **79**, 531–554.

Cressie, N. (1989). Empirical Bayes estimation of undercount in the Decennial Census. *Journal of the American Statistical Association* 84, 1033–1044.

Cressie, N., and Read, T. R. C. (1989). Spatial data analysis of regional counts. *Biometrical Journal*, **31**, 699–719.

Dunn, R. (1988). Framed rectangle charts or statistical maps with shading. An experiment with graphical perception. *American Statistician*, 42, 123–129.

Eaton, M. L. (1985). The Gauss-Markov Theorem in multivariate analysis, in *Multivariate Analysis–VI*, ed. P.R. Krishnaiah. Elsevier, Amsterdam, 177–201.

Forster, F. (1972). Use of a demographic base map for the presentation of areal data in epidemiology, in *Medical Geography*, ed. N.D. McGlashan. Methuen, London, 59–67.

Harville, D. A. (1985). Decomposition of prediction error. *Journal of the American Statistical Association*, **80**, 132–138.

Journel, A.G., and Huijbregts, C. J. (1978). *Mining Geostatistics.* Academic Press, London.

Kidron, M., and Segal, R. (1984). *The New State of the World Atlas.* Simon and Schuster, New York.

Lindley, D. V., and Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society B*, **34**, 1–41.

Louis, T. A. (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. *Journal of the American Statistical Association*, 79, 393–398.

Prasad, N. G. N., and Rao, J. N. K. (1990). On the estimation of mean square error of small area predictors. *Journal of the American Statistical Association*, 84, 163–171.

Wallin, E. (1984). Isarithmic maps and geographic disaggregation, in *Proceedings of the International Symposium on Spatial Data Handling* (Zurich, Switzerland, 1984). Geographisches Institut, Universitat Zurich-Irchel, Zurich, Switzerland, 209–217.

Zimmerman, D. L., and Cressie, N. (1991). Mean squared prediction error in the spatial linear model with estimated covariance parameters. *Annals of the Institute of Statistical Mathematics*, 43, forthcoming.

POSITION PAPER
for

NCGIA I-7 Specialist Meeting, June 8-12, Castine

VISUALIZATION OF DATA QUALITY

Ferenc Csillag
Department of Geography,
Syracuse University, Syracuse, NY 13244-1160
[fcsillag@sunrise.bitnet]

What is the problem?

My first problem is exactly this: "What is the *accuracy*?" type of questions do not seem to lead to meaningful answers, because a certain piece of information can be used for inference, prediction, or even as a guess in many ways, where judgement of quality may be different.

There are two consequences of the first problem: (1) The need for specific accuracy measures (like soil pH of a spot, population of a census block, or high risk of groundwater contamination); and (2) The need for general data descriptors (also referred to as metadata) which, in turn, can be incorporated in such specific measures.

Both of the previous needs generate further requirements in terms of visualization, i.e. how can one *see*, for example, a specific accuracy measure.

Finally, all this can be viewed, and reconsidered, taking into account current GIS software, existing and envisioned standards, databases and their use.

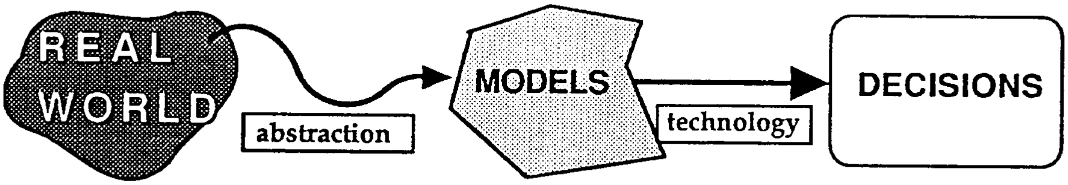
Some general comments

A digital thermometer on my wall displays only the following message: "Do not use this instrument to find out temperature, because it always gives false information!" If I knew that the thermometer on my wall showed 100F more than the actual temperature, I would easily reduce this type of error by subtracting 10 from the actual reading. This primitive example reveals two important aspects of the quality of information: (1) the nature of providing quality information, and (2) the nature of using this information.

In the context of current geographic information systems the nature of quality information may not fit well into the processing strategies of softwaremakers. I think the primary reason for this is that *data* are rarely considered to be an integral part of the system, consequently the software vendor is attempted to sell a "never-breaking' software. That seems to be impossible. Certainly, users contribute to this concept when they try to go for 'error-free' data sets and 100% accuracy.

Errors and data models

One can find a number of key references to attack the general problem of errors, including definitions of terms like precision, accuracy, reliability, uncertainty, etc., and their nature in GIS (Burrough 1986, Chrisman 1986, Chrisman 1989, Goodchild 1990). Beyond the 'obvious', their major message is that one should know a lot. It's not always clear, however, what should we know. I tend to decompose this knowledge into two components:



This Position Paper is partly based on my contribution and experience at the panel held at the AutoCarto-10, in Baltimore, March 27, 1991 (Buttenfield and Beard 1991). The depth of the sections is heterogenous by intention.

where the first step on the left is somewhat philosophical, requiring understanding of the nature of data, while the second step is more technical/mechanical. This is because abstraction, model creation, really works with real world entities and processes, so imagination and creativity can take an important role in the construction of the models. The application of these models is generally based on some form of mathematics.

My interest is focused on a subset of all possible, geographically relevant abstractions, namely the ones which deal with *neighborhood(s)* in relationship to time, space and attributes (Sinton 1978, Dueker 1979, Csillag 1991). This type of abstraction leads to statistical models based on assumptions about the data. The sharp separation between the above mentioned two steps must not be missed: It’s nature that dictates the kind of models, and statistics provides the tools to handle them. (In many cases the ’thought-about’ models do not lead to easily solvable mathematical problems. Then one can go for a simpler case with reducing the number of unknowns, but it’s never statistics that tells which family of models to choose from.)

The family of data models I find rather powerful in environmental mapping is based on the estimation of covariance functions. In this particular case, having a sample of size n, the model is concerned with the Stochastic variables having joint normal distribution. It is crucial to everyday practice that we hardly have any tools to check this assumption. It is especially difficult, because the sample taken at n locations is a single realization of the variables. Furthermore, it is assumed that the expected value of this distribution is zero, and the variance is finite. So with this model we are confined in our prediction to the case, when, somehow, our original problem has been reduced to a zero-mean variable. With these assumptions we can prove that the covariance exists and it is positive semi-definite. It is our task now to construct an estimate of our distribution so the variance of the difference between the model and the estimate should be minimum. It is only due to the joint-normality assumption that our search for the estimate can be restricted for linear functions, i.e. in the form of weighted sum:

$$\hat{\xi}(x) = \sum_{i} \lambda_{i} \xi x_{(i)}$$

The major problem in constructing our estimate is that we may not have sufficient information about the covariance, therefore further assumptions will be necessary. For instance, stationarity is a quite frequent assumption in order to reduce dramatically the number of elements to be estimated in the covariance matrix:

$$COV[\xi(x)] = \mathfrak{f}[\Delta x)$$

with eliminating everything except distance.

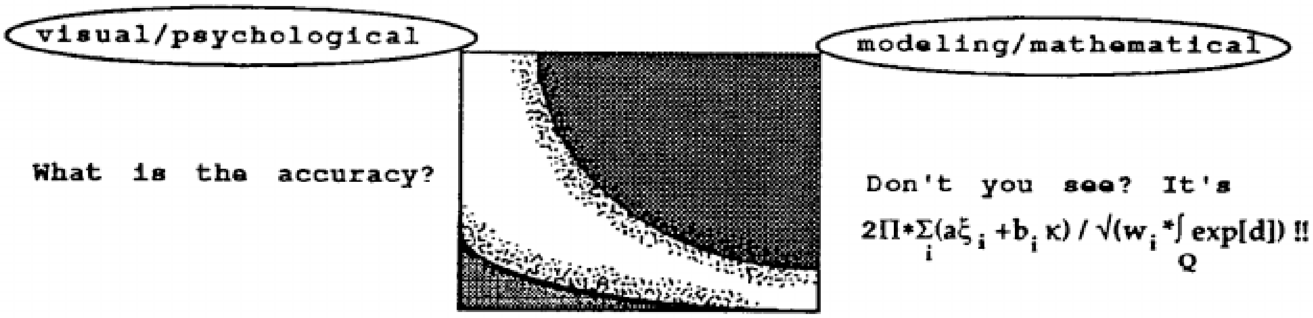
There is a huge variety of other kinds of data models applicable in a GIS environment. Unfortunately, it’s impossible not to realize that due to its ease statistics based on independent sampling is by far the most widely used. (A clear case when technology overrides abstraction.) Also there are numerous other (e.g. local vs. global, geometric vs. statistical) aspects of the data model problem.

Error-sensitivity analysis of applications

Application-specific accuracy measures can be viewed as a function where general data characteristics are one of the arguments. When complex models are applied on complex data it is difficult to immediately understand how sensitive the result is to (a) input data, (b) processing. This requires the incorporation of further models on propagation in addition to the data characteristics. The major challenge in this step is to understand the relationships between the 'ultimate answer to the question' and the actual data manipulation (Beckett and Burrough 1971, Webster 1977, Stein et al. 1988).

Is WYSIWYG?

When having a closer look at visualization tools another abstraction/ technology boundary can be found. Once a choice is made about what should be seen, the appropriate representation should be found (McCormick et al. 1987). The situation suggested by the figure below is not rare:

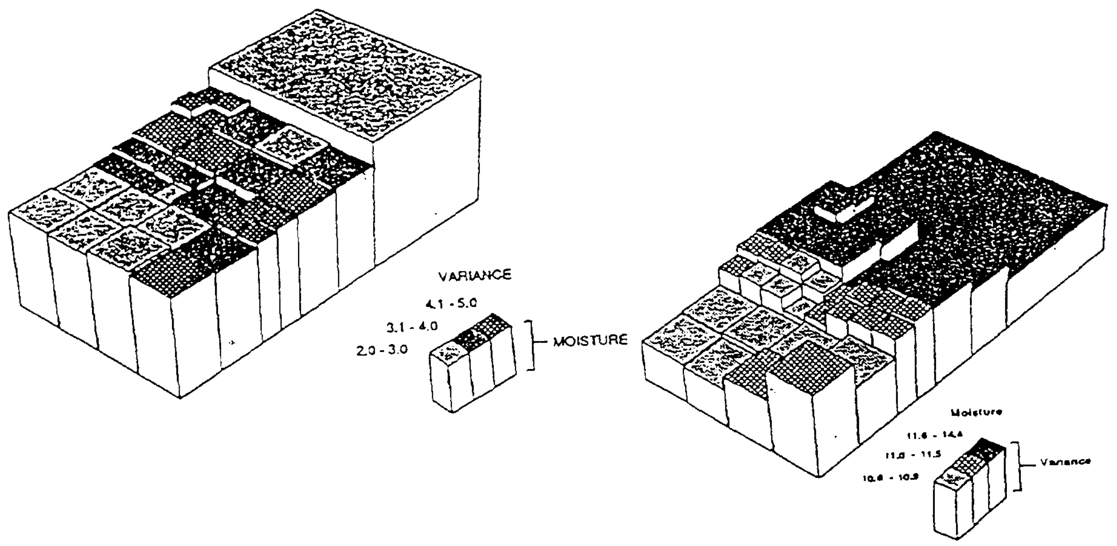


Favourite questions with illustration

Some examples of the types of questions related to data quality I get excited about are as follows (with short reference to the envisioned technological solution and field of application):

- Having measured a variable at a number of locations, what is the most informative way to ’construct areal units? [interpolation, variability; typical in soil mappping]
- Having a partition on attribute values, how can I define a spatial partition so that spatial neighbors will be neighbors in the feature space as well? [classification, variability; frequent in remote sensing image processing]
- Having a partition in the mapping space, what is the amount of information that can be derived about the efficiency of feature space partition corresponding to it? [interpolation, classification; frequent in landuse mapping]

Finally, as a complex illustrative example, the figure below shows a representation where statistical model, data structure and visualization tool have been integrated:



The map is based on a sample measured on a regular grid. Block kriging was used as interpolation with block sizes corresponding to quadtree-decomposition. A leaf was kept iff its estimation variance was lower than the limit specified by the user. Note that resolution and variance minima do not always coincide.
[from Csillag and Kertész 1989]

References

Beckett, P.H. and Burrough, P.A. (1971) The relation between cost and utility in soil survey: V. The cost-effectiveness of different soil survey procedures; Journal of Soil Science, 22(4):481-489.

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

Buttenfield, B.P. and M.K.Beard (1991) Visualizing the quality of spatial information; Proc. AutoCarto-10 pp.423-427, ACSM-ASPRS, Bethesda.

Chrisman, N.K. (1986) Obtaining information on qulity of digital data; Proc. AutoCarto London, 1:350-358.

Chrisman, N.K. (1989) A taxonomy of error applied to categorical maps; International Cartographic Assoc. World Congress Budapest, (manuscript).

Csillag, F. (1991) Resolution Revisited; Proc. Auto-Carto 10. pp. 15-28., ACSM-ASPRS, Bethesda.

Csillag, F. and M. Kert6sz (1989) Spatial variability: Error in natural resource maps?; Agrokemia es Talajtan 37:715-726.

Dueker, K.J. (1979) Land resource information systems: spatial and attribute resolution issues; Proc. AutoCarto IV 2:328-337. ASP-ACSM, Falls Church.

Goodchild, M. (1990) Modeling error in spatial databases; Proc. GIS/LIS. San Antonio.

McCormick, B. H, T.A. DeFanti and M.D. Brown eds. (1987) Visualization in scientific computing; Computer Graphics 21, ACM SIGGRAPH, New York.

Sinton, D. (1978) The inherent structure of information as a constraint to analysis: mapped thematic data as a case study; Harvard Papers on Geographic Information Systems (ed. G. Dutton) Vol. 7. Addison-Wesley, Reading.

Stein, A., Hoogerwerf, M. and Bouma, J. (1988) Use of soil map delineations to improve (co)kriging of point data on moisture deficits. Geoderma 43:163-178.

Webster, R. (1977) Quantitative and numerical methods in soil classification and survey; Oxford University Press, Oxford.

Probability Filtering for Fuzzy Features

Geoffrey Dutton
Spatial Effects
150 Irving Street
Watertown MA 02172 USA
qtm@cup.portal.com

Abstract

Inaccuracies in digitized map features may be both analyzed and visualized. Some, positional errors derive from circumstances of data capture, others from the ways in which features are represented in a database, and others stem from how data are manipulated (particularly when datasets; are merged or graphic scales change). The first two of these are quite intimately related, as map digitizing - whether via manual tracing or by scanning and vectorization - generates lists of coordinates which are the basis for further data structuring. This paper addresses the ways in which this (vector) mode of representation affects the inherent quality of digital cartographic lines, and by implication, the quality and usability of geographic databases based on digitized map features. It demonstrates how locational uncertainty about point, line and areal cartographic features derives from mislocation of the coordinates that define them, regardless of what causes these mislocations in the first place (hardware error or operator blunders during digitizing, finite line widths on source maps, numerical instability in computations, etc.). An analysis of positional error is offered which allows one to characterize the uncertainty of digital map features in terms of regions of constant probability. This seems to provide a rational basis for automated selection of features when changing the scale of map displays, and for resolving conflicts when features are integrated or otherwise compete for space.

Prologue

One of the earliest forms of spatial data structures was the "location list" (Peucker and Chrisman, 1975), in which cartographic lines and regions are represented by lists of x,y or x,y,z tuples. Each cartographic object encoded in a location list either a point, a line or a polygon, and each is stored separately. The "sliver problem" first arose when maps were digitized in this format. The problem actually derives from the digitizing process, not from the data structure itself, although the data structure offered no practical way to overcome it. Each object owned a set of coordinates, for which there was no cross-references to other objects, even if they had points in common. There was no modeling of boundaries between regions, and no idea that features could share space. This data model came to be called "spaceship polygons" (although it also encompasses point and linear features). It is still widely used, especially in thematic mapping, but is disappearing from GIS databases.

With the adoption of a topological framework, map features no longer had to exist as isolated entities. This primarily applied to polygons; unconnected points and lines have no adjacent objects, although they can be identified via regions that enclose them. Still, by imposing necessary and sufficient topological predicates, encoding the connectivity and adjacency of map features solved a number of nagging problems, and became the basis for a whole generation of GIS databases.

Analyzing space in a GIS requires integrating cartographic features that may or may not be in the same layer and may or may not come from the same source. Whether the operation involves map sheet matching, feature identification or overlay, one challenge is always present: two or more independent but unified sets of features must be combined into a single web. Although in some instances information may exist about common identifiers (feature codes) or locations (such as control points, or survey monuments), brute force geometric rectification (with or without tolerance filtering) normally is used to integrate separate coverages. While the objects created may all be assigned unique and consistent identifiers, nothing about the process assures that all of them are in fact meaningful map features, or at which scales they can be said to exist. Slivers have struck back with a vengeance.

What’s My Line?

How faithfully can spatial entities be digitized? More specifically, how certain are locations along digitized lines? A brief thought experiment: Suppose a curvalinear map feature such as a stream segment is digitized by selecting points along it and “connecting the dots" with straight line segments; an infinity of representations are possible, some more faithful to its shape than others. Visualization (V) I depicts this process of abstraction. The density of selected points will tend to vary with the line complexity in their neighborhoods. If sufficient care is taken in sampling points along curves, one's digitized approximation may not visibly deviate from the graphic source. But as the source is itself only an rendering of a real-world entity (which has a finite and possibly variable width), digital caricatures of it are inherently uncertain, however faithful. Given this, one may assume that every point recording a feature's shape is to some degree uncertain, as are the line segments that their sequencing defines. Call this assumption a (for accuracy).

One may further assert that there is *no systematic bias* (assumption b) in representing the stream, which is to say that its center, rather than one of its banks, has been symbolized (this is reasonable to assume when tracing many source maps produced by government agencies). Given these assumptions, the positions of points selected for digitizing are subject to variation within some defined limit (the width of the stream or its approximation). If the stream is tidal, if its volume varies seasonally, or if its channel contains boulders, coulees, beaches or bars, its width (at least for navigational purposes) will vary, and in places may become rather indeterminate. If one's goal is to represent the real world (rather than lines on a map) in a database, there is no precise way to express this with a single line.

One way to operationally handle such variability and indeterminancy is to assign to each point a (circular) "locus of uncertainty" that represents some limit to how far a point may wander from the center of the stream and still be in the water. At any given vertex the actual extent of this radius may be hard to know, but estimates of it may suffice. As assumption b excludes systematic bias, one may draw a circle around the centerline at each digitized point, within which any location can be considered a reasonable alias for that point. Although it is not necessary, one can further posit that sampled points are drawn from *a normal distribution* (call this assumption c). As any location within the circular locus has a known probability of being selected in digitizing a particular point, and is equally representative of it, there are a multitude of line segments that can be generated to connect adjacent points. V. 2a shows a set of 100 such sample points representing possible termini of a hypothetical line segment.

If one were to connect random pairs of such endpoints, some resulting segments would lie to the left of the median line, others to its right, while others would cross it at some point between its endpoints, as V. 2b illustrates. By tabulating, at regular intervals, the distance from each segment to the median line it represents, one can generate dispersion statistics to characterize displacements along the segment. V 2c illustrates such an analysis of 100 trials: pairs of endpoints were independently drawn from circular normal distributions centered at two locations; squared distances from each segment to the median line were computed at 11 equally-spaced locations along the median line, and a standard deviation was computed at each of these places. The bold concave lines in V. 2c depict one standard deviation of distance from the median line for the experiment. Results of the experiment are shown in V. 2d, in which contours are drawn around the median line at one, two and three standard deviations of error.

Sausages and Snakes

The fact that dispersion is greatest at endpoints and least around the midpoint may seem odd at first. That is, when line segments are digitized under assumptions a, b and c, the displacement error tends to be greatest near the "known" endpoints and least midway between them, where there is no known coordinate! No matter how much care is taken in selecting its endpoints, the segment's centerpoint will prove to be a more reliable location, even though it is more fictional. This result may have application in line generalization, as it points to approaches that smooth lines by inserting vertices at segment midpoints and displacing existing vertices in certain directions. Dutton (1981) describes a line enhancement algorithm which does this; Buttenfield (1985), discusses this approach in the context of line generalization. The ability to construct probability regions (c.f. V. 24 around line segments may have a number of graphic and analytic applications, some of which are explored below.

This model in part derives from that of Perkal (1966), with extensions by Chrisman (1983), in which the center of an error locus of radius *epsilon* is rolled along a cartographic feature, sweeping out a sausage-shaped band (see V. 3). In its full fuzziness, (the feature is considered to occupy all locations within the band. In accordance with the simulation described above, one may generalize Perkal's Sausage by allowing the size of the locus of error to vary at each known point along the feature, and shrinking it as the locus nears the middle of segments. The locus of error for the feature will thus resemble a snake that has just eaten a chain of lumpy beads more than it does a sausage. V. 3 illustrates Perkal's method and compares it with that of the author. While their motivations and approaches are similar, the two methods have some distinctive properties. A good review of literature on errors in spatial databases, see Veregin (1989).

Probability Contours and their Application

One might think that providing an error term for every coordinate in a database is overkill. For 2D coordinates, doing this could expand storage requirements by fifty percent, and would complicate the data structures needed to communicate spatial objects. Still, this capability may be worth providing, even should it rarely be used. Even if error terms are but crude estimates, they provide some intelligence where none existed before. This information can provide a parametric basis for procedures that merge and rescale entities in spatial databases under conditions of uncertainty (i.e., normal conditions). While it is unlikely that the size of error loci will vary among the points that constitute a feature (and thus need not be stored for each of them), being able to handle such variability is useful. Consider that in thematic overlay operations, new features are constructed from portions of two or more original sets of features; if the parent features had different error loci specified for them, all the derived features will have mixed loci, information which should be preserved and made use of in analysis and display. The following paragraphs explore some approaches to using positional error information in normalized form.

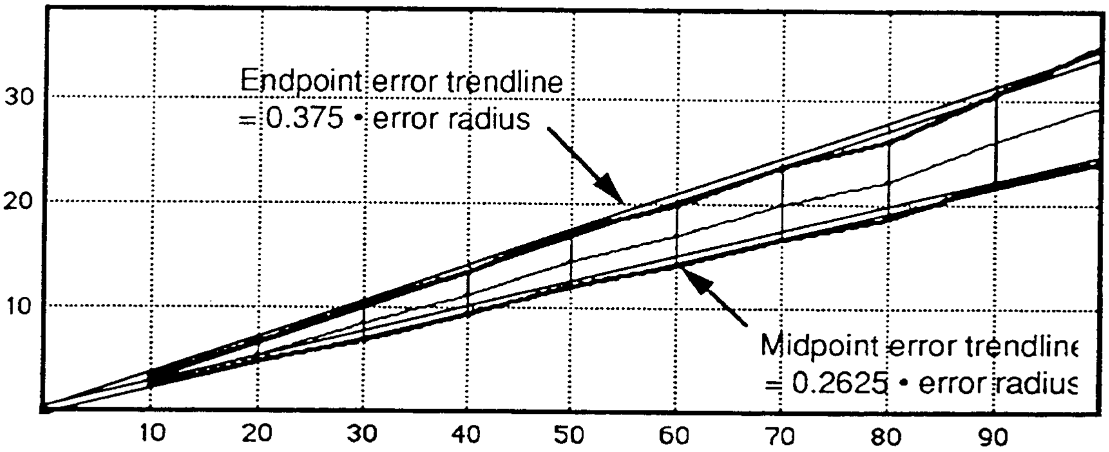
As Visualization 2 shows, a model derived from assumptions a, b and c above can be used to generate contours of probability around each line segment. V. 3 and V. 4 show what this model looks like when applied to linear and areal features. At any vertex or at any point along a segment one may compute the width of a band representing a some number of normalized units of error, and thus the area of the band can be computed by integration or approximation. If one wishes, for example, to know where a line might lie with 95% confidence, then the 2 standard deviation contour can be traced. The more confident one wishes to be, the broader the figure becomes, according to the (one-tailed) area under the normal curve. That is, bands grow quickly at low levels of probability, but as the probability approaches 100%, little growth in area occurs. If one is overlaying two features, one can then choose the probability with which to evaluate near-miss intersections. This probability may be selected according to the scale of either or both features, it can be based on what each feature represents in the real world, or on other contextual criteria. The smaller the probability, the more sliver polygons are likely to be generated.

The area of a feature's error band can also be used to determine whether to display it at a given scale. If the feature is a polygon, half of the area of its error band may be subtracted from its nominal area, and this adjusted area compared with a minimum area criterion (which may be a function of scale, feature class or more complex factors, such as feature density). If the adjusted area is smaller than the criterion, the polygon can be caused to vanish. One way to do this is to contract it to its centroid, which becomes a new node for features adjacent to the deleted polygon. V. 4 shows what a shrunken "probability polygon" looks like, although it isn't really a polygon.

For operations such as the above, there is no apparent necessity to generate error contours connect them into a polygon and measure its area as V. 4 illustrates. Shortcuts and heuristics can be developed that can approximate it from the segment and error data. Approximation is appropriate because the error terms for coordinates are likely to be estimates - if not proxies - themselves. The details of the shapes of error loci can probably be disregarded. This paper has stressed their shape in order to communicate the model that underlies the probability filtering approach. However, if one wishes to visualize positional accuracy across a digitized map coverage, it may be worthwhile to generate and shade probability contours to make a realiabity diagram for feature boundaries. Existing algorithms for casing or buffering lines can be applied to producing such renderings, or one that is more faithful to the concave shapes of contours could be developed. The author has developed an approximate method using a HyperCard testbed.

The experiment illustrated in V. 2 has been performed many times by the author. A line of nominal length of 256 (pixels) was simulated with trials of 25, 50, 75, 100, 125, 150, 200 and 256 drawings of random points to represent it, and repeated with error radii of 10, 20, 30, ... 100 pixels at each end. The results of these trials were tabulated and weighted by sample size to determine the relationship between the error locus radius and the width of the one standard deviation probability contour at 11 locations along each set of lines. The graph in Figure I displays the empirical relationship found between error radius and the minimum, maximum and median width of this contour. The minimum radius is always found near the midpoint of the segment, the maximum near one or both endpoints (almost all probability contours generated were concave).

Figure 1



Minimum, Maximum and Median Radius of One Standard Deviation Error Contour as Function of Radius of Locus of Point Uncertainty

These relationships are linear functions, in which the minimum (midpoint) Standard deviation of error is equal to 0.2625 times the endpoint error radius, and the maximum (endpoint) error radius equals 0.375 times this radius. That is, 0.2625 is the approximate slope of the lowest line in F. 1, and 0.375 is that of the uppermost line (the intercepts are by definition zero). These parameters can be used to estimate the location of error contours based on the radius of uncertainty of the coordinates of a feature at segment endpoints, midpoints or anywhere in between (the standard error at the midpoint is 0.7 as large as the average of the endpoint errors). Note that this empirical relationship is between equal units of distance, the metric of which has no bearing on the slope parameters or their ratio. This allows the model to be easily applied to any set of planar coordinates at any scale.

Epilog

No branch of science can bear fruit unless its findings can be qualified by the various uncertainties to which measurement and analysis of its data are subject. Only recently have the consequences of error propagation for GIS data integration and analysis been explored in detail. Many problems in handling spatial data are in some degree attributable to imprecision and inaccuracy in describing where things are on the surface of the Earth. A significant number of such errors may derive from how spatial phenomena are modeled in digital databases. In particular, the construct of coordinate tuples, relied upon to encode the geometry of points, lines and areas, may be at the heart of many difficulties. Uncertainties inherent in this approach complicate both analytic operations - such as feature identification and spatial overlay - and cartographic feature generalization algorithms. It is tempting - and clearly necessary - to respond to the challenges these impediments present by refining algorithms, and interesting approaches are being made by a number of researchers. But in addition to inventing ways to visualize error in spatial data, perhaps we should step back to consider how uncertainty is built into our paradigms of digital cartography, and what kinds of problems this may cause.

Using "fuzzy tolerances" when integrating coverages can help to overcome discrepancies in digitizing map features, but may not prevent the formation of (often foursided) blobs and slivers, spurious artifacts of the procedure. The fact that algorithms have grown better at handling geometric inconsistencies is encouraging, but should not blind us to the fact that the problem is recurring, and owes much to the convention of encoding features using strings of coordinate tuples. That is, despite the information in a topological database about what things exist in various universes (layers), it has little to offer when one attempts to relate things that exist in different universes. It may be the case that many difficulties encountered in maintaining a complete feature topology (even at a single graphic scale) derive from data models which behave as if features owned coordinates, rather than acting as though they share locations on their planet. This distinction, while fundamental, has yet to be made operational in a commercial GIS; the time may have come to shift paraDIMEs.

While approaches such as described in this paper can mitigate some of the errors arising from imprecisely-defined coordinates by providing a probablistic basis for integrating digitized map features, they ultimately fail to provide a suitably general model of spatial location that connects objects in the real world to representations of them in digital databases. Slivers and blobs will continue to appear and will have to be painstakingly dealt with as long as spatial databases represent features on maps rather than the entities that maps portray.

References

Buttenfield, B. (1985). Treatment of the Cartographic Line. Cartographica, vol. 22:2, p. 1-26.

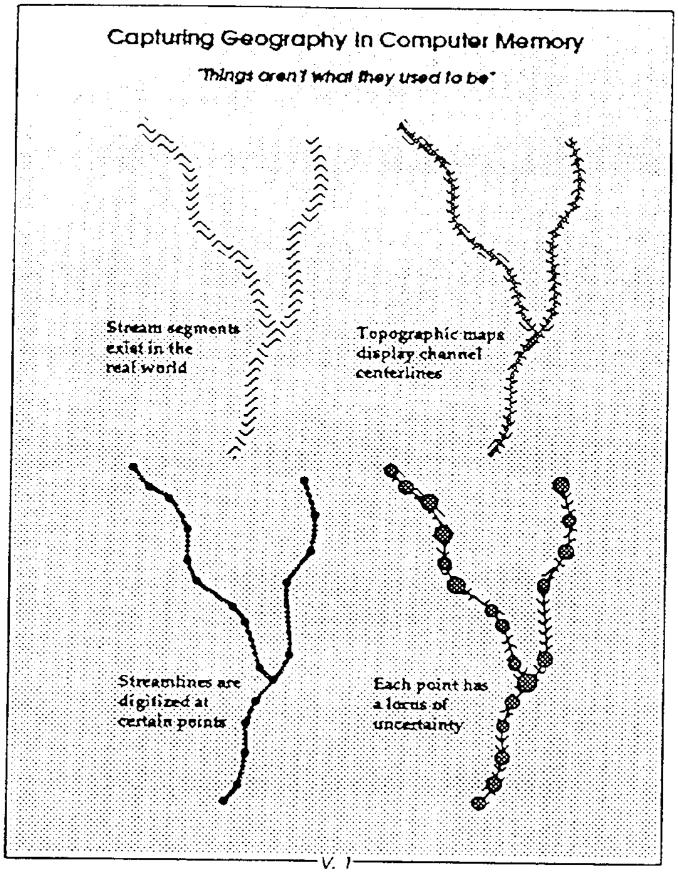
Chrisman, N.R. (1983). Epsilon Filtering: A technique for automated scale changing. Proc. 43 rd A nn. Meeting ACSM, p. 322-331.

Dutton, G. (1981). Fractal enhancement of cartographic line detail. American Cartographer, vol. 8:1, p. 23-40.

Perkal, J. (1966). On the length of empirical curves. Michigan Inter-University Community of Mathematical Geographers, Discussion Paper #10.

Peucker, T.K. and Chrisman, N.R. (1975). Cartographic Data Structures. American Cartographer, vol 2, p. 56-69.

Veregin, H. (1989). A taxonomy of error in spatial databases. NCGIA Technical Paper 89-12, 115 p.



Capturing Geography in Computer Memory

"Things aren't what they used to be"

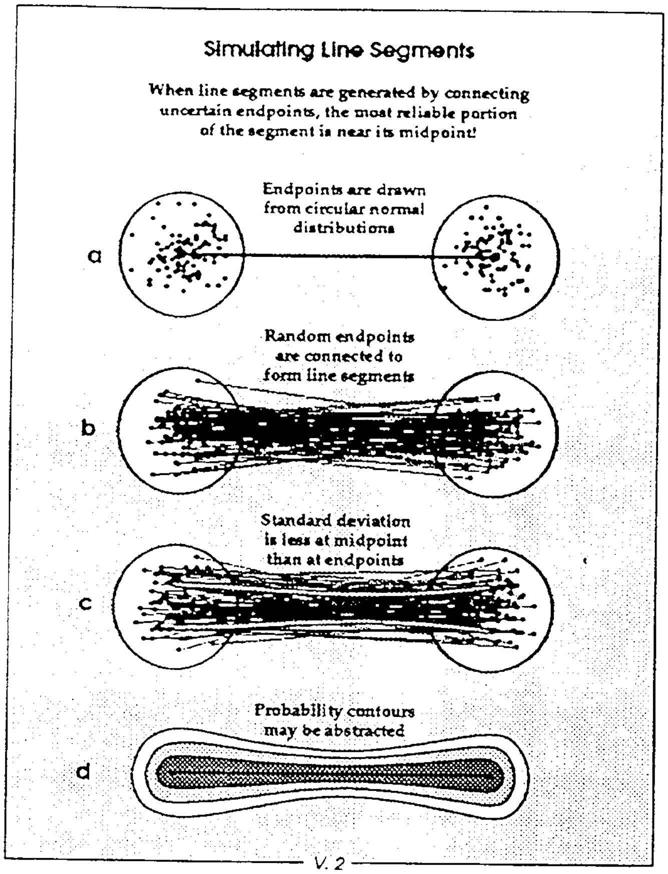
Stream segments exist in the real world

Topographic maps display channel centerlines

Streamlines are digitized at certain points

Each point has a locus of uncertainty

V. 1

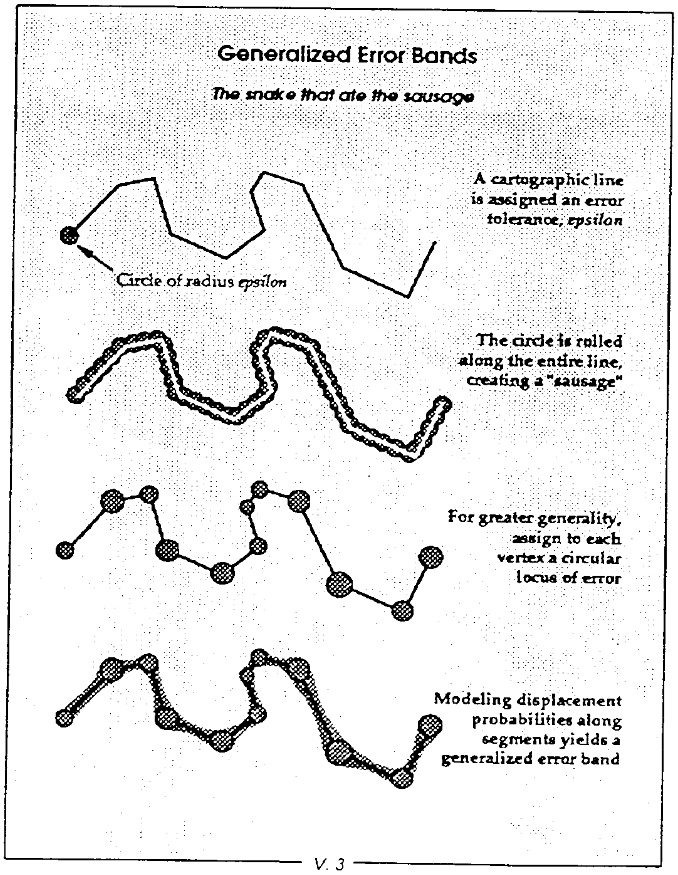


Simulating Line Segments

When line segments are generated by connecting uncertain endpoints, the most reliable portion of the segment is near its midpoint!

a
Endpoints are drawn from circular normal distributions
b
Random endpoints are connected to form line segments
c
Standard deviation is less at midpoint than at endpoints
d
Probability contours may be abstracted

V. 2



Generalized Error Bands

The snake that ate the sausage

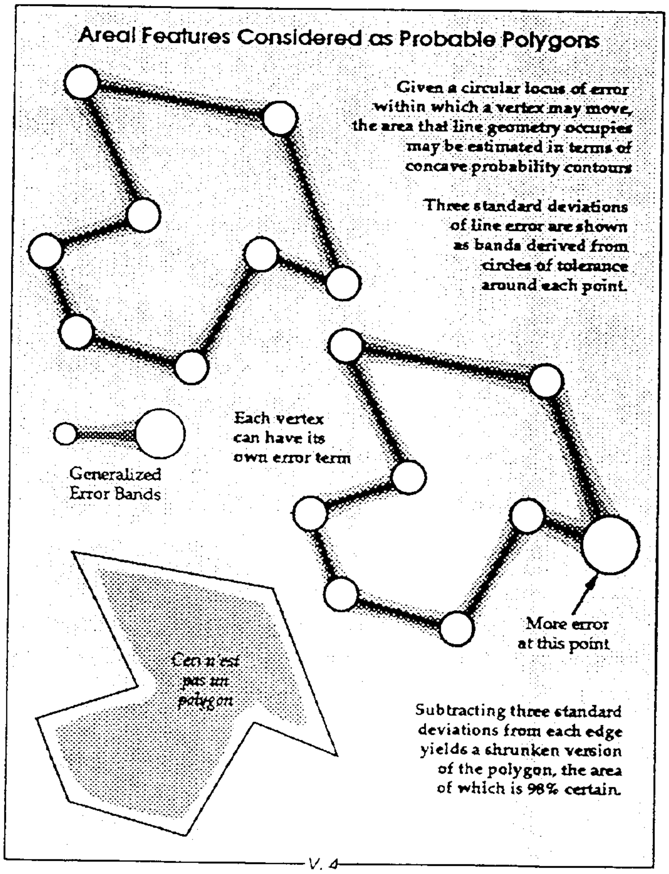
A cartographic line is assigned an error tolerance, *epsilon*

The circle is rolled along the entire line, creating a "sausage"

For greater generality, assign to each vertex a circular locus of error

Modeling displacement probabilities along segments yields a generalized error band

V. 3



Areal Features Considered as Probable Polygons

Given a circular locus of error within which a vertex may move, the area that line geometry occupies may be estimated in terms of concave probability contours

Three standard deviations of line error are shown as bands derived from circles of tolerance around each point.

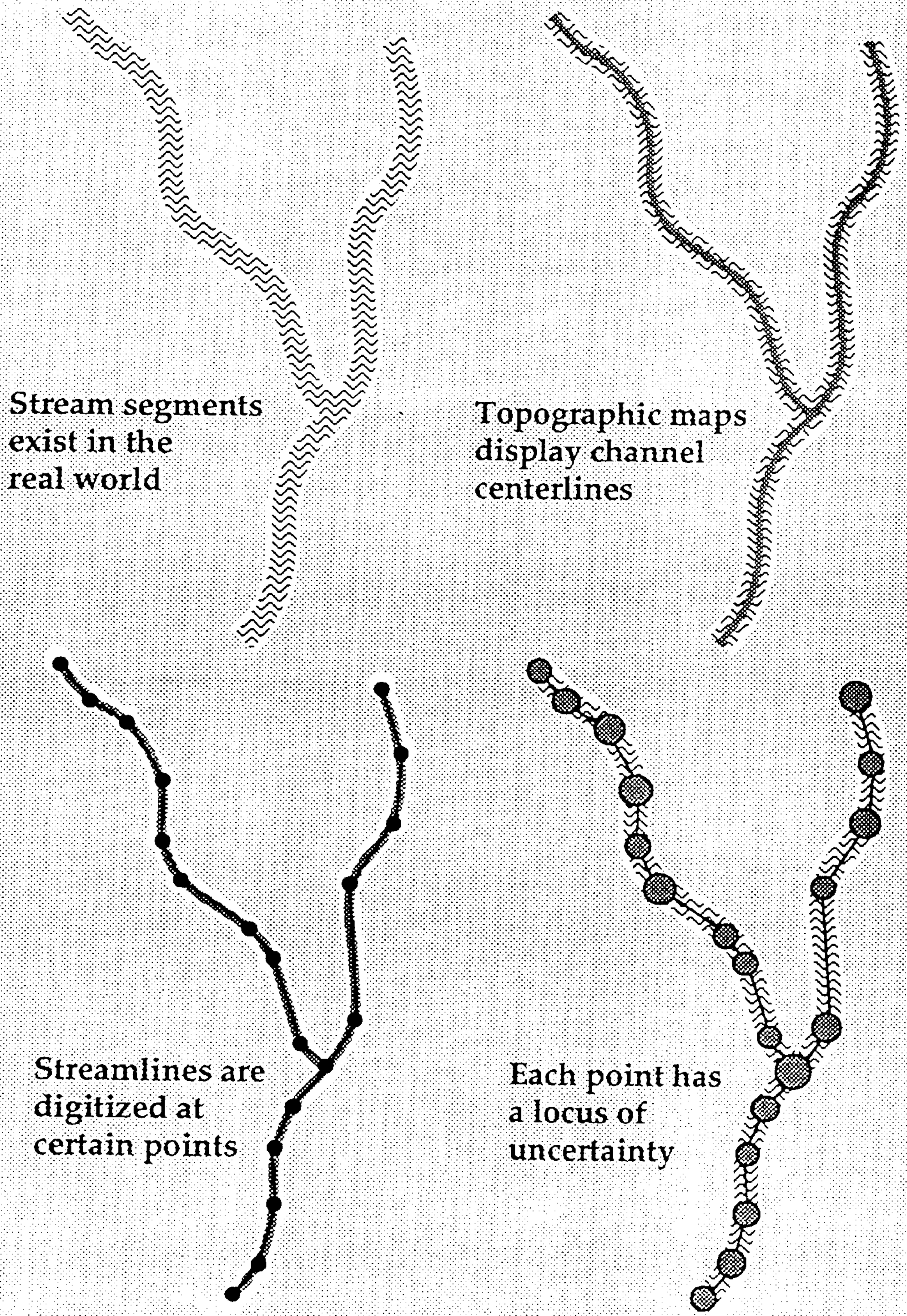
Generalized Error Bands
Each vertex can have its own error term
More error at this point
Ceci n'est pas un polygon

Subtracting three standard deviations from each edge yields a shrunken version of the polygon, the area of which is 98% certain.

V. 4

Capturing Geography in Computer Memory

"Things aren't what they used to be"



Stream segments exist in the real world

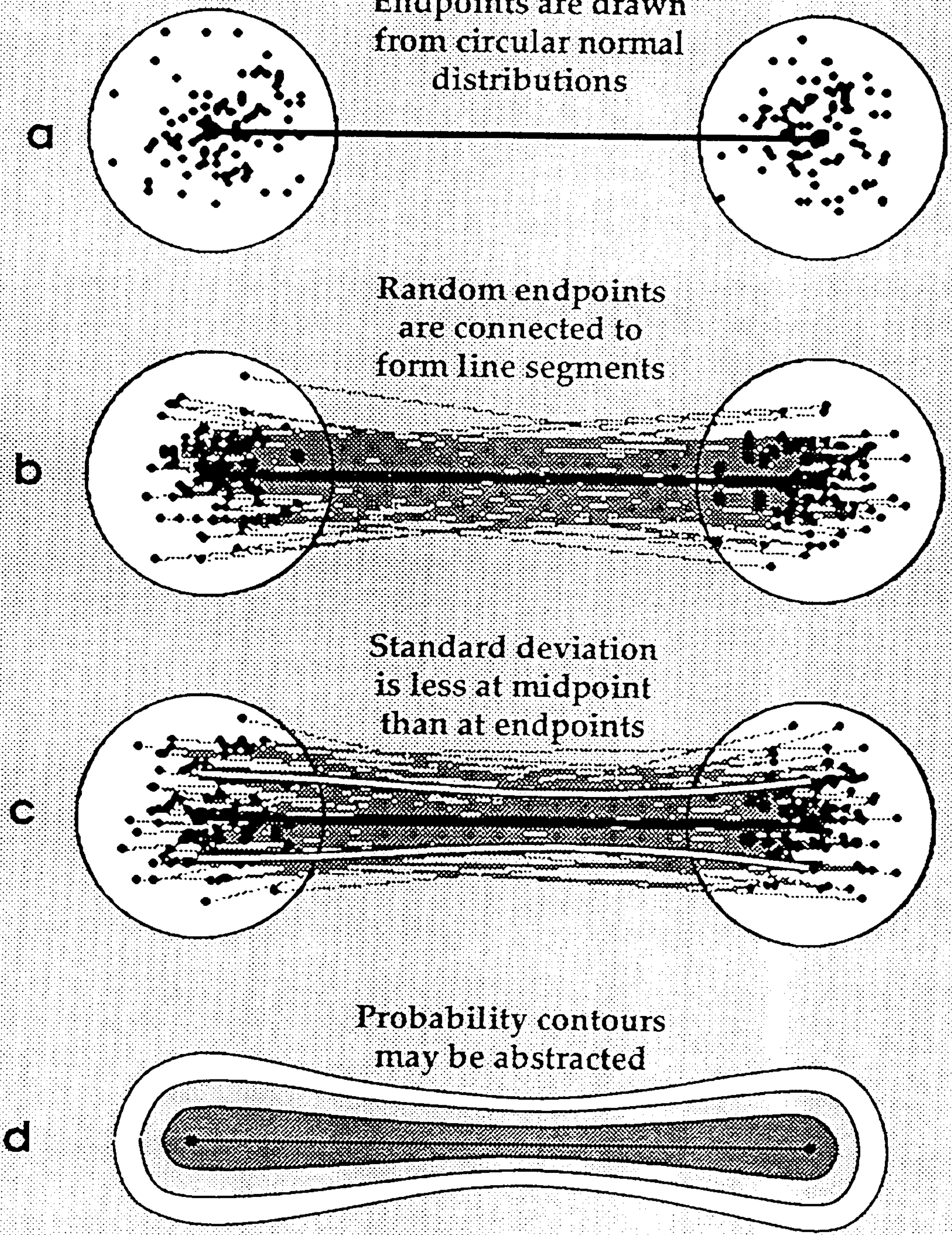
Topographic maps display channel centerlines

Streamlines are digitized at certain points

Each point has a locus of uncertainty

Simulating Line Segments

When line segments are generated by connecting uncertain endpoints, the most reliable portion of the segment is near its midpoint!



Endpoints are drawn from circular normal distributions

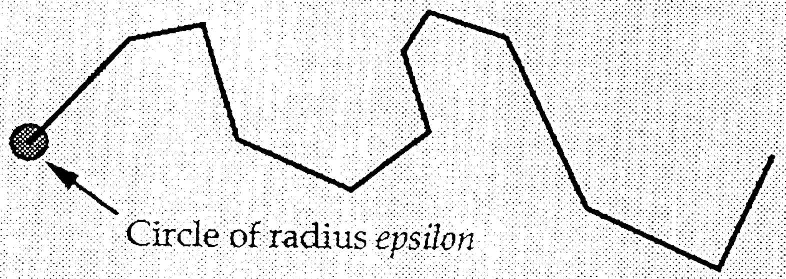
Random endpoints are connected to form line segments

Standard deviation is less at midpoint than at endpoints

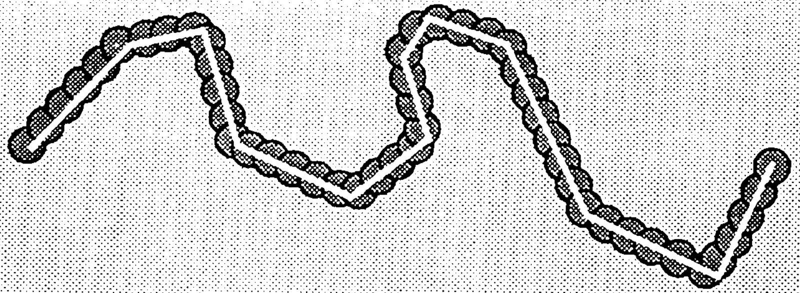
Probability contours may be abstracted

Generalized Error Bands

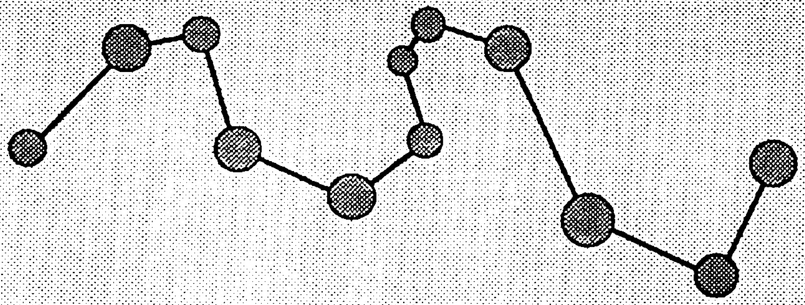
The snake that ate the sausage



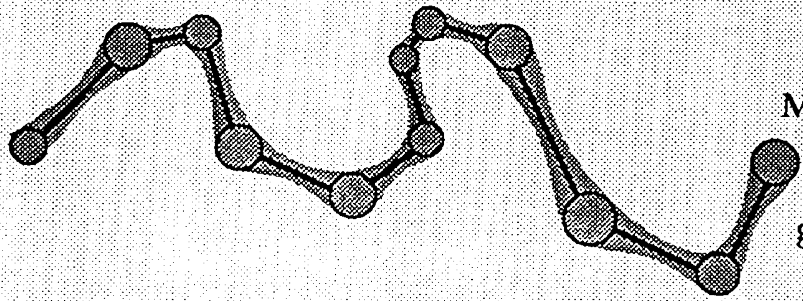
A cartographic line is assigned an error tolerance, *epsilon*



The circle is rolled along the entire line, creating a "sausage"

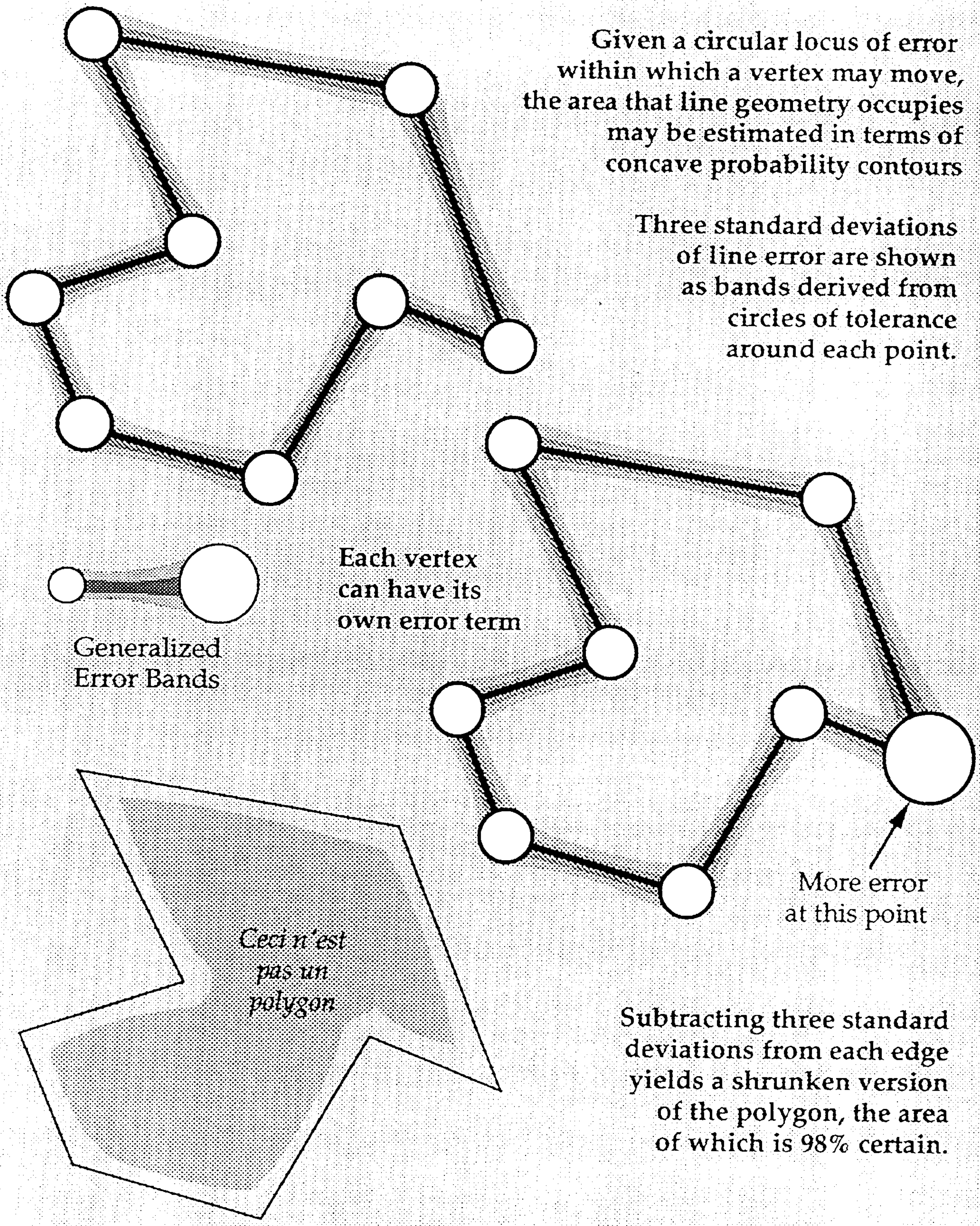


For greater generality, assign to each vertex a circular locus of error



Modeling displacement probabilities along segments yields a generalized error band

Areal Features Considered as Probable Polygons



Given a circular locus of error within which a vertex may move, the area that line geometry occupies may be estimated in terms of concave probability contours

Three standard deviations of line error are shown as bands derived from circles of tolerance around each point.

Each vertex can have its own error term

Subtracting three standard deviations from each edge yields a shrunken version of the polygon, the area of which is 98% certain.

MODELING AND VISUALIZING UNCERTAINTY IN GEOGRAPHIC DATA

Position Paper for the Specialist's Meeting of Research
Initiative Number 7, National Center for Geographic
Information and Analysis, Castine, ME, June 8-12 1991.

Peter. F. Fisher
Department of Geogaphy
Kent State University

THE PROBLEM.

At the root of the problem of visualizing geographic data is the fact that in neither a theoretical nor a practical sense is either the location or the attribute of any phenomena certain. The extent of this problem is only recently beginning to be understood, but it is endemic in geographic data, as is argued below for a number of different types of geographic data. This being the case, the method of presenting that geographic data is inherently concerned with how to present uncertainty. Historically, the tendency has been to present the dominant taxon, or the most important class as occurring at a location or across an area. This denies uncertainty, and implies a level of definition that often cannot be matched in the field, and is unsubstantiated by detailed, or even cursory, examination of the phenomenon mapped.

There are a number of methods for presenting accuracy information in spatial databases, but they are inherently a-spatial. Thus the Root Mean Squared Error (RMSE) for USGS DEM's or the National Map Accuracy Standards, or the Soil Map Accuracy specifications, etc. are all lacking any spatial information. Similarly, the misclassification matrix of remote sensing fame (whatever name it is known by) is without any spatial connotation. That is not to say that these measures could not be visualized, they can, but it is inherently more appropriate to construct that visualization around the distribution of uncertainty, and to develop methods to display the distribution.

On the whole this has not been attempted in traditional cartography, although outstanding examples can be quoted, but, with increasingly powerful tools for scientific visualization, the desire to display the distribution of uncertainty is increased. In spite of this desire, however, it is not certain that enough is known of the structure of uncertainty in most traditional geographic information to make it possible to map it. This position paper presents arguments relating to recent research on the nature of the error in geographic information derived from remote sensing, maps, and other sources.

THE NATURE OF UNCERTAINTY IN GEOGRAPHIC DATA

Land Cover Classification

Consider a land cover classification from a Landsat image. The normal practice is to condense the multiple spectral bands into a single image of land covers, mimicing the traditional class-area map of land use. Among the most common classification methods used, however, there is a statistical likelihood of ever land cover type being in every pixel.

In the Parallelepiped classifier the chance is 0 or 1, but, in the Minimum-Distance-to-theMean, or the Maximum Likelihood procedures, the likelihood (distance or probability respectively) is assessed for all cover types. A simple decision rule is usually implemented which places a pixel in the closest, or most likely cover type. Another cover type may be only a little further away or less likely, but that information is completely lost in the usual approach. Similarly, an important but small feature, such as a building may be dismissed because it is sub-pixel. Fisher and Pathirana (1990), among others, have presented fuzzy mathematical approaches to defining the likelihoods of all cover types, and retaining all those likelihoods. Gong and Howarth (1990) have explored a similar approach using the Maximum Likelihood method. The important point in both these methods is that the result is a new, derivative image for each cover type which records the likelihood of each pixel belonging to the particular cover type. With coarse resolution sensor information (AVHRR, MSS) this is obviously an acceptable approach, and the meaning of the membership values in the pixels has been demonstrated by Fisher and Pathirma (1990). An argument can be put forward that the approach is valid no matter the resolution of the sensor, since as sensor resolution decreases attempts are made to detect increasingly small objects, and, furthermore, the statistical uncertainty of the system is worth recording.

Uncertainty in Soil Maps

In the report which accompanies a soil map, it is commonly acknowledged that within each polygon delimited on the map, a number of different soil types may actually be present. Thus it is true to say that, in a theoretical/statistical sense at least, any location has a particular likelihood of any of the soils mapped in the map-area in general being present, and possibly even some that are not actually mapped. Certainly almost all soils have a likelihood of occurring in locations other than those they are depicted as being in, and where they are delimited, there is a finite likelihood that they are not present. The best way to record this information is again to develop a model of likelihood for each soil type over the whole area. i.e. in an area where 20 soil types occur, there are actually 20 different maps of the likelihood of the occurrence of soil types (one for each soil) (Fisher 1989). Maps of uncertainty (probability of being correct) are an integral part of the Kriging algorithms widely used in goestatistics and in mapping of single soil parameters, deriving two spatial databases (a contour map and a probability correct map) for every parameter (Webster and Oliver 199 1).

Uncertainty in the Built Environment

Features in the built environment are commonly more certainly present. Thus many people feel that they are certainly within a building. But when do they enter it? At the storm door, or the main door, or when the get onto the porch? There is a zone of uncertainty around any building or other object, as to where that feature begins and ends. Of course, certainty is improved if the porch is defined as a separate entity from the house, or as a constituent object of the house, but at what point does the usual geographic database stop breaking down objects.

Subjective definition of built objects is inherently poor, and clearly again there is a case for recognition of the uncertainty, depending on the perspective of the human. Objects are definable, but their extent is less clear. While it would be extreme to state that any location in the city of Kent, Ohio or even the US, could be perceived as being the Fishers', there is obviously a zone that is the greater Fishers' residence. If this is extended to the concept of nearness, then there is a still greater case for saying that depending on the human using the term, any location may or may not be near to some other location, and that one human's definition of near is different from another's (Robinson 1990; Fisher and Orf 1991).

MODELING THE UNCERTAINTY

Models in Remote Sensing

Statistical models of uncertainty are available. Most common among these is the maximum likelihood classifier used in remote sensing, but this method is commonly used to absorb uncertainty in the classification process to enable an optimal derivation of an image of Boolean cover types, not to allow the user to overtly manipulate the uncertainty. Indeed, a few years ago the current writer made exhaustive inquiries among several leaders in the field of image processing of satellite data, both vendors and academics, and none could quote a single system that saved the full probability fields in the maximum likelihood classification process. One system (ERDAS) has an optional saving of a single probability layer which records the probability values associated with the most likely cover type in each pixel (a product which is very little improvement over the Boolean image). Derivation of the uncertain images has only been attempted in the research domain, because of the large amounts of data storage required, but those methods do present a method for deriving uncertain spatial information.

Many other methods in the realm of image processing may also be viewed as measures of uncertainty. Thus an edge detection filter passed over an image derives a measure of the possibility of an edge being present, and a vegetation index map can be viewshed as an image of the likelihood of high biomass being present.

Models in DEMS

The error component in spatial data bases is reported. Error in USGS Digital Elevation Models from 1:24,000 maps, for example, is reported as a Root Mean Squared Error (RMSE), while the DEMs from 1:250,000 maps are reported as conforming to a vertical and horizontal (circular) accuracy standard. This information has been used by Fisher (1991) in modeling the uncertainty of the viewshed, one of the derivative products of the DEM. The Fuzzy Viewsheds defined are more accurate representations of the world experience of the human beings, and provide a more useful planning tool than do the traditional Boolean Viewsheds. An error simulation model has been used in this research to date, but other algorithmic approaches are under development. Most interesting is the ability to use Fuzzy set theory to combine the fuzzy viewsheds, and the possibilities inherent in the alternative fuzzy products.

Models for Soil Maps

Similarly for traditional soil maps it is actually possible to develop statistical models based on the soil types that occur within particular mapping units, and the Soil Conservation Service mapping standards, or error rates in the accompanying reports (Fisher in press).

VISUALIZING THE UNCERTAINTY

Fuzzy Land Cover

Some of the uncertain spatial products mentioned above are inherently displayable, using conventional imaging methods. Thus the fuzzy membership of land cover type 'A' being within present within a cell is a good candidate for a grey scale image. Darker areas report high memberships, while paler areas distinguish lower values. The images can then be overlaid in threes to give three color composites, where areas in white would have high memberships in all cover types and those shown in black would have low memberships in all cover types. This is using conventional image processing techniques, and three color composite display, but the number of possible 3 color composites is enormous, even for a dataset with 4 cover types (24 possible if we include all possible 3 color combinations of the 4 cover types). The user of land cover information, however, expects to be presented with all land cover types in a single image. Are we therefore to develop a system with as many primary colors as there are cover types in an area? Theoretically this would allow the viewer to see all cover types, and the uncertainty of each at once. The information content of the image smacks of overload, however, and who but the most experienced user (probably only the people who developed it) would understand the image if it was made? The same problem exists in display of uncertainty in soil maps, using the data model suggested above.

It would be possible to display the land cover map with a background measure of the separation from the most likely cover type and next most likely. Large separations indicate purer pixels, but this is reporting a small part of the information in the error database.

Simulations

The most interesting alternative to the present writer is develop a simulation approach to display, where the likelihood of a pixel being displayed as a particular cover type is deter-mined by a combination of its likelihood of being that cover type and a random process. This would cause the display to be different each time the image was displayed. Indeed, animation techniques could be used to change the display whilst it was on the screen, according to the same selection process. Thhis would convince the user of uncertainty in the data, but I believe that many users would find it unacceptable, since they inherently believe that the data is correct The same approach could be used in the display of uncertainty in DEM data, but the image would actually give the impression of a moving surface, not a desirable feature of the land surface, unless an earthquake is in progress.

SUMMARY

To date very little has been done with quality of spatial information, although there is an increasing literature on this important subject. Without the research on the nature of the uncertainty, its distribution, and development of models of its occurrence, it is very difficult to develop methods for visualizing that uncertainty. In some areas, it is possible to make suggestions and ask questions as to possible lines of development, but some of these verge on the ridiculous, and may have very low acceptability among a user community with unpredictable perceptual problems. There is a very real need to develop these methods, however, but they rely on a desire to study and model spatial error, and a user community ready for some innovative displays. Unfortunately the time tested paradigm of the map mitigates against display of quality. 'Me major consideration in this area of research is to establish the human perception of whatever methods are developed.

REFERENCES

Fisher, P.F. (1989) Knowledge-based approaches to determining and correcting areas of unreliability in geographic databases. In M.Goodchild and S.Gopal (eds.) Accuracy of Spatial Databases, pp. 45-54.

Fisher, P.F. (1991) Simulation of the uncertainty of a viewshed. Proceedings Auto Carto 10, pp. 205-218.

Fisher, P.F. (in press) Modelling soil map-unit inclusions by Monte Carlo simulation. International Journal of Geographical Information Systems.

Fisher, P.F. and Orf, T.R. (1991) An investigation of the meaning of near and close on a university campus. Computers, Environment and Urban Systems, 15, 23-35.

Fisher, P.F. and Pathirana, S. (1990) The evaluation of fuzzy membership of land cover classes in the suburban zone. Remote Sensing of Environment, 34, 121-132.

Gong, P. and Howarth, P.J. (1990) The impreciseness in landcover mapping, its determination, representation and application. IGARSS'90, pp. 929-932.

Robinson, V.B. (1990) Interacfive machine acquisition of a fuzzy spatial relation. Computers & Geosciences, 16, 857-872.

Toward a Theory for the Visualization of Data Quality in Cartographic Display[1]

Position paper for Initiative 7 "Visualization of Data Quality"

Andrew U. Frank
Department of Surveying Engineering
National Center for Geographic Information and Analysis
University of Maine
Orono ME 04469
(207) 581-2174 (Fax: (207) 581-2206)
FRANKC@MECAN1.bitnet

1. Introduction

The visualization of uncertainty in geographic data is a very complex problem. In this position paper, a mathematical framework for the design and assessment of visual communication of data quality is proposed and a research direction sketched.

2. Communication of data quality in paper maps

The problem of communicating uncertainty has not been an overwhelming consideration in cartography, because the medium communicated an implied accuracy standard (e.g. National Map Accuracy Standard). Cartographers are supposed to select scale justified by the quality of the data available, and map users assume that the data are as good as shown (but refrain from trying to measure from a paper map distances with more precision than say 1/10 mm). The graphical medium enforces this constraint. Distributing data sets in computer readable form and the technical means to show them with any scale desired voids that tacit assumption and opens the door for misunderstandings between map producer and user. It thus becomes necessary to communicate data quality directly.

3. Communication of data quality in a GIS display

The question to be addressed is how to select a method to communicate data quality and specifically how to assess different options. Goodchild and Schweitzer have tested, for the case of positional uncertainty in point location, a number of possibilities and found that the human visual system understands some of them as uncertainty information, but that most are given another interpretation. From this initial work, a full fledged systematic program can be developed, constructing systematically all technically feasible displays for uncertainty and then testing them with human subjects to see if they are effective. It is very obvious that such a program would be a large and difficult enterprise, with a large number of -variables to be controlled. In addition, it was found in Human-Computer Interface studies, that detail studies do not combine well and often are not good predictors for behavior in complex environments (probably due to interaction between components).

An alternative to such an empirical approach is based on the mathematical concepts of algebra and morphism to assess the potential for communication of a specific message through a channel. It is based on the following observations:

- A class of objects, e.g. weight measures or uncertainty measures in the position or the value of a feature, is characterized by the operations that one can perform with them. For example, uncertainty in position is different from a physical quantity like weight, because addition of uncertainty does not follow the same rules as addition of weight. The operations and their behavior can be described as an algebra and the differences formalized.

- Visual variables in a graphic display, e.g., a map, can also be characterized by the operations applicable to them. Symbol size can be gradually increased and symbols can be compared for size, whereas symbol type is (usually) an ordinal scale variable and no order relation is available.

- Communication of meaning is achieved by a correspondence between the behavior of the conceptual and the visual object. This means the same operations with the same properties should be available. For example, it is not a good practice to use the symbols A, B, and C to show towns which are located at low, medium, and high altitudes on a map. The characters as symbols do not visually convey the ordering.

[1] Funding from NSF for the NCGIA under grant SES 88-10917 and from Intergraph Corporation is gratefully acknowledged.

- If both conceptual and visual objects are described as an algebra, the correspondence should be a morphism. Morphisms are structure-preserving mappings between algebras. Given two algebras (A, *) and (B, +). A mapping f is a homomorphism if it maps objects from A to B, f (a1) -> (b1), and maps the operation * on A's to the operation + on B's such that f (a * b) = f (a) + f (b) (as an example recall the rule that the logarithm of the product of two positive numbers is the sum of the logarithm of the two numbers log (a * b) = log a + log b).
The human visual system has a strong tendency to interpret pictures in terms of the familiar environment. For example, there is a preference to see three dimensional objects, even if the stimulus presented is purely two dimensional (Nekar cubes). One could read the above assumptions of a detailed statements of this general rule, saying that the behavior of the familiar environment determines our understanding of pictures.

4. Communication of data quality using visual variables with similar algebraic structure

Applying these assumptions to the visual communication of data quality, I reach the following conclusions:

- Data quality can only be communicated effectively using visual variables that have a similar behavior than the quality measure to be shown.

- Different aspects of data quality (e.g. precision, completeness, timeliness) each have different behavior. Compare how to 'add' descriptions of each, given two datasets; that are merged.

- Most visual variables cannot communicate data quality measures. The search for candidates for visual variables that can communicate data quality could start from searching our visual environment for situations where uncertain data is visually acquired (e.g. objects in a distance, fog, etc.).

5. Some evidence

A few observations that show how I think these concepts apply:

- The traditional method to communicate data quality, especially uncertainty in position, is selecting the appropriate scale such that one does not read more precision than was initially available in the data. The visual metaphor is 'objects seen from a distance' and one cannot see all detail (and the current problem is due to the potential for 'zoom' in the GIS). Hint for a solution: limit the possibility to change scale ('zoom') when the limit warranted by the data quality is reached (this will not be popular with users).

- Goodchild and Schweitzer found that moving point symbols are perceived as moving objects and do not convey an uncertainty in location (perhaps unless a high enough frequency is reached so the images blur, which is hard to achieve on a CRT).

- The cloud of points seems to work if it is a cloud, not a collection of points. It uses the visual metaphor of 'cloud' or 'fog'.

- Filled circles are seen as a measure of the magnitude of the object. Unfilled circles may work for positional uncertainty, because the distance between two points with uncertain location is - common sense reasoning - between the longest and the shortest distance between any two points in the two circles.

6. A tentative research direction

My proposal for a research plan would be the following:

- Study the algebraic properties of data quality measures, especially how do they compare, add etc.

- Study the algebraic properties of visual variables in order to identify variables that have a similar algebraic structure.

- Explore the natural environment for situations where acquisition of visual data with limited quality applies. Study their algebraic properties.

- Find mappings for the data quality measures to visual variables which preserve the algebraic structure (morphism).

References

These concepts have developed from the results from research initiative 2 (spatial languages) and the NATO workshop held last summer, where metaphors were extensively documented, mostly based on the work of George Lakoff and Mark Johnson. The ’morphism’ concept was detailed in Kuhn, Werner, and Andrew Frank, "A Formalization of Metaphors and Image Schemas in User Interfaces", in Cognitive and Linguistic Aspects of Geographic Space, David Mark and Andrew Frank (eds.), NATO ASI Series, Kluwer Academic Press. In press.

POSITION PAPER FOR 17 SPECIALIST MEETING

Michael F. Goodchild
NCGIA
UC Santa Barbara

SUMMARY

The first section gives a brief summary of work under Initiative 1 on Accuracy of Spatial Databases, as it is relevant to I7. The second section then looks at how statistical error models might be used as the basis for cartographic display of uncertainty. More detailed work under I1 on uncertainty in area class maps is difficult to describe in this position paper, but I will bring some slides and copies of the latest paper on the work to the meeting. The last section comments on the four subthemes of the meeting.

1. Background: Initiative 1

Among other things, Initiative 1 (Accuracy of Spatial Databases) focused on the development of error models for GIS, and the propagation of errors through GIS processes. The case for I1 was based on two assumptions:

- the major source of information for GIS at this point is maps;

- the current generation of spatial databases fail to capture the uncertainty that is inherent in maps, or to propagate it-through GIS processes to the final product.

It was argued that the ideal GIS would not only provide ways of describing error or uncertainty in the datasets stored in its database, but also provide ways of tracking this uncertainty through processing and of reporting appropriately calculated confidence limits on GIS products, such as measures of area.

All geographical data is subject to uncertainty, with the possible exception of geographical facts that have a mathematical basis, such as the latitude of the North Pole. Thus a geographical feature can be regarded as one of a population of possible, distorted versions of the same true feature, just as the Gaussian distribution is used to describe a population of observable values of some scalar measurement. A feasible definition of an error model for geographical data is a stocastic process capable of generating such a population of features. The circular Gaussian distribution provides a widely used error model for the geographical location of a point. However, a substantive conclusion of I1 is that using this definition, it is not possible to write down suitable error models for wide classes of map features. Examples include contour lines, and edges in the boundary networks of area class maps (e.g. soil maps, land use or land cover maps).

Contour and area class maps are examples of the cartographic representation of fields. In the contour case, the relevant variable (topographic elevation) is measured on an interval scale; in the area class map,' the variable is usually discrete. In such cases the objects seen on the map, and coded in the database, are not representations of identifiable, discrete features in the real world but artifacts of a process of discretization. In order to capture an infinite amount of continuous variation, it has been necessary to create discrete objects that approximate the real variable by isolines or areas of approximately constant value. The solution to the problem of error modeling for such objects lies in writing the error model for the field from which the objects were derived, not for the objects themselves. Thus one can model error in a contour map by generating distorted versions of the underlying elevation field. The resultant samples of contours differ not only in their positions but also in their topological properties (numbers of objects, etc.). Much of the effort in I1 went into developing suitable field error models for continuous and discrete variables. Arbia and Hatning formulated a very general autoregressive model for the continuous case, and have been working with a model for the discrete case.

Given suitable error models, it is possible to store information on uncertainty in the database, as attributes of individual themes, objects, regions or points as appropriate. Uncertainty can then be propagated through GIS operations and reported in products. David Lanter has developed a system at UCSB for tracking uncertainty as one attribute of each dataset's lineage. Yang Shiren has implemented the discrete case error model in GRASS and found confidence limits on products using simulation techniques. At Newcastle, Stan Openshaw's group has been adding the epsilon band model to standard GIS operations in ARC/Info. In Peter Burrough's group in Utrecht, Gerard Heuvelink has worked out analytic solutions to the propagation problem using Taylor series expansions, and implemented them with Monte Carlo simulation in a general language compiler for error propagation.

Selected bibliography.

Arbia G, Haining R P (1990) Error propagation through map operations. Unpublished manuscript

Carver S (1991) Adding error handling functionality to the GIS toolkit Proceedings, EGIS 91 EGIS Foundation, Utrecht: 187-196

Goodchild M F, Gopal S (1989) Accuracy of Spatial Databases Taylor and Francis, New York

Goodchild M F, Sun G, Yang S (1991) Development and test of an error model for categorical data. Unpublished manuscript

Heuveliak G B K Burrough P A, Stein A (1989) Propagation of errors in spatial modelling with GIS. International Journal of Geographical Information Systems 3: 303-322

Lanter D (1990) Lineage in GIS: the problem and a solution Technical Paper 90-6 NCGIA, Santa Barbara CA
Veregin H (1989a) Accuracy of spatial databases: annotated bibliography Technical Paper 89-9 NCGIA.. Santa Barbara CA

Veregin H (1989b) A taxonomy of error in spatial databases. Technical Paper 89-12 NCGIA, Santa Barbara CA

2. Display of uncertainty

Why is information on uncertainty so absent from maps (and generally from cartographic representations of geographic data) when uncertainty is so ubiquitous? Why is the information on uncertainty embedded in many map legends not transferred to the digital representation of the map?

Suppose we were to adopt a simple theory of the map, that its design reflects an optimal communication between cartographer and user under the constraints imposed by mapping technology (fixed scale, paper, pen strokes of uniform width etc.). Then the distinct absence of uncertainty in conventional map content suggests two alternative inferences: that communication of uncertainty is incompatible with communication of more important content; or that there is a tacit understanding between cartographer and user that uncertainty should not be communicated. In the first case, it may be that human perception simply cannot handle uncertainty as a dimension orthogonal to all the other information that has to be communicated about an object. Or there may be no available means, within the constraints of the technology, for communicating another dimension.

The move to digital mapping technology has added a series of new dimensions to visual communication. Under this theory of the map, therefore, we might expect the new technology to offer ways of getting around the old barriers that kept uncertainty off traditional maps. The new technology adds:

- time dependence - displays can be animated;
- three dimensions - either by stereo viewing or by simulating the appearance of movable solid objects;
- multiple media - sound, images and text all easily accessible;
- hierarchical access - interactive zooming, and logical linking of representations at different scales;
- continuous spatial transition - techniques for showing continuous change through gradations of color and density.

Any one of these has the potential to expand the bandwidth of communication between cartographic object and observer so that uncertainty can be communicated effectively.

At Santa Barbara, Diane Schweitzer and I have been experimenting over the past month with a number of these options. We have taken what may be the simplest case of all, uncertainty in the location of a point. Here are some preliminary conclusions:

- The eye always interprets a moving object as representing movement in reality, never as representing uncertainty of position.
- If we simulate the uncertainty as a cloud of dots, there appears to be a critical density: above that density, the eye sees a fuzzy object; below it, a swarm. The first case communicates uncertainty but the second does not
- If a set of points are represented by a set of clouds, the points must be generated independently in each. If not, the eye identifies a standard pattern and perceives the clouds as instances of one standard symbol.

- If the circular Gaussian distribution is simulated using a piecewise approximation to the density function, the eye is very clever at detecting concentric rings where the density gradient changes.

- The eye often perceives a simulation under a Gaussian distribution as having two distinct domains - a central one of high, constant density, and a peripheral one with decreasing density. In other words the eye is not good at perceiving a probability density function from a simulated point distribution-

- Where density functions overlap, point densities should approximate the greatest individual density, not the sum-

- Height is broadly acceptable as a representation of probability density. For example, two perpendicular sections of a bell curve can be drawn above a point However this is not reasonable for a dense point set.

- Ile diameter of filled circle is more likely to be interpreted as a measure of the object's magnitude than its positional uncertainty. However the reverse may be true of the diameter of an unfilled circle.

- We are currently experimenting with composite symbols, using a conventional symbol to denote the location and attributes of the object (e.g. through size, weight or color) and an unfilled circle to denote uncertainty in location.

- When errors are positively correlated between points in a set (i.e. large absolute errors but small relative errors) it may be appropriate to blur the frame (other fixed objects in the field of view) rather than the points.

These are very preliminary conclusions, but might form the basis for some interesting, testable hypotheses, or for further discussion at the meeting. of course they deal only with point objects, and only with locational uncertainty.

One widely used convention for uncertainty in the location of a political boundary is a dashed line (e.g. South Yemen and Saudi Arabia). The context is sufficient to avoid confusion with other uses of dashed lines (e.g. ephemeral streams) but the convention gives no impression of level of uncertainty.

3. Uncertainty in area class maps

This has been the main focus of my work under I1 at Santa Barbara over the past two years. It deals with a class of data that is very common in GIS: a discrete field, typified by a soil map or land cover map. I assume that accuracy can be defined as the difference between the actual class at some arbitrary location (xy) and the class returned by the database. A common discretization of this class of data reflects the features depicted on maps, and consists of a partitioning of the space into polygons, each assumed to have some constant value. A cynical view of such maps and databases is that they contain "lines that do not exist surrounding places that have nothing in common".

To model error we re-discretize the space into objects that are more compatible with the concept of field, i.e. raster cells. In each cell we assume the existence of a probability vector {p1,p2,….,pm} defining the probabilities that the cell belongs to each of m classes. These probabilities might be obtained by classifying an image using a fuzzy classifier, or might quantify the uncertainties described in the legend of the map (eg- "A: soils in this class consist of loam with 10% inclusions of clay"). The observed map is assumed to be one realization of a multinomial process defined by these probabilities. In addition, within one realization the outcome in each cell is dependent on neighboring outcomes; across realizations, however, outcomes are independent.

In this model the familiar polygon objects of the map representation are not part of the stochastic process itself, but must be inferred from its outcomes. Thus it is not possible to visualize the uncertainty inherent in the model by modifying the representation of objects such as boundary lines. Moreover, since outcomes are spatially dependent, uncertainty can only be visualized by displaying the variation across realizations. The only feasible approach seems to be to focus on displaying the parameters of the stochastic process itself (the probability vector fields) rather than on its realizations.

4. The proposed subthemes

The comments above relate most strongly to the first (Data Quality Components - error modeling and derivation of indices of data quality) and third (Representational Issues -visual tools to facilitate internal representation and graphical display). I see the third as the key to this research initiative, whereas the first is more like a revisiting of I1. The second subtheme (Data Models and Database Issues - management of data quality within databases during manipulation and update) is very important, but seems to stray away from the Visualization theme of the initiative, unless a good case can be made. Finally the fourth (User Needs - assessment of the tools and algorithms, and analysis of user demands for data quality information) seems to combine two very different issues -the need for a focus on perception, which seems fundamental to 17, and user needs assessment, which seems highly peripheral. Of course I don't want to limit discussion unduly, but I think ifs important to maintain a strong sense of a central focus in a meeting of this nature, if we are to arrive at a tight, feasible and significant research agenda. I hope we can review the four themes and their interpretation and relative importance at the Saturday evening session.

DATA QUALITY AND VISUALIZATION:
A POSITION PAPER

by
Daniel A. Griffith
Geography Department & Interdisciplinary Statistics Program
Syracuse University

1. SCOPE OF THE PROBLEM

Quality of data alludes to various types of errors that can be lurking about in figures. These errors stem from different sources, and may be classified as being numerical, measurement, sampling, specification, and stochastic (see Griffith and Amrhein, 1991). Numerical errors arise from the computer being unable to represent the entire real number line, from rounding/truncating of strings of places to the right of the decimal point, and from arithmetic mistakes (presumably computers have minimized the occurrence of these), to name a few. Measurement errors refer to incorrect numbers (no quantity can be measured with complete certainty), and arise from the use of an inappropriate measure scale, from measuring aggregated rather than individual objects, or from instruments that are imprecise (unacceptably large variation in numbers occurs when the same item is measured repeatedly), invalid (the wrong or a surrogate variable is measured), or inaccurate (systematic bias is present in values). Sampling error refers to the difference between sample statistics and their corresponding population parameter values; this error arises from the use of a subset (hopefully one that has been judiciously selected) rather than the entire parent population. Specification error refers to the use of incorrect assumptions and/or the application of incorrect mathematical formulae or equations. Finally, generally speaking, stochastic error refers to "natural" variation in phenomena; spatial autocorrelation is linked to this type of error.

This paper will discuss issues concerning the interaction between measurement error and sampling error, specifically focusing on the accompanying attribute and locational components of these errors. For the most part, little is known about the nature, degree, or distribution (both spatial and aspatial) of error in a geo-referenced database. This contention is the basic justification underlying research projects concerned with field-checking and ground truthing (as is done with some remotely sensed data), and sensitivity analysis (see the contents of International Journal of Geographical Information Systems). If researchers know what the error is in a given database, then two courses of action may be taken. First, appropriate corrections could be made in order to eliminate it (e. g., adjustments made to raw data from satellite sensors, or data editing to remove coding mistakes). Second, when elimination is not feasible (e. g., the use of sampling to conserve on resources), then the error needs to be managed in such a way that its impact becomes negligible. Principally, though, researchers do not enjoy the luxury of knowing anything about existing error. The issue then becomes one of getting a handle on potential error, in order to manage it and in order to understand how it propagates through analyses.

Visualization should aid researchers in better understanding known error and identifying potential error in geo-referenced databases. In Part this contention is based upon one experience scientists had with pseudo-random numbers. Over the years considerable effort was devoted to identifying "best" generators of strings of pseudo-random numbers. Essentially very many sample sequences were shown to "pass" a battery of statistical tests. More recently, though, when 2-dimensional sequences from these generators were plotted in 3-dimensional space, and this space rotated, repeatedly a very clear systematic pattern was uncovered in the supposed pseudo-random numbers. Without scientific visualization capabilities this problem may never have been recognized. With regard to geo-referenced databases, visualization in terms of a map also will allow the locational component of error to be treated. This approach has a long traditional in geography; for decades, now, geographers have been analyzing maps of regressi6n residuals.

2. BACKGROUND

Two classes of problem are of interest here. In the first, at least something is known about error; in the second, only educated suspicion about the presence of error is available. Georeferenced socio-economic data provided by government agencies often are accompanied by a description of the affiliated sampling error. Measurement error contained in remotely sensed data is understood well enough to be handled, at least partially, with filtering techniques. But most databases can have nothing more than diagnostics applied to them. In this context, where measurement and sampling errors interact, researchers may wish to exploit such statistical notions as outlier, influential point, and leverage point. For geo-referenced databases, these diagnostics need to be transformed into their spatial statistical counterparts. A useful visualization of such indices might be achieved by creation and display of a frequency distribution, in which global extremes are highlighted, and a map, in which locally extreme values are highlighted.

2.1. The spatial sampling error problem

Anselin has devoted considerable effort over recent years to the notion of heterogeneity of spatial data, and has found it to be markedly important in spatial data analysis. This result is a natural outcome of the sampling schemes used for socio-economic data collected by government agencies.

Let P be the number of items located in a geographic landscape, with p_i being the number situated in areal unit i ($P = \sum_{i=1}^{n} p_i$). Suppose that a random sample of size r is drawn from P (with replacement, order important), with areal unit samples being of size r_i ($r = \sum_{i=1}^{n} r_i$), and the data tabulated by the n existing areal units. Then

Case I: if $p_j = p = P/n \in I^+$, i = 1, 2, ..., n (a uniform distribution of the items over the areal units), and $r_j = r.$ (a constant), then the central limit theorem

implies that $E(\bar{x}_j) = \mu$, $\sigma_{\bar{x}_j} = \sigma/(r.)^{1/2}$, and the sampling distribution for each areal unit is exactly the same (there are $p^{r.}$ possible samples);

Case II: if $p_j = p = P/n \in I^+$, i = 1, 2, ..., n (a uniform distribution of the items), and $r_j \neq r.$, then the central limit theorem implies that $E(\bar{x}_j) = \mu$, $\sigma_{\bar{x}_j} = \sigma/(r_j)^{1/2}$, and the sampling distribution for each areal unit may well be noticeably different (there are p^{r_j} possible samples for areal unit j); and,

Case III: if $p_j \neq P/n \in I^+$, i = 1, 2, ..., n, and $r_j \neq r.$, then the central limit theorem implies that $E(\bar{x}_j) = \mu$, $\sigma_{\bar{x}_j} = \sigma/(r_j)^{1/2}$, and the sampling distribution for each areal unit is capable of differing even more dramatically (there are $p_j^{r_j}$ possible samples for areal unit j).

Case III is closest to what occurs for much of the socio-economic geo-referenced data. It is not a stratified random sampling design!

2.2. An example of the model management of error problem

Error in remotely sensed data originates from at least two sources. First, the infinite number of points in a pixel is "averaged" into a single value for the pixel. This measure becomes increasingly representative as the internal variation of a pixel decreases (measurement error disappears). Refinements in sensors that have reduced the size of pixel have attempted to minimize, this source of error. Second, instrument induced attribute error arises from the effects of light scattering so that what is recorded for pixel i includes a weighted average of the attributes in neighboring segments of ground truth. The form of this error, which may be characterized with a "point spread" function, is analogous to a weak spatial filter being passed over the surface.

Arbia and Haining explored this problem during their visit to the NCGIA/UCSB. They have specified the point spread function for this situation in parallel with a standard spatial autoregressive model scheme. They note that this error impacts upon an analysis in three different ways. First, the mean statistic for a surface may be altered by an inappropriate pixel size. Second, spatial autocorrelation of this spillover error is induced by the point spread function process. Third, systematic attribute error may appear in the mean statistic due to its being strongly correlated with the form of the underlying terrain. All three of these sources often are convoluted, further complicating error analysis of remotely sensed data.

2.3. Statistical diagnostics for geo-referenced data

Potential error in geo-referenced data may be indexed by statistics used to earmark data anomalies or aberration. Possible indices include those conventionally used to analyze outliers, leverage points, and influence functions[1]. One important difference here is that geo-referenced data differ from traditional data so much that these statistical measures must be adjusted [Wartenberg (1990) provides selected demonstrations of this point]. In addition, the presence of spatial dependence that is uncompensated for may mask some, and erroneously uncover other outliers, leverage points, and influence functions.

Outliers, which are unusual pieces of data, stand apart from the rest of a data set; they are extremely large or extremely small values when placed within the frequency distribution of all data points. Moreover, they appear to be inconsistent with the remaining observed values. Outliers become troublesome in a statistical analysis because they may unduly influence summary descriptors, rendering misleading characterizations of a data set. For geo-referenced data, outliers must be defined locational as well. Even when a value is not a pathological extreme of some sample, it still could be an outlier when it is placed into a geographic context. This situation is further complicated by the presence of spatial dependence, which may amplify or dampen underlying extreme values.

If an outlier contributes disproportionately to a summary descriptor, then it is viewed as a leverage point. If an outlier is an observation whose deletion would cause a disproportionate change in a parameter estimate, then it is viewed as an influential point. Griffith has investigated influential observation diagnostics used to identify leverage points and outliers, with special reference to Hoaglin and Welsch's index (which seeks to reveal data points that are potentially influential by virtue of their location in the variable space), and has reported preliminary findings from this work based upon simulation experiments. Algebraically, for a data vector X the leverage measure for an error covariance structure of $V\sigma^2$, where $V^{-1} = \Gamma\Gamma$ is a spatial filter, becomes $H = \Gamma X(X^t V^{-1}X)^{-1}X^t\Gamma$. This matrix is altered both by the filtering ΓX, and by the embedded V matrix, implying that spatial dependence makes a considerable difference here, too. Meanwhile, spatial dependence effects were found to compromise the calculation of this index measure. Further, these impacts are not necessarily removed by generalized least squares estimation, and seem sensitive to the filter used for this purpose. And, the power of this index seems to be dramatically impacted upon by the presence of ignored spatial dependence.

Another problematic feature for geo-referenced data has to do with the relative location of an areal unit; peripheral locations can end up having high leverage, as has been shown by Unwin and Wrigley (1987).

This evaluation of diagnostics can be extended to other popular indices, too. For example, Daniel and Wood have proposed a statistic to determine whether or not data points are remote in the variable space. This statistic reduces to a squared z-score for a single variable. Again, algebraically, for a data vector X having an error covariance structure of $V\sigma^2$, where $V^{-1} = \Gamma\Gamma$ is a spatial filter, this index becomes $(X - 1^t V^{-1}X/ 1^t V^{-1}1)^2/[(X - 1^t V^{-1}X /1^t V^{-1}1)^t V^{-1}(X - 1^t V^{-1}X/ 1^t V^{-1}1)/(n-1)]$. Except for a constant of proportionality, this index is exactly the Cook's distance (which seeks to summarize both leverage and influence) for a single variable situation; Cook's distance takes a jackknifing approach to diagnostics. Once more one should expect that spatial dependence makes a considerable difference here.

3. IMPORTANT OBSTACLES AND IMPEDIMENTS

A number of important barriers stand in the way of sound geo-referenced data quality assessment, and its accompanying visualization. These deterrences will be classified into theoretical, computational, and technical problems.

[1] These results appear to extend to other diagnostics, too. For example, Griffith has reported on Mallows C_p index, which was developed in conjunction with the "all possible regression" problem, where an analyst fits a regression equation for each of the possible combinations of candidate predictor variables (for p variables, there are 2^P possibilities). If a specific variable X_j has $\beta_j \neq 0$ then excluding it from the model results in b being biased, unless the retained variables are orthogonal to the deleted variables. The C_p index measures this bias, which is present due to the absence of variables, allowing an analyst to examine the trade-off between bias in the regression equation and reduction in the average error of prediction over the set of all possible subset regressions. But results of this index are extremely sensitive to the error estimate used, which will be incorrect if any latent autoregressive structure in the data is overlooked. Numerical results for this index suggest that as both the sample size and the degree of spatial dependence increase, C_p increases, even though all X variables are included in a regression equation. In other words, bias being detected strictly due to missing autoregressive terms can be substantial and serious. Griffith also has explored the lack-of-fit (which evaluates the assumption of a linear structure) and pure error (which tries to give a model-independent estimate of error) diagnostics. lie has reported preliminary simulation results implying that spatial dependence impacts upon these diagnostics, too, both when the exact pure error and the approximate pure error calculations are involved. tic has corroborated this finding with results from an empirical example. Apparently the Type I error probability for lack-of-fit is most conspicuously affected by spatial dependence.

3.1. Theoretical problems

Three conspicuous theoretical spatial statistics impediments can be gleaned from the preceding discussion. First is the issue of specifying the inverse covariance matrix V^{-1}, or its decomposition form $\Gamma\Gamma$. Thus far spatial researchers have found that either a conditional autoregressive (CAR) or a simultaneous autoregressive (SAR) formulation are appropriate. The CAR form of this covariance matrix is consistent with the popular negative exponential characterization of the semi-variogram found in kriging. Ripley even argues that the jury is in now on this matter, and that the CAR is the correct model for spatial statistics! Arbia (1986) argues just the opposite, though; he contends that rejection of specifications other than the CAR and SAR for geo-referenced data mostly is because of the defective manner in which spatial data are analyzed. Findings reported for the leverage index appear to be sensitive to the specification of these matrices, implying that this specification issue is a serious obstacle to advancements in geo-referenced data quality analysis; identification of potential error should not be overly dependent upon V^{-1}.

The second impediment concerns the problem of edge effects. Because peripheral locations can end up having high leverage, some type of adjustment may well need to be made to diagnostics for these locations. Otherwise, a map of potential error will allow visualization of something that is intuitively obvious without doing a single calculation! Ripley discovered useful edge correction techniques for point data; unfortunately, according to work completed by Griffith (1988a) and others, spatial dependence latent in attribute data seems to defy simple correction.

A third impediment deals with the conversion of standard diagnostics to spatial statistics. Simulation studies need to be conducted with, for example, the Daniel and Wood index, and the Cook's distance, to better understand their spatial statistics properties and behavior. Additional diagnostics need to be developed that explicitly probe spatial autocorrelation situations. Recent work by Wartenberg (1990) and by Getis and Ord (1991) seems to be addressing this second need. But their findings should be couched in a multiple comparisons test framework, if useful yardsticks for decomposed spatial autocorrelation statistics are to be established.

3.2. Computational problems

A single troublesome computational impediment is affiliated with the preceding discussion. Spatial statistics calculations are numerically intensive; calculating spatial statistics diagnostics also will be numerically intensive. In part this feature of spatial statistics is evident from matrix multiplications involving V^{-1} and Γ, which are n-by-n. In part it is due to estimation of the spatial linear operator matrix Γ. If a CAR model is employed, then $V^{-1} = (I - \rho C)$, where ρ is a spatial autocorrelation parameter and matrix C is a spatial weights matrix; if an SAR model is employed, then $\Gamma = (I - \rho W)$, where matrix W usually is the stochastic version of the spatial weights matrix. These models are estimated using maximum likelihood techniques, which involves calculation of the determinant $|V^{-1}|$. The eigenvalues of this determinant are needed in order to simplify estimation. This extraction of n eigenvalues dramatically increases the computational requirements of spatial statistics; supercomputer work by Griffith (1990) found that this single task alone can account for as much as 90% of computation time for modest values of n.

The important obstacle here is presented by a need to calculate the determinant $|V^{-1}|$ for very large n. Remotely sensed images may have millions of pixels. GIS databases may have many thousands of areal units. In these settings, especially given the finding that peripheral areal units by definition can be identified as outliers, computation of meaningful diagnostics for visual display remains problematic.

3.3. Technical problems

Two technical impediments are associated with the preceding discussion. First is identification and transformation of a battery of useful diagnostic indices. Those by Daniel and Wood, by Hoaglin and Welsch, and by Cook have been mentioned above; there are many others. The transformation from conventional statistics to spatial statistics will require considerable algebraic manipulation, especially in order to capture the most general cases.

A second technical obstacle pertains to the visualization of potential error. A researcher moving across a map will need a window that focuses on a local portion of the map. This window will need to he able to highlight extreme values in terms of the local map context. The position of a highlighted value in the aspatial frequency distribution for the entire map will need to be displayed simultaneously, allowing the researcher to determine whether or not it is a global extreme. Essentially, then, the problem is one of producing these two displays nearly instantaneously and continuously. The recent software by MacDougall. and suggestions made by Monmonier in his position paper, come very close to what I envision.

4. RESEARCH AREAS

Interaction of two components of data quality, namely measurement error and sampling error, combined with links to visualization have been examined in this position paper. The general strategy suggested here is that, first, an attribute error index should be constructed. This index should suggest the probability of observational error in an aspatial context. A practical visualization tool to aid in its interpretation Would be placing the index value on a scale that identifies the extreme values of this statistic, as well as where it falls in its parent sampling distribution. Second, the local geographic context of spatial data values should be inspected. This notion can be quantified by decomposition of spatial autocorrelation statistics for individual values. A practical visualization tool to aid in its interpretation would be a mobile or roving window. Third, the geographic arrangement of potential error should be indexed and exhibited. Of course visualization here would be achieved by a map of error probabilities, which could be computed from the aforementioned global and local considerations. These three aspects of quality assessment should be developed interactively with users, for the primary goal here should be to respond to user needs.

Data model needs refer to (1) specifying the inverse covariance matrix, and (2) positing a useful underlying probability model. Presently at least the CAR and SAR specifications of the inverse covariance matrix can be studied. Besag (1974) contends that only a few selected probability models are consistent with spatial autocorrelation, including the normal, the Poisson, and the multinomial. The normal model underlies most of the arguments put forth in this position paper; considerably more work needs to involve these other two models, which historically have been found to be consistent with certain types of geo-referenced data.

5. REFERENCES

Arbia, G. 1986. "Problems in the estimation of the spatial autocorrelation function arising from the form of the weights matrix," in D. Griffith and R. Haining (eds.), Transformations Through Space and Time, The Hague: Martinus Nijhoff, pp. 295-308.

Anselin, L., and D. Griffith. 1988. "Do Spatial Effects Really Matter in Regression Analysis?," Papers of the Regional Science Association, 65: 11-34.

Arbia, G., and R. Haining. 1990. "Error propagation through map operations," paper submitted to Technometrics.

Besag, J. 1974. "Spatial interaction and the statistical analysis of lattice systems," Journal of the Royal Statistical Society, 36B: 192-236.

Getis, A., and J. Ord. 1991. "The analysis of spatial association by use of distance statistics," unpublished manuscript, Department of Geography, San Diego State University.

Griffith, D. 1988a. Advanced Spatial Statistics, Dordrecht: Martinus Nijhoff.

Griffith, D. 1988b. "Interpretation of standard influential observations regression diagnostics in the presence of spatial dependence," paper presented to the 35th Regional Science Association, Toronto.

Griffith, D. 1990. "Supercomputing and Spatial Statistics: A Reconnaissance," The Professional Geographer, 42: 481-492.

Griffith, D. 1991 "Pure error and lack-of-fit regression diagnostics in the presence of spatial dependence," Sisteini Urbani, in press.

Griffith, D., and C. Amrhein. 199 1. Statistical Analysisfor Geographers. Englewood Cliffs, NJ: Prentice Hall.

International Journal of Geographical Information Systems

Heuvelink, G., and P. Burrough. 1989. "Propagation of errors in spatial modelling with GIS," 3: 303-322.

Chrisman, N. 1987. "Efficient digitizing through the combination of appropriate hardware and software for error detection and editing," 3: 265-277.

Bolstad, P., P. Gessler, and T. Lillesand. 1990. "Positional uncertainty in manually digitized map data," 4: 399-412.

Dunn, R., A. Harrison, and J. White. 1990. "Positional accuracy and measurement error in digital databases of land use: an empirical study," 4: 385-398.

Lodwick, W., W. Monson, and L. Svoboda. 1990. "Attribute error and sensitivity analysis of map operations in geographical information systems: suitability analysis," 4: 413-428.

Unwin, D., and N. Wrigley. 1987. "Control point distribution in trend surface modelling revisited: an application of the concept of leverage," Transactions of the Institute of British Geographers, N. S. 12: 147-160.

Wartenberg, D. 1990. "Exploratory spatial analysis: outliers, leverage points, and influence functions", in D. Griffith (ed.), Spatial Statistics: Past, Present, and Future, Ann Arbor, MI: Institute of Mathematical Geography, pp. 133-156.

An Approach to Spatial Data Quality Using Standard Visualization Tools

Virginia R. Hetrick, Ph.D.
Office of Academic Computing
University of California, Los Angeles
Sneakernet address:
2260 Linda Flora Drive
Bel Air, California 90077
Bellnet address: (213) 471-1766
Electronic mail address:
until 1 July 1991: CUSGRAF@VM.OAC.UCLA.EDU
after 1 July 1991: DRJUICE@NERVM.NERDC.UFL.EDU

Draft Paper for
GIS/Visualization Specialists Conference
Castine, Maine

8-12 June 1991

An Approach to Spatial Data Quality
Using Standard Visualization Tools

Using Standard, public-domain visualization tools to evaluate and record the quality of spatial data can provide a powerful, yet extremely cost-effective, alternative to custom-developed software. Why might such a capability be useful? With the development of advanced techniques for using data, it is possible to make effective use of data which are flawed in some way and which might otherwise have to be discarded or reacquired. However, an important component of these techniques is that the quality of the data needs to be recorded and used in the analytical techniques. Finding a mechanism to accomplish this is sometimes difficult. This short paper demonstrates how certain aspects of a quality evaluation might be carried out on selected types of spatial data. In principle, this process could be applied to aspatial data as well. For the particular case at hand, we are interested in determining, first, whether all of the items in a particular dataset fall within "normal" ranges and, second, whether the values presented in the dataset are internally coherent and consistent. The toolkit demonstrated here accomplishes those objectives and helps the researcher derive conclusions more quickly and, in some cases, more accurately.

The Toolkit

The toolkit used to prepare this paper is the NCSA visualization suite. Developed at the National Center for Supercomputing Applications located at the University of Illinois, this toolkit is widely used to prepare visualizations of scientific results for viewing on Macintoshes, PS/2s, and X-stations, where the results can be animated or captured on hardcopy devices, video, or a film recorder, as well as being examined in more detail over a longer period of time on the workstations. The objective of the toolkit is to provide rapid visualization of many types of scientific data.
<footnote 1 goes here>

< begin footnote I >

The software is available from FTP.NCSA.UIUC.EDU by anonymous FTP. The complete software not only includes the workstation software but also includes software to generate visualization riles in the appropriate format on several kinds of host machines, from workstations to Crays and IBM 3090s.

The primary components for the host fall into three categories: a C-callable subroutine library, a Fortran-callable subroutine library, and a series of utility programs designed to minimize the need to write custom software. The primary components for workstations are available for three classes of machine: most workstations running X-Windows, IBM PS/2s running DOS, and Apple Macintoshes running Mac/OS. These components fall into three categories: 'image' software for examining and analysing raster riles, 'datascope' for scientific data riles, and 'dicer' for three-dimensional image files.

Commercial versions of the software which can be run Macintoshes are available from the Spyglass Corporation, Champaign, Illinois. The three programs available from Spyglass are: View, Transform, and Dicer, which correspond to the original components available from NCSA.
< end footnote 1 >

To use any of these tools first requires that the data being evaluated be converted to the appropriate file format. All of the tools in this set use a self-describing file called an HDF (Hierarchical Data Format) file which can be created in many different computational environments. For this paper, since all the datasets were already on UCLA's IBM 3090-600J run by the Office of Academic Computing, the conversion to HDF was carried out in the MVS/ESA environment available there. Two Fortran programs were used to convert the original data.

< begin footnote 2>

These two programs are available from the author as examples of how to build your own programs to do the file conversions.

<end footnote 2>

For the purposes of this paper, two types of data will be considered:

- data distributed as an isosurface

- data with some type of periodicity

Isosurface data.

Typical of this class of data are topographic, bathymetric, and temperature data. All of these data vary continuously over geographic space. Thus, if a particular data value falls outside an expected range or the data do not seem to vary as an isosurface, the researcher should conclude that the data point is incorrectly entered or recorded or that there is an unexpected perturbation in the data which should be investigated more thoroughly. In addition, for isosurface data, displaying the data visually rather than in tabular form can help the researcher identify aberrant data more rapidly. For example, in topographic data, if the display shows an exceptionally low value in an area of high values, without the normal gradation of tone, the researcher might reasonably suspect that the low value was a data acquisition error and should be more closely examined. The dataset used to illustrate this particular application is a series of thermal images of a weather system acquired by satellite.

Periodic data.

The major characteristic of periodic data is that similar data values occur for the same observation point at some time interval. For example, using the monthly average of daily maximum temperature should show rises and falls with the appropriate season. The value for corresponding months in different years should be relatively similar. Also, the values for nearby stations at like elevations should be comparable. This would normally be confirmed statistically by examining means, standard deviations, and variances of the data, among other measures. However, no statistical procedure in common use can identify specific outliers in this type of data as effectively as visualization of the data using the toolset demonstrated here.

Examples of this class of data are monthly average rainfall data or monthly average temperature data over some lengthy period of time. Periodic data are somewhat more difficult to examine, yet the toolkit can help identify problems in such data as well. While this type of data need not necessarily be spatial distributed, they frequently are. For example, the data used to illustrate this particular application are the average monthly maximum temperature data for several stations in southern California.

The Data Quality Matrix and Its Creation

To allow the use of data of different quality by the informed researcher, data quality matrix (DQM) can be developed. In this frame of reference, the DQM is simply an 8-bit raster file where the quality of the data in the "real" dataset can be represented. Each element in the DQM corresponds positionally and informationally to an element in the original clataset. The values in the DQM can vary, at the researcher's choice, over a range as great as 0 to 255 and the quality value for each element in the dataset is stored in the corresponding position in the DQM as a binary number. The DQM can be displayed and searched visually or numerically just as any other HDF file. The DQM can be made up of any values the researcher desires. However, normally the matrix is displayed with a palette (color lookup table) defined such that the brightest values are high quality data while low quality data are represented by low values in the range and are darkest. If desired, the specific data quality, values (DQV) can represent specific quality aspects of the data, i.e., a "1" represents one quality type, a "2" represents another type, and so forth, such that the data quality values have completely arbitrary meanings. A better alternative is to develop a more general table, called a data quality descriptor table (DQDT), which can be used for all of the datasets created during the course of a single project or in the scientist's ongoing research program. When this latter alternative is chosen, these should be additive so the components can be broken out if necessary. An example of how this could be done is shown in the table below.

<begin Table 1 >

Table 1. Sample data quality table.
10 Missing data
20 Data acquisition Incomplete
30 Data acquisition error; probable sensor/platform fault
40 Bad data; acquisition fault
50 Bad data; unknown cause
1 Not replaced - no estimate possible
2 Not replaced - estimate possible
3 Replaced with estimate - average of incomplete data
4 Replaced with estimate - average of preceding and following observations
5 Replaced with estimate - reconstructed by formula (specify formula)
6 Replaced with estimate - reconstructed by regression with - ---

< end Table 1 >

For example, missing data, not replaced with an estimate and no estimate possible would be 11; bad data of unknown cause replaced by an estimate averaging the preceding and following observations for that data element would be 54. Handling the data in this way allows preservation of the original data quality when that is desired but allows researchers to understand which data elements may be somewhat less reliable than others. The example table, in fact, is the data quality table used by the author to construct a data quality matrix for the thermal data used to illustrate the isosurface example in this paper. The data quality table specification is also included as a set of comments within the HDF file containing the DQM when it is created.

How the Toolkit Works

As noted earlier, the data can be converted into HDF files on any system of the researcher’s choice. For the isosurface data, the HDF file is a raster file of the size required by the researcher. The file can contain 1-bit, 8-bit, or 24-bit rasters, also as appropriate to the research. For the current isosurface data, the files are B-bit files because that is the form in which they were available. For the periodic temperature data used as an example here, the HDF files are in the scientific data format. Data in this type of file is stored as IEEE-floating point data, regardless of the type of system which generates the HDF file. As a consequence, moving HDF files from one system to another only requires treatment as any ordinary binary file with no character set translations required to move the file among various systems. As with the raster files, the file size is appropriate to the researcher’s task and is limited only by the size of the various system elements in the researcher’s workstation.

<begin footnote 3>

In addition to the 8-bit raster and scientific data format files, two other rile formats are available. They are the 24-bit raster format and the annotation format riles. The 24-bit raster files allow the user to create HDF files containing ’true-color’ data, where each pixel is represented by three 8-bit bytes to impart more information. The annotation format riles allow specification of annotation information for the data files.

<end footnote 3 >

Examples of data quality evaluation.

Isosurface data.

The isosurface data were made into an HDF file containing the images themselves as well as the twelve DQMs, one for each of the twelve frames in the sequence. These data are the infrared representation of the storm that crossed much of California on 19 March 1991. The twelve images as they were originally acquired are shown in Figure la through Figure 11. These were acquired at half-hour intervals beginning at 3am GMT.

< begin footnote b2 >

2. U.S. Department of Commerce. National Weather Service. (1991) Western
GOES Satellite, Infrared Observations for 03 March 1991, 0300-1600.:.

<end footnote b2 >

The data in most of these images are excellent quality. The sole exception is in the first frame (Figure 11a). So, the DQMs for these images would indicate the uniformly high quality except for that single data line in Figure la.

We expect to find such problems from time to time in remotely sensed data and would probably attribute the bad line to a sensor fault. We could replace the bad line in one of several ways. The most common replacement technique is to average the preceding and following lines on an element by element basis and use the resulting vector to replace the bad data line. Another common replacement technique is to copy the preceding or following line into the position of the bad data line.

Using Table 1 as the DQDT, we would probably assign a DQV of 43, indicating a data acquisition error with a probable sensor/platform fault and replacing the line with the average of the preceding and following observations.

For this particular task, we have a small Fortran program, REPL, which is run on the single frame with the bad data line. The program extracts the frame with the bad line from the HDF file containing the full sequence of 12 images, finds the bad data line, and averages preceding and following lines on a pixel by pixel basis. After replacing the individual elements in the bad data line, it replaces the repaired frame in the HDF file. REPL then extracts the corresponding DQM from the HDF file and replaces the individual elements in the DQM line with a uniform series of elements having a value of 43.

At this point, the repaired HDF data file can then be used as desired.

<Figures 1 a through 1I and Figure 2 go here.>

Periodic data.

Each set of the periodic data was made into a separate HDF files together with its own DQM contained in the appropriate file. The data are the monthly average of daily maximum temperatures and were calculated from data acquired through the Climatedata CD-ROM facility.

< beg- footnote b3>

4. USWest Optical Publications. (1988) Climatedata: Summary of the Day. Denver: USWest Optical Publications.

<end footnote b3 >

Only data from 1950 through 1988 are used in this example. The datasets are:

<begin indent>

Burbank	Valley Pumping Station
Los Angeles CC	Los Angeles Civic Center
Los Angeles	Weather Service Office, Los Angeles Airport
Pasadena	Near Cal 134 and Orange Grove Boulevard
Santa Monica	Santa Monica Pier
UCLA	UCLA campus

<end indent>

The visualizations of these datasets are shown in Figure 2a through Figure 2h. The average for each month is represented as a single rectangle. Darker values indicate lower temperatures while brighter colors indicate warmer temperatures. The horizontal axis represents the months while the vertical axis represents the years, with January, 1949, at the top left and December, 1990, at the bottom right. Consequently, one would expect that the left and right sides of the visualizations would be darkest while the area near the center would be lightest. In fact, that is what is shown in the data with several exceptions. Those several exceptions can be analysed in the manner described below.

Examining the Santa Monica data, we first note some areas are black, indicating missing data for several months. These data are missing in the original data and cannot be reconstructed directly.

So, in the data quality matrix, those particular data elements are coded as 11, missing data with no estimate possible. Were the researcher to decide that it was important to have a complete data series, it may be possible to replace the missing elements with observations from a nearby station.

For example, in the instance of Santa Monica, the Los Angeles Airport Weather Service Office (LAX) is only about 10 km. distant and the UCLA station is 17 km. distant. Further, for the observations present, a simple correlation analysis reveals that the coefficient of determination for the correspondence between the Santa Monica and LAX is datasets is 73 percent while the coefficient of determination for Santa Monica and UCLA is slightly less than 68 percent. So, one way to deal with the missing elements would be to simply replace them directly with elements from the LAX dataset which is most highly correlated with the dataset we are trying to repair. We could be reasonably certain that the repaired dataset would not "look" very different. The other alternative would be to perform a regression of the Santa Monica data with the LAX data, since it was more highly correlated with Santa Monica than the UCLA dataset and use the regression coefficients to calculate a new value for each of the missing elements in the Santa Monica data. This is, in fact, what was done. The Santa Monica visualization both before and after repair is shown in Figure 3.

<Figure 3 goes here>

However, having done such a replacement of these elements, this fact can be recorded in the DQM with an annotation that the replacements were performed and that the replacements were calculated from the regression of the Santa Monica data with the LAX data. We can further record the regression equation used to repair the dataset. So, now in the elements of the DQM which we have replaced, we have the value 16, which is annotated as

Missing data, replaced with estimate, reconstructed by regression S
TA-MONICA = 14.8180 + 0.7467 * LAX

Conclusions

In managing data for a scientific project, two issues are paramount: data integrity and specification of data integrity in some manner which is understandable to researchers and others who use the same data. Using the HDF formats to combine the data with the spatial data quality information in the same file which is easily ported from one platform to another provides an accessible and easily managed first step to approaching the problem of visualizing spatial data quality.

NOTE:

Reproduction of the illustrations is not yet completed. If you would like a copy of the illustrated version of this paper, please give me your paper mail address while we're at Castine. vh

The Quantitative Approach to Error Propagation in GIS[*]

Gerard B.M. Heuvelink
Department of Physical Geography
University of Utrecht
PO Box 80.115
3508 TC Utrecht
The Netherlands
email: iaaheuv@cc.ruu.nl

Introduction

This paper gives an overview of the research on error propagation in GIS, as it is carried out at the Department of Physical Geography of the University of Utrecht. Perhaps the main difference between this research and the work done by others in this field is that it Is solely restricted to *quantitative* errors. The research thus exclusively deals with spatial variables that are *interval* or *ratio*. Clearly the restriction to this type of data is a severe one, because much of the data stored in GIS is *nominal* or *ordinal*, such as topography, landuse, cadastral data, geologic maps and soil maps. However, many of the problems in physical geography do relate to quantitative data (such as elevation, precipitation, infiltration, concentration of pollutants, depth to groundwater, clay content of topsoil), and this is increasingly so because quantitative methods are becoming more and more popular in physical geography (and perhaps not only there). The decision to concentrate on quantitative errors was also led by the fact that they are relatively easy to deal with, making it possible to get much further-reaching results on how these errors affect the quality of GIS results. An immediate advantage of the quantitative approach is that both the error modelling process and the error propagation analysis can adopt well-developed techniques from (geo-) statistics and standard error propagation theory.

The first part of this paper contains a highly condensed description of the theory and methodology used to handle the propagation of quantitative errors in GIS. It is intended to serve as an introduction to the second part of the paper, in which attention is directed to how to integrate these techniques with GIS and to how this may substantially improve the functionality of GIS.

Modelling Quantitative Errors in GIS

To model the presence of quantitative errors in a spatially distributed variable z, the standard statistical approach is to represent it as a continuously distributed *random field Z.* Important properties of Z are its mean μ and standard deviation σ, which are themselves spatially distributed. Thus for each location x, Z(x) is a random variable with a mean μ(x) and standard deviation σ(x). Of the two maps μ and σ, the former is just the ‚ordinary' map, while the latter is a measure for its 'error'. The error map σ will generally be spatially variable. For instance, when a map is obtained from interpolating point measurements, the interpolation error will be smaller in the neighbourhood of measurement points and larger further away from them. When Z refers to a choropleth map, areas with a larger within-unit variance will be assigned larger standard deviations than areas where the within-unit variance is smaller.

An obvious way to visualize data quality is in this case just to display σ(preferably together with μ), or perhaps to show the relative error σ/μ. One might also use a DRAPE function to display both die data and its error in one picture. Two alternatives to these rather common techniques have recently been given by Bregt (1991).

The mean and standard deviation are properties of Z that *always* need to be known' in order to calculate how the error in Z propagates to the output of a GIS operation, but knowing these will not always be sufficient. For instance, to use the Monte Carlo method one also needs to know the type of probability distribution. The Gaussian distribution is a natural candidate for many situations, but other distribution types may also occur (for instance, infiltration capacity of soil is highly skewed and better fits a lognormal distribution).

Another important property of Z that ought not carelessly be neglected is correlation. Presence of correlation can have an important influence (both positive and negative) on the results of an error propagation analysis. Three types of correlation can be distinguished (Wesseling and Heuvelink 1991). The first type refers to the correlation between the values of two maps at the same location (say the correlation between Zi(x) and Zj(x)). It is often assumed to be spatially invariant. The second type is spatial auto-correlation, which defines the correlation between two values of the same variable but at different locations (correlation between Z(xk) and Z(xl). The third type, spatial crosscorrelation, defines the correlation between two variables at different locations (correlation between Zi(xk) and Zj(xl). For continuously varying spatial variables (those that are most often stored in a raster GIS) one

[*] This position paper is hardly concerned with the visualization of data quality (the author's knowledge on this subject is rather poor) but more with error modelling and error propagation.

will often make certain stationarity assumptions on the correlation functions (as is done in geostatistics by means of the (co-)variogram, Oliver and Webster 1990).

Summarizing this section, to completely identify an error-contaminated map it is necessary to specify its mean, its standard deviation, its distribution type, its spatial auto-correlation and its correlation with other maps stored in the GIS. Clearly this is a lot to be known and rarely will this information be available in practice. This lack of information thus creates an additional problem which will be addressed later on.

Error Propagation Techniques

Having presented a model for characterizing errors in spatial variables, the next step is to calculate how these errors propagate through a GIS operation. We concentrate on GIS operations (spatial models) that can be generally expressed as

$$R = F(Z_1, Z_2, \ldots, Z_n, \alpha_1, \alpha_2, \ldots, \alpha_m)$$

where the resulting map R is obtained by applying expression F on input maps Z, and model coefficients αi. Here model coefficients are added because very often a GIS operation is in effect (a part of) a model. F may be a simple regression model but it can also stand for a dynamic, distributed erosion model. To take the uncertainty of a model into account, some of the αi will be stochastic as well. For instance, for a regression model the αi represent the regression coefficients together with the residual noise term, which are all stochastic. If F represents a crop growth model, then the αi should not only represent the model parameters but should also comprise model error.

The aim of an error propagation analysis is to determine the statistical properties of the random field R, given the properties of the Zi, the αi and the type of function F. Only in a few special cases can this be done analytically. In the majority of cases one will have to use other methods, such as numerical, Monte Carlo, or finite order (Taylor) methods. We will not discuss these here but only say that each method has its drawbacks. The numerical and Monte Carlo method are time consuming and the results do not come in a nice form. The Taylor method is only practical for relatively simple models and it involves approximations that cannot be ignored when the operation F is not sufficiently smooth. For complex models the Monte Carlo method seems the only workable method.

Potentially of great value is die ability to use error propagation techniques to determine how much each of the inputs contributes to the output error (Heuvelink et al. 1989, Heuvelink et al. 1990). In this way one is able to calculate how much the quality of the output improves if the accuracy of a particular input increases. It can also be used to tell whether it is the model itself or the model input that is the main source of error. Error propagation techniques are thus not only useful to quantify the output error, but they also provide information with which rational decisions can be made on how best to improve the quality of the output.

Some disadvantages of error propagation techniques must also be mentioned. First, to use them the user must specify a lot of statistical information, which is generally not readily available. Second, they often involve approximations of which it is hard to tell whether these may be ignored. If one wants to avoid approximation errors by all means, then one is forced to use time consuming alternatives. This is the third disadvantage, because users are used to getting answers quickly and will be tempted to omit an error analysis if it takes too long. Fourth, an error propagation analysis generally only renders information about the mean and standard deviation of the output, not about correlation or distribution type.

Linking Error Propagation Tools with GIS

The theoretical developments have by now reached a stage at which it seems the right moment to seriously work on implementing the error propagation techniques in GIS. This brings about the problem of how this can best be done. The view taken by the Utrecht research group on this subject was recently addressed in a paper presented at EGIS '91 (Wesseling and Heuvelink 1991). This section briefly summarizes the main conclusions of that paper.

In the EGIS paper a prototype error propagation tool (ADAM) is presented that can be viewed upon as a first step towards the full integration of error propagation techniques in GIS. ADAM is composed of two parts, a model description language and the associated compiler. The model description language serves as an intermediate between user and tool, it provides the user with a means to supply all the information needed to perform the error propagation analysis. Next, the compiler applies the chosen error propagation technique to the particular problem. It generates a string of GIS operations that embody the error propagation analysis. This string of operations must be carried out by the GIS to yield the mean and standard deviation of the model output. Thus ADAM does not actually perform the calculation, it merely translates tile problem into an appropriate sequence of basic GIS operations. Separating the compiler from the map processor was purposely done to achieve an easy linkage with any of the current GI systems.

The only requirement to be met by the GIS is an appropriate functionality of its script language, and this forms no serious obstacle. One of the goals for the nearest future is to establish the link between ADAM and commercial GIS at selected test sites, and to evaluate its performance in practice.

Quantifying Input and Model Error

As mentioned before, to carry out an error propagation analysis, the user must specify all errors involved. In practice this still is a major problem because a user rarely knows how accurate his map really is. It is obvious that one cannot demand a map user, who was not involved in the map making process, to know the statistical parameters of such an error-contaminated map. These parameters need to be provided by the map maker. Consequently, it would be advisable if users only purchase a map /I if it is accompanied by its associated error map a, and/or any other information that reveals the accuracy of the map.

For quantitative data, several methods can be used to compute an error map. When a map is constructed from interpolating point measurements, the interpolation error can be obtained by using geostatistical interpolation (Oliver and Webster 1990). The error of a choropleth map can be quantified by the within-unit sample variances. When these approaches do not apply, other methods need to be sought to obtain the error map. This may sometimes even imply the use of informed guesses, because one often does have some idea about the magnitude of errors. But it remains the task of the map maker to supply the error map. For instance, not the farmer must tell how reliable a soil map is in predicting the texture of his land, but the particular soil survey institute should provide that information.

Apart from errors in the input of GIS operations it is also important to pay attention to model error. This should not be ignored because there is no sense in only paying attention to the error of GIS data, when the data is substituted in a model of which one has no idea of its accuracy. Quantifying model error is often difficult and should be done by experts in the particular discipline to which the model applies. It is also important to realize that a model has no universal accuracy: an erosion model which is developed in the United States may behave poorly when it is applied in Europe. These are all problems that go beyond the scope of 'error modelling in GIS', but they need to be considered when a particular error propagation analysis is performed.

Towards Intelligent GIS

The present situation is thus one in which error propagation techniques are not available as a standard tool in GIS, nor is the accuracy of data stored in GIS, nor are models and their accuracy linked to GIS. If this situation were to change for the better, then there potentially comes a GIS available which is much more powerful than the present one. It could lead to a GIS which is able to advise the user in what way optimally to achieve a desired aim.

When error propagation techniques are incorporated in GIS, the user will be able to calculate the accuracy of his result. If the result does not satisfy tile desired accuracy level, the same techniques can be used to decide how best to proceed to improve accuracy. The extended GIS can tell which of the inputs should be collected more accurately, and for which inputs measurement costs can be cut down without noticeable consequences. Such a GIS can also advise on how to balance input and model error. It can show that it is unwise to spend a lot of efforts on collecting data if what is gained is immediately thrown away by using a poor model. It can also show why a simple model is as good as a complex model if the latter needs lots of data that cannot be accurately obtained.

If GIS can do all this and also store and use expert knowledge on how to deal with specific problems, then we are on the way towards what one might call an intelligent GIS (Burrough 1991). Clearly at present we are still far away from such a GIS. But many of its tools are available in one way or another, they only need some refinement and implementation. One of the feasible goals for the near future is the integration of error propagation techniques with GIS. If this is accomplished, already will we have provided the qualified GIS user with a valuable tool.

References

Bregt, A.K. (1991). Mapping uncertainty in spatial data. In: J. Harts; et al. (eds), Proceedings EGIS '91, pp. 147-154. EGIS Fouodation, Utrecht.

Burrough, P.A. (1991). The development of intelligent Geographical Information Systems. In: J. Harts et al. (eds), Proceedings EGIS '91, pp. 165-174. EGIS Foundation, Utrecht.

Heuvelink, G.B.M., P.A. Burrough and A. Stein (1989). Propagation of errors in spatial modelling with GIS. Int. Journal of GIS, 3, pp. 303-322.

Heuvelink, G.B.M., P.A. Burrough and H. Leenaers (1990). Error propagation in spatial modelling with GIS. In: J. Harts et al. (eds), Proceedings EGIS '90, pp. 453-462. EGIS Foundation, Utrecht.

Oliver, M.A., and R. Webster (1990). Kriging: a method of interpolation for geographical information systems. Int. Journal of GIS, 4, pp. 313-332.

Wesseling, C.G., and G.B.M. Heuvelink (1991). Semi-automatic evaluation of error propagation in GIS operations. In: J. Harts et al. (eds), Proceedings EGIS ’91, pp. 1228-1237. EGIS Foundation, Utrecht.

POSITION PAPER
NCGIA SPECIALIST MEETING

Geoffrey R. Loftus
Department of Psychology
University of Washington
Seattle, WA 98195
Gloftus@milton.u.washington.edu

I would like to accomplish two things in this position paper. First, I'll provide a brief sketch of who I am, why I'm at this conference, and the kind of research I do. Second, I'll suggest some possible contributions that an experimental psychologist might make in grappling with the issues that have motivated this conference.

Personal Data

At the risk of sounding patronizing and/or narcissistic, I'll begin by providing some information about myself and the kind of research I do. This section is motivated by my observation that very few of the participants in this conference are psychologists.

Who I am and Why I'm Here

I'm an experimental psychologist doing research in the fields of visual perception and visual memory. I'm interested in the issues covered in this conference for two reasons. First, the display of visual information has always been of intrinsic interest to me: I discover, for instance (somewhat to my dismay) that when I write a manuscript I seem to spend as much time obsessing about how a manuscript looks as to how it reads. I hope to justify these proclivities by discovering, at meetings like this, that such a strategy is actually reasonable in terms of conveying the information that I'm attempting to convey.

Second, I see design of graphical displays as being a potentially interesting application of the somewhat sterile basic research that I normally do. (As an aside, it continually amazes me that many of my vision -scientist colleagues seem not to realize that the eyes they study are generally attached to real people who use them to do real things such as trying to read the low-contrast, tiny-typeface, and generally incomprehensible slides that appear with startling regularity during talks given at vision-research conferences).

Research in Visual Perception and Memory

My interests in vision-related topics began with my PhD dissertation in which I investigated the relationship between where people look in a visual scene, and what they later remember about the scene. (For purposes of this paper, the term "scene" refers to any static visual display). From that project evolved a general framework that has guided my research for the past twenty years. This framework is rooted in the fact that scanning of static visual scenes takes the form of a series of periods during which the eye is relatively immobile (called *fixations*) separated by quick jumps of the eye from one place to another (called *saccades*). Fixations are variable in duration, but typically last on the order of 250 - 300 ms. It is known that visual information is extracted from the fixated portion of the scene during fixations, whereas vision is essentially suppressed during saccades. This means that the visual system is continually presenting the brain with a series of "still snapshots" that must be integrated into one coherent representation of the picture.

Eye Fixations and Research Questions

Given this organization of the visual system, three research questions present themselves. The first is: how does the system determine which portions of a complex visual scene are to be fixated and which are to be ignored? The second is: what is the time course by which the system acquires information within a fixation, and what factors influence this time course? The third question, already alluded to, is: how is the information acquired over a series of fixations integrated into a unified "internal mental model" of the scene being observed?

Perception vs Memory

In studying visual memory, it is convenient to distinguish between perceptual processing and conceptual processing. Perceptual processing refers, roughly speaking, to the original acquisition of visual information from some scene. The output of perceptual processing is some short-term, perceptual representation of the scene that is sufficient for identifying the scene's gist. It, requires a minimum of about 50-100 ms for perceptual processing to produce useful output.

If the scene is to be remembered later on, then conceptual *processing* is necessary. Conceptual processing refers to a variety of different mental events, including rehearsal of the scene, determination of the relationships among different parts of the scene, association of the scene with other related scenes, etc. The output of conceptual processing is a long-term memory representation that can be retrieved and utilized at any subsequent time.

A Typical Ongoing Project

My present research involves a variety of projects, all of which are couched one way or another within this framework. As an example, one such project is designed to address the question: what is it about *degraded stimuli* (e.g., photocopies of photocopies, or slides viewed with the room lights still on) that makes them seem harder to see and understand relative to their undegraded or *clean* counterparts? I considered two hypotheses. The first is that degradation affects perceptual ("front-end") processes only: i.e., once perceptual processing has been affected by stimulus degradation, there is no additional degradational effect on conceptual processing. More specifically, this hypothesis yields the prediction that a briefly presented clean stimulus will engender a memory representation identical to that engendered by a degraded stimulus that is presented for a longer time period.

The second hypothesis is that degradation affects not only perception, but also the conceptual processing that results in the ultimate memory representation. This means that, ultimately, the memory representation of degraded stimuli would be qualitatively different from (and presumably poorer than) memory for clean stimuli.

In a series of experiments, we are trying to determine the circumstances under which one hypothesis or the other applies.

Optimizing Visual Perception of and Memory for Visual Displays

In this section, I will provide very rough suggestions about how this kind of psychological research might be fruitfully applied to optimal display of visual information.

There are many existing strategies for designing such displays. Excellent examples are found in Edward Tufte's wonderfully sensible and intuitive volumes, which I reread in preparation for attending this conference. It's struck me, however, that the rules and suggestions provided in these and other books, while appearing eminently reasonable, are almost entirely lacking in theoretical and/or empirical foundation. I believe that the establishment of such theory and data would serve (at least) two purposes: first it would undoubtedly verify a number of the rules and suggestions that have been proffered and utilized over the years. Second, however, it might provide some surprises, showing that some intuitively reasonable rules and suggestions are, in fact inimical to the goals for which they were designed - or, that while optimal for achieving one goal, may be suboptimal for achieving other goals. In what follows, I will try to explain what I mean by this.

An Example

Suppose I am presenting a talk describing my research that is, as usual, profusely illustrated with slides. What are the goals I'm trying to achieve by using the slides, and how should I design the slides so as to optimally achieve these goals?

Let's start by considering more generally what I'm trying to accomplish in the talk. If I'm a typical speaker, I'm (1) trying to make my audience understand what I've accomplished in my research, (2) trying to make the audience interested in my research, (3) perhaps trying to entertain the audience (i.e., keep them from falling asleep), (4) getting across as much information as possible in the limited amount of time I've been allotted , and (5) ensuring that the audience remembers what I've said after the talk is finished.. Note that some of these goals compete with one another: for instance getting across as much information as possible is, as we've all discovered many times, inconsistent with optimizing understanding.

Slide-Design Goals

Presumably the goals I have in mind when I design my slides are isomorphic to these more general goals. So, for instance, to make the members of my audience understand, I'd make my slides as clear, simple, and direct as possible. To try- to get them interested in my topic material, perhaps I'd use the slides for a demonstration of whatever perceptual or memory effect is the focus of the research I'm talking about. To try to keep them entertained, perhaps I'd make the slides colorful and jazzy, or I'd sporadically include slides that, while interesting, are not central to the talk's main focus (for example, one of my colleagues routinely includes photographs of individuals whose research she alludes to in her talks). To keep the information -transmission rate as high as possible, I'd put as much information as possible on each slide. To get my audience to remember the information later on, I'd be redundant: I'd present the key slides multiple times throughout the presentation.

Like the general goals, the corresponding slide-design goals are, alas, mutually incompatible to a significant degree. For example, putting as much information as possible on a slide, or providing slides that are tangential to the main talk, or making a slide jazzy and colorful is often contrary making a slide clear, simple, and direct. I think that the reason that many bad talks are bad is that, when designing the slides, the speakers were trying to accomplish all the goals at the same time without prioritizing them. The impossibility of so doing results in a miserable mishmash in which none of the goals are accomplished.

If one accepts these arguments, then there are several lessons to be learned. The first is that one should think carefully about what one's goals are in designing and presenting slides. This point seems trivial and trite; however as I have noted a couple of times, it is ignored to a rather surprising degree. At least in the field of Psychology, I'd judge that in at least 50% of the talks I see, the slides could probably have been better designed by my pet cats.

Theory

The second lesson is that optimization works best when you have a theory to guide what you're doing. What such a theory could accomplish depends, of course, both on the breadth and sophistication of the theory itself, and on the precision with which one can define and organize one's goals. The latter is beyond the scope of this paper. I would, however, like to list some possible ways in which ongoing work in various subfields of cognitive psychology could potentially contribute to decisions about how slides should be designed given that the goals are reasonably well-specified. This list is not by any means meant to be exhaustive; it is only meant to be illustrative.

Degradation

Earlier, I briefly described one of my research projects designed to investigate the effects of stimulus degradation on perception and memory. Theory issuing from this and related research could guide the design of both individual slides and slide sequencing given potentially degrading circumstances. Suppose, for instance, I know that during an upcoming talk the room lights will have to be left on, e.g., to allow note taking (thereby leading to contrast reduction in the slides). What, if anything, should I do to compensate? One possibility is that I can design my slides (taking font size and contrast into account) such that room light won't have a noticeable effect. What if this isn't possible? Then I have to decide how to compensate for the contrast reduction. If such reduction merely slows down processing, then my strategy is to increase time per slide. If, however, contrast reduction qualitatively changes the potentially acquirable information from the slide, then I must compensate in some other way, e.g., by changing the manner in which I describe the slide while it is present.

Information Density

How much information should be placed on a given slide? One is tempted to put as much information as possible on the slide for two reasons: first because the more information per slide, the fewer slides you need, and second the more information per slide, the less inter-slide information integration your audience has to carry out, and the less will be the corresponding memory demands. On the other hand, putting more information per slide has costs: there is more intraslide information integration to be carried out, and the information itself (e.g., the typeface) must be physically smaller, which makes it harder and slower to read. Thus, there is, in theory, some *optimal* amount of information per slide. What this optimum is can only be specified within the context of some specific theory.

Font Design

With the advent of WYSIWYG displays and laser printers, has come an avalanche of interest in font design. Again an interesting tradeoff becomes apparent which is as follows. In principle, one would expect that the more different are the individual letters of a font from one another, the more readable they would be. However, the more different the letters are, the more information is required in specifying the font. A critical ingredient in the resolution of this tradeoff is a collection of questions about what information is used by the perceptual and cognitive systems for letter perception and what isn't. These questions can be asked on several different levels, for instance:

1. What information is lost before it gets to the retina simply as a result of the eye's optical resolution?

2. What information that makes it to the retina is actually used in any way for reading?

3. What are the different ways in which information in letters is used during reading? For instance, different features of letters and words are used for actual information acquisition on the one hand, vs determining where the next fixation should be on the other. This mean, for instance, that different information is important during the reading of long passages of text vs reading of single words or short phrases.

Perception and Memory

A great deal of research has focussed on the amount of time that is required for perceiving and remembering visually-presented material. Theory issuing from such research bears on three issues in graphical display:

1. How much time is needed to perceive the necessary information in a single slide?

2. How much time is required to perceive and organize the information from a single slide such that this information can be adequately retrieved and utilized later in the same talk when it is needed in order to understand some subsequent slide?

3. How much time is required to perceive and organize both single slides and also slide sequences such that the information conveyed in the slide sequences can be remembered after the talk is finished?

Position Statement on Visualization of the Quality of Spatial Information

A personal perceptive

John W. Kick
U.S.D.A. Soil Conservation Service
Syracuse/Buffalo, New York

The Soil Conservation Service strategy for implementing GIS technology is to put GIS tools in the hands of local decisions makers. This will be done by making GIS as transparent as possible. It is recognized that not all of the powerful features of GIS will be made available to all users. Several front ends to GIS will allow access to GIS to produce products and reports frequently needed by the local decision makers.

It would follow that the use of visualizing data quality should also be transparent to the user, to communicate intuitively the aspect of quality. The use of colors (intensity), graphic defaults (fuzziness, colors), and system imposed limitations of data manipulation, display, and analysis all have potential for visually communicating data quality. The latter would require an expert system type of environment for implementation.

The issue of communicating data quality is a concern in the Soil Conservation Service, but traditionally, priority is given to data collection and mapping. These activities are more easily measured when in a production mode. GIS technology is forcing the agency to address questions of data quality.

There are many aspects of data quality that are a part of mapping and data collection. These activities are often cited as both an art and a science. The "art" part of data are judged in relative, subjective terms. Those doing the mapping have a "feel" for the quality of the information. The quality of data are accessed by subjective pass-fail tests.

The development of qualitative statements of data and the quality of the information are seen as beyond the mission of mapping activities. These issues are seen as basic research. With the advent of present technologies the integration of other data sources makes the issue even more relevant for the need for basic research.

There is a reluctance on the part of research institutions to critically look at the issue of data quality of an agencies data base, and there is a reluctance of agencies to have their data critically examined. This is probably politically motivated. Soil mapping which is a result of art and science, may not stand up well to stringent statistical examination. The data was not intended to withstand such examination, but is intended to guide land use decisions. Visualization of data quality should take this into account, but there is certainly a need for information to be used as intended. Agencies are frequently skeptical of another agencies data. Understanding the data and communicating the data quality would make more efficient use of data between agencies.

Critically looking at data quality is seen as an issue which is not a money making issue. Mistakes due to databases may have negative economic consequences that may only become evident in legal litigation resulting from inappropriate use of data.

Applications utilizing hardcopy maps allows the data to be messaged by experts, problems addressed and caveats added when needed. With GIS technology the manipulation of data occurs so quickly that results may not readily be understood.

Data quality are often communicated in narrative terms, experience has shown that it has not been a successful method of communicating data quality. Soil maps are used as "stand alone" sources of data, with users not knowing how soil polygons are defined. Attempts have been made in soil survey to make statistical statements from random transects, but these attempts have not been accepted and methods of communicating the statistical information have not been pursued. The attempts at statistical descriptions have been made even more difficult due to the need to make subjective judgements on data used in analysis. The use of block diagrams or 3-dimensional representations of landscapes have been used to aid in the description of data quality.

There is a need to have a framework or strategy for guiding the degree of spatial accuracy and precision for a particular application. There are great differences of quality required of spatial data that are used for guiding air traffic through rugged terrain, or navigating a ship through narrow channels, and that needed for forest inventory, or water quality planning. The technology may allow the precise measurement and delineation, but to impart high degrees of precision to databases where it is not needed requires unnecessary inputs of resources.

What the limits of quality are for particular databases and applications are not well understood. Taking soil surveys as an example, soil boundaries are delineated in the field, sketched with a pencil on non-rectified aerial photography. The boundaries are transferred by a visual cartometric technique or map compilation onto a rectified photobase. This transferring process is a judgement process that is difficult to quantify. The soil polygons are then digitized with a digitizing tablet with a tolerance specification of .005

m1s.. There many are differences in magnitude of possible sources of error. The entire process of data development needs to be analyzed from start to finish, with sources of errors identified.

Natural resource spatial data, such as soils, geology, land use, and land cover, require delineating polygons that contain features that are not included in the name of the polygon. There are guides as to how to make decisions to combine dissimilar soils, or how to handle areas too small to map at the desired scale. The decision to separate one type of polygon from another are also guided by sets of rules. These rules help the mapper to decide where to place boundaries between polygons. These rules are subject to subjective judgement. Methods of determining land use cover using an image processing system instead of manual interpretation may minimize subjective judgements, but there are always places where human judgements are made.

Possible Visualization Techniques

The use of 3-dimensional models (block diagrams) to represent components of polygons that can not be shown, due to scale, can communicate visually that which as been traditionally in text form.

Add attributes of data quality to databases which can be related back to spatial components. The attributes may be 3-dimensional, which could be draped over graphic displays of spatial data. The attributes could be 2-dimensional which could be toggled between the spatial data and the data quality. The use of colors, color intensity, or fuzziness could be used to show relative differences of quality.

Many graphics used in GIS applications use thematic maps to represent spatial data. Natural resource managers are accustomed to working with aerial photography as a base for carrying out planning applications. The photographic background aids in the guiding of decisions. This is a visual mechanism to guide the decision making process. If an agricultural field is the area of concern, the photographic background identifies the field boundaries much more intuitively than a thematic background. The use of such backgrounds such as digital orthophotography, and digital satellite imagery should be investigated as to how it influences the quality of decisions compared to other graphic displays. The Soil Conservation Service plans to use digital orthophotography as a base to digitized off the CRT, features such as field boundaries. It is thought that this method would be more accurate than compiling field boundaries on a planimetric base and then digitizing or scanning.

In the same way that aerial photographs more closely approximates the model of the real world, so does the use of digital elevation models. The use of 3-dimensional images of landscapes can be used to aid in the interpretation of image quality. If soil data in an agricultural field has a short steep slope not represented by the soil datum, having the image in 3-dimension, the short steep slope would be obvious to the user. Mechanisms could be built in to highlight such inconsistencies between databases.

Manipulating the scale of databases in GIS can quickly result in data being used in ways not intended. Spatial data has a range of scale that the information can be used reliably. Detailed soils are mapped at scales ranging from 1:15840 to 1:24000. Displaying and manipulating soil maps at a scale of 1:7900 is commonly done for farm planning without the user realizing that the interpretive value may be compromised. Graphic defaults could be used to communicate the potential problems.

GIS technology is relatively new in the Soil Conservation Service. The technology is not embraced equally throughout all levels of the organization, and program activities. Intuitively communicating quality, and just what is happening in the GIS to the data, will facilitate the acceptance of the technology, and allow the greatest benefit to be gained.

Visualization of Data Uncertainty: Representational Issues

working paper for the June, 1991
NCGIA Specialist Meeting on Visualizing the Quality of Spatial Data

Alan M. MacEachren
Department of Geography
302 Walker
The Pennsylvania State University
University Park, PA 16802
e-mail: NYB@PSUVM

Data quality is a critical issue in geographic visualization due to the tendency of most people to treat both maps and computer derived analyses as somehow less fallible than the humans who make decisions they are based upon. When a GIS is used to compile, analyze, and display information, the chance for unacceptable or variable data quality is high due to the merging of multiple data layers.

There is a strong tradition in cartography of attention to data quality. Only rudimentary steps, however, have been made thus far to deal with the complex issues of visualizing data quality for multidimensional data displays used in image analysis and GIS applications. The importance of pursuing the topic is evinced by the decision of the National Center for Geographic Information and Analysis to make visualization of data quality the first visualization initiative undertaken by the Center as well as by the recent attention of the Environmental Protection Agency to representing uncertainty (Rejeski and Kapuscinski, 1990).

As I have pointed out elsewhere, the term *visualization* has a number of definitions (MacEachren and Ganter, 1990; MacEachren, et. al., 1991). It is important to know what we mean when it is used. My personal perspective is that visualization is, first and foremost, an act of cognition. It is a human ability to develop mental images, often of relationships that have no visible form. This ability can be facilitated and augmented by use of tools that produce visible representations. These representations allow our visual and cognitive processes to focus on the patterns depicted rather than on mentally generating those patterns. I believe, therefore, that this NCGIA initiative should include attention to the cognitive issues of what it means to understand attribute, spatial, and temporal "quality" and the implications of this "understanding" on decision making as well as to the symbolization issues of how to represent these aspects of quality and the methodological, technical, and ergonomic issues of generating displays and creating interfaces that work. We of course need to understand how to assess and measure quality before we can represent it. This issue, however, is the focus of NCGIA initiative number 1.

The first question I had to ask myself when preparing these brief comments was just what is meant by data quality? Kate Beard and Barbara Buttenfield (1991) indicate that quality of spatial information "relates to accuracy, error, consistency, and reliability." These aspects of quality are meant to apply to more than locational verity. It is useful to begin with the framework of the Proposed Digital Cartographic

Data Standard (Moellering, 1988), incorporating locational accuracy, attribute accuracy, logical consistency (i.e., a data structure whose topology makes sense), completeness (comprehensive data and systematic ways of dealing with missing values) and finally lineage. I would argue that these categories are important, but to use a GIS effectively for either scientific inquiry or policy formulation, we need to broaden our scope. In risk assessment circles, the term **uncertainty** has gained some acceptance and I suggest that we might be better off if we follow their lead (Morgan and Henrion, 1990). We never know the precise amount of error in any particular data object -- or we would correct the error. We are more - or less - uncertain about the available characterization of particular data objects. From this perspective alone, the term *uncertainty* might be a better description of what we have been calling quality. In addition, however, I feel that uncertainty includes something of importance beyond the narrow definition of quality that the Initiative seems to be directed toward. Let me use a brief example to illustrate the difference between a focus on quality and on uncertainty and why I think it is the latter that should guide our efforts.

Imagine a single census block in a city. You have sent an enumerator out to take the census. In this particular case, the response rate is 90%. In data quality terms, we might say that our population and income information for this block is of less than perfect quality because of the lack of "completeness" in the data. Further, there may be "attribute inaccuracy" in the data collected due to misunderstanding of the survey questions or deliberate misinformation about items such as income or education, or "spatial inaccuracy" due to address coding errors by the census enumerator. If, in the adjacent census block we somehow achieved 100% participation in the census, everyone understood the questions and gave truthful responses, and the enumerator made no mistakes, a data quality assessment would label that unit's data as perfect. What we will be leaving out of this assessment is the issue of variability (over both space and time and within categories).

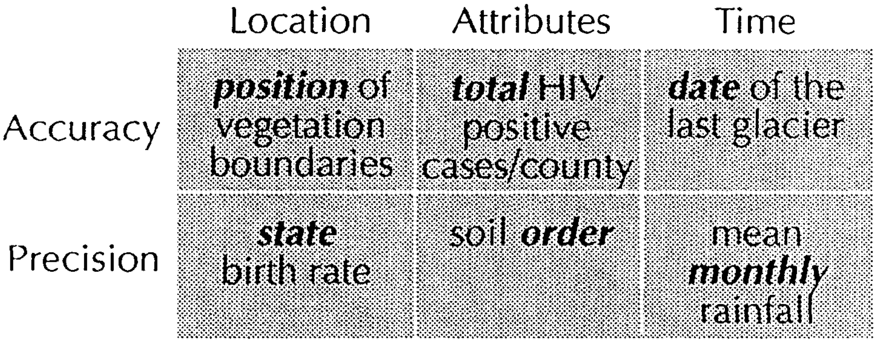
All data are categorized. In some cases, the categories might be quite narrow in relation to the problem investigated. For example, temperatures might be measured to the nearest degree, or even 1/10 of a degree. Most data in a GIS, however, will be grouped into much broader categories (e.g., soil classifications, income brackets, whether a house has indoor plumbing or not, etc.). In all of these cases, the categorization introduces uncertainty even when the data are of high quality.

We can only be certain that a particular location -- a particular data object -fits somewhere within the attribute bounds of the categories and the spatial bounds of the enumeration unit to which it is aggregated. The aggregate totals for our census blocks disguise the variability within those census blocks. Our level of uncertainty about those aggregate values will be a function not only of the quality of values (as defined above), but of statistical variance around the mean values we typically use to represent the unit, and of spatial variance within the unit.

In addition to data quality and variability, a final uncertainty to be dealt with is temporal. The data, even if accurate and homogeneous, represent a snapshot at one point in time. Our uncertainty about their veracity will increase with both the time separating us from the event being studied and with the temporal window through which we view the data.

When we use a GIS, the important issue is quality of the decisions we make (e.g., about a research course to follow or an urban development policy to impose). Whether we use the term data quality or data uncertainty matters less than whether the tool we give the GIS user is adequate for deciding how much faith to put in any particular piece of information extracted from the database. The main initial point that I want to make, then, is that we can have highly accurate data while still having imprecise data. This lack of precision is at least as important an issue as a lack of accuracy.

One initial research question should be whether visualizing accuracy and precision require different visualization tools and whether similar visualization tools can be applied to representation of uncertainty about time, space, and attribute categorization. The figure below provides examples of data about which we are likely to have each kind of uncertainty.

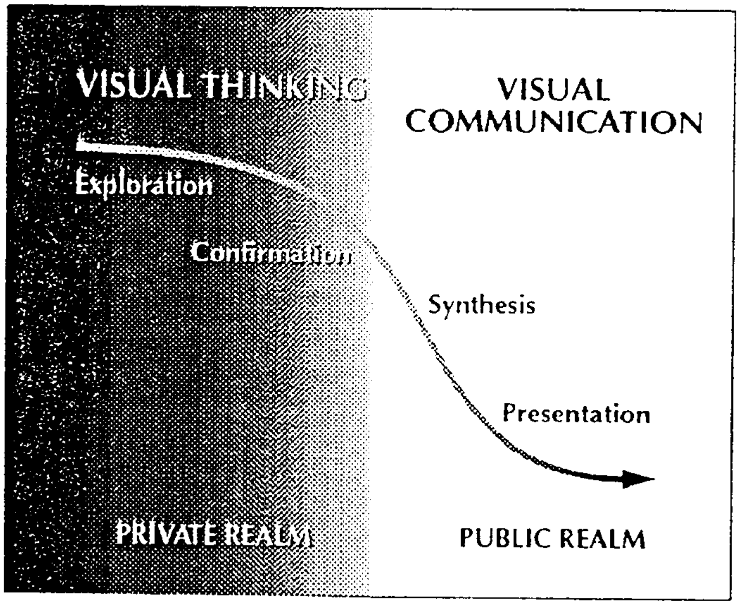


	Location	Attributes	Time
Accuracy	***position*** of vegetation boundaries	***total*** HIV positive cases/county	***date*** of the last glacier
Precision	***state*** birth rate	soil ***order***	mean ***monthly*** rainfall

REPRESENTATIONAL ISSUES

Varied goals and needs - categories of interaction with data

If we continue to attack cartographic questions with our communication model visors on, we will fail to take advantage of the power that GIS and visualization tools provide. The search for the "optimal" data quality visualization tool might prove as fruitless as the search for the optimal graduated circle map. It is critical to recognize that GIS and visualization tools attached to them are used for a range of problem types that may have quite different visualization needs in general and visualization quality needs specifically. DiBiase (1990) recently developed a graphic model of the range of uses to which graphics might be put in scientific research. I believe, that his basic model is relevant, not only to science, but to applied spatial decision making with a GIS.



A second important question for the research agenda is how visualization uncertainty tools will differ across this range (from the use of GIS by an EPA scientist exploring the spatial distribution of a pollutant to the use of GIS for emergency vehicle routing).

Graphic variables - logical symbol-referent relationships

Because many or most GIS users are not trained in cartographic symbolization and design, it will be necessary to design expert systems that translate information into graphic display form in logical ways. Jacques Bertin, the French cartographer/ graphic theorist, has had a tremendous impact on our approach to the problem. The Robinson, et. al., text (1984) (that, according to Fryman's (1990) survey in the latest *Cartographic Perspectives*, is used by 50% of introductory cartography courses in the country) cites Bertin's basic system of graphic variables as the fundamental units we can use to build a map image. Weibel and Buttenfield (1988) in their paper on map design for GIS and Muller and Zeshen (1990) in a paper on expert systems for map design, both accept this system as a base to build from in designing expert systems for map symbolization.

A second representation issue for visualization of data quality, therefore, is how Bertin's graphic variables (with possible additions or modifications) might be logically matched with different kinds of data uncertainty. One additional variable, beyond Bertin's original seven, that we are looking into at Penn State is "focus." [This variable appears to have been originally suggested by David Woodward in a seminar at Wisconsin.] Presenting data " out of focus," or simply at lower spatial resolution, (as you would see it with an out of focus camera) might be an ideal way to depict uncertainty.

Linking visualization tools to models of uncertainty.

Different uncertainty representation issues will arise when dealing with different kinds of data (e.g., qualitative data on land use/land cover versus quantitative data from the census). When data are quantities aggregated to units such as counties, we should consider the spatial characteristics of the phenomena represented by these quantities as we select symbolization methods to depict the uncertainty about them. A distinction can be made about the nature of quantitative distributions between discrete and continuous phenomena. Both stepped and smooth continuous functions are possible. From a practical point of view, the distinction between discrete and continuous is best treated as a continuum because many phenomena exhibit continuity at some scales but not others. A second continuum relates to *character* of variation in the phenomenon across space. Some phenomena (e.g., tax rates) can vary quite abruptly as political boundaries are crossed while others (e.g., average farm size in Kansas) can exhibit a relatively smooth variation quite independent of the units to which data are aggregated. Following George Jenks' (1967) lead, David DiBiase and I have proposed a series of data models that represent locations in this continuity-abruptness phenomena space (MacEachren and DiBiase, 1991).

Three research questions suggest themselves here: a) is it safe to assume that the spatial characteristics of uncertainty will mimic those of the data that uncertainty is being estimated for, b) do particular symbolization methods actually communicate the particular spatial characteristics that we as cartographers associate them with (e.g., is a layer tinted isarithmic map depicting uncertainty in air pollution estimates interpreted as a smooth continuous surface or as discrete regions) and c) what approach should be followed when a data set has multiple kinds of uncertainty associated with it and the spatial characteristics of that uncertainty vary.

User interfaces - How to merge data and uncertainty representations

Beyond the basic issue of how to represent uncertainty is the question of how and when to present the representation. This is complicated by the likelihood that GIS representations are often products of a combination of measured and model-derived multivariate data. There seem to be three choices that could be used separately or in combination:

a) map pairs in which a data map is depicted side-by-side with a map of uncertainty about that data

b) sequential presentation in which a user might be warned about uncertainty with an initial map which is followed by a map of the data. Alternatively, the user could be given a tool that allows toggling between the data and the uncertainty representations

c) bi-variate maps in which both the data of interest and the uncertainty estimate are incorporated in the same map.

Most attempts that I have seen thus far to graphically depict spatial data uncertainty have used the map pair strategy. Cartographers have spent relatively little time investigating the impact of sequential information presentation. Possibilities of interactive mapping and GIS, however, have recently attracted several of us to the issue of sequential information presentation MacEachren, 1989; Slocum 1988). One clear avenue to explore here is the potential of hypertext to allow user's to navigate through the potential maze of data and uncertainty representations that we might be able to provide them with.

In relation to the third possibility, bi-variate maps, the census Bureau's bivariate choropleth maps from the 1970 census are perhaps the best known example. Experimentation with those maps by several researchers indicates that untrained readers have considerable trouble reading bi-variate maps (e.g., Olson, 1981). There are, however, a number of bi-variate mapping possibilities that have not yet been investigated. One of our current graduate students, Alan Brenner, for example, is investigating the use of color saturation (or intensity) as a graphic variable for depicting uncertainty on maps in which hue or value is used to represent the data values of interest. The variable of focus might be used in similar ways. Another possibility might be to combine sequencing and bi-variate techniques and allow a fade from a data map, through an uncertainty map, back to the data map.

Evaluation of the utility or affect of providing uncertainty information

This leads me to my final point. The representation of uncertainty about information in a GIS provides a unique opportunity to determine whether our efforts at map symbolization and design research over the past 40 years have provided the tools required to develop a representation system. If past perceptual and cognitive research along with the conceptual models of symbol-referent relationships based on semeiotics are really useful, we should be able to draw on the conceptual models and use past experience to design appropriate experiments in our quest for answers about representing uncertainty.

This possibility may tempt some of us to go back to our roots in the communication model approach to cartography. Communication of data quality or uncertainty seems to be the ideal case for which the communication model was developed. Uncertainty can be treated as a precisely defined piece of information that we want a GIS user to obtain. I am afraid, however, that if we follow this narrow information processing communication model approach we will hit the same dead ends that we found a decade or so ago.

This time around we need to be aware of the range of human-user interactions with graphics that occur from initial data exploration to presentation, the fact that even seemingly *precise* data quality information is conditioned by the social-cultural context in which decisions about what to represent are made, and our limited ability as cartographers to determine the relative importance of various kinds of quality or uncertainty information in a particular context. Maps are REpresentations and as such are always one choice among many about how and what to REpresent. There is always uncertainty in the choice of representation and representation method, therefore, representing the uncertainty in our representations is an uncertain endeavor at best.

Acknowledgments

I would like to thank David DiBiase and Tony Williams for suggestions on an earlier draft of this paper and Alan Brenner for sharing ideas on bi-variate maps of uncertainty.

References

Buttenfield, Barbara P. and Beard, M. Kate 1991. Visualizing the quality of spatial information. Technical Papers 1991 ACSM-ASPRS Annual Convention, Volume 6, Auto-Carto 10. 423-427.

DiBiase, David W. 1990. Scientific visualization in the earth sciences, Earth and Mineral Sciences, (Bulletin of the College of Earth and Mineral Sciences, The Pennsylvania State University, 59(2): 13-18.

Fryman, James F. and Sines, Bonnie R. 1991. Anatomy of the introductory cartography course, Cartographic Perspectives, 8: 4-10.

Jenks, George F. 1967. The data model concept in statistical mapping. International Yearbook of Cartography, 7, 186-190.

MacEachren, Alan (in collaboration with Barbara Buttenfield, James Campbell, David DiBiase, and Mark Monmonier) Visualization, in Abler, Marcus, and Olson (eds.) Geography's Inner Worlds (in press).

MacEachren, Alan M. 1989. Learning a City Using an Interactive Map: A Comparison of Route Versus Landmark Based Learning paper presented at the 14th International Cartographic Conference, Budapest, Hungary, August 17-24.

MacEachren, Alan M. and DiBiase, David W. 1991. Animated maps of aggregated data: Conceptual and practical problems, Cartography and Geographic Information Systems (in press).

MacEachren, Alan M. and Ganter, John H. 1990. A pattern identification approach to cartographic visualization. Cartographica, 27(2): 64-81.

Moellering, H., Fritz, L., Nyerges, T., Liles, B., Chrisman, N., Poeppelmeier, C., Schmidt, W. and Rugg, R. (executive committee) 1988. The Proposed National Standard for Digital Cartographic Data, The American Cartographer, 150): 9-142.

Morgan, M. Granger and Henrion, Max 1990. Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis, Cambridge: Cambridge University Press.

Muller, Jean-Claude and Zeshen, 1990. A knowledge based system for cartographic symbol design, The Cartographic journal, 27(2): 24-30.

Olson, Judy M. 1981. Spectrally encoded two-variable maps. Annals of the Association of American Geographers, 71(2): 259-276.

Slocum, Terry A. 1988. Developing an information system for choropleth maps. Proceedings of the Third International Symposium on Spatial Data Handling, August 17-19, 1988, Sidney, Australia, 293-305.

Rejeski, David and Kapuscinski, Jacques 1990. Risk modeling with geographic information systems: Approachcs and Issues, Report of the U.S. Environmental Protection Agency, Office of Information Resources Management.

Robinson, A. H., Sale, R. D., Morrison, J. L., and Muelircke, P. C. 1984. Elements of Cartography, fifth edition, New York: John Wiley & Sons.

Weibel, Robert and Buttenfield, Barbara P. 1988. Map Design for Geographic Information Systems, GIS/LIS'88 Proceedings: Assessing the World, Volume 1, San Antonio: ACSM, ASPRS, AAG, URISA, 350-359.

Storing Data Reliability Information within the Digital Chart of the World Database

Robert J. Maki
Environmental Systems Research Institute, Inc.
Redlands, California

Background

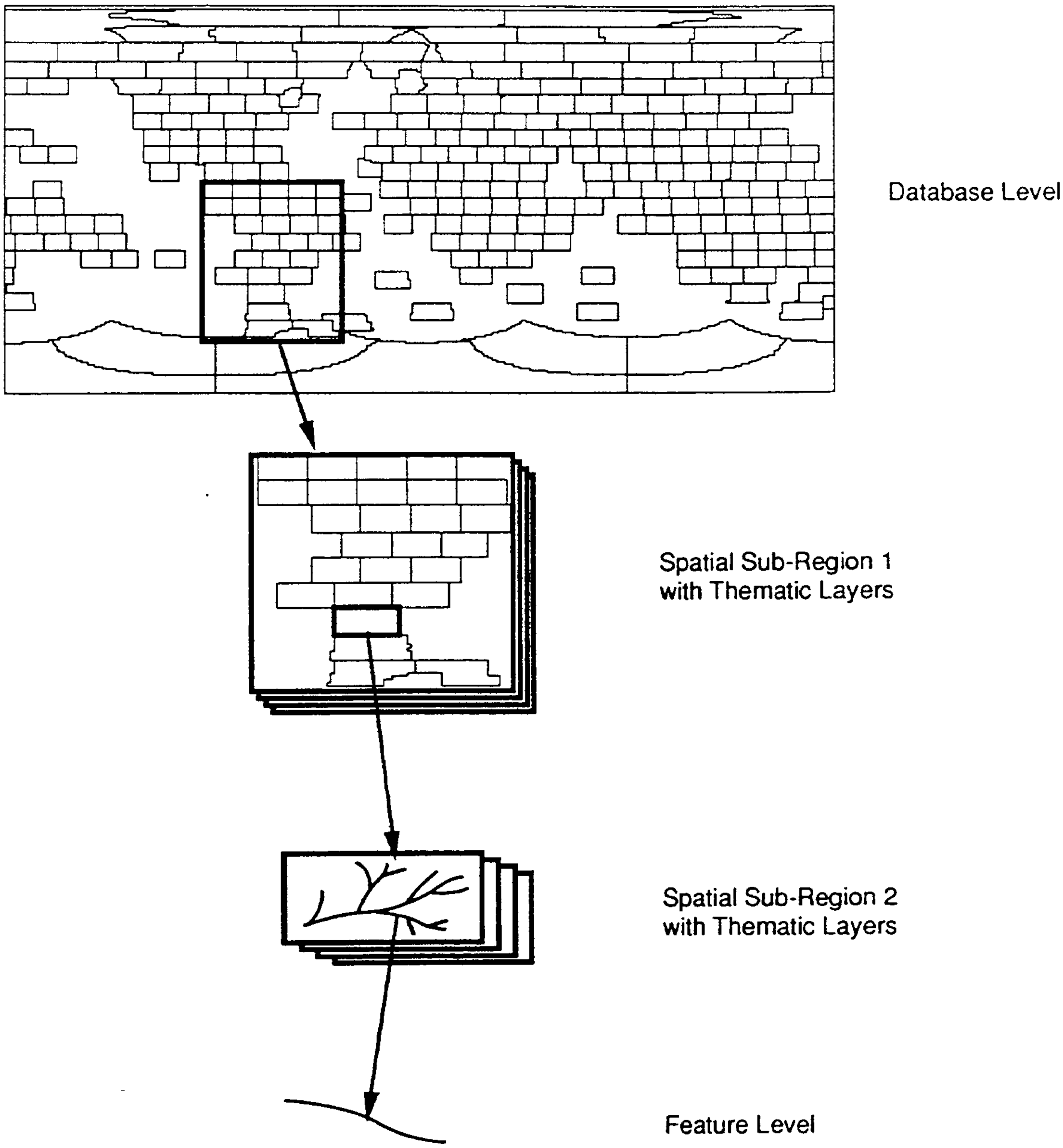
In September of 1988, Environmental Systems Research Institute, Inc. (ESRI) was awarded a contract by the Defense Mapping Agency (DMA) to create a suite of products and associated standards to support DMA's digital Mapping, Charting, and Geodesy (MC&G) program. An important component of the contract is the development of a 1: 1,000,000 scale database of the land area of the planet based on DMA's Operational Navigational Chart (ONC) series. The product is commonly referred to as the Digital Chart of the World (DCW). In conjunction with this effort, ESRI has been directed to investigate techniques by which data reliability and digital marginalia information can be stored within the database. From this requirement has emerged a perspective on the nature of reliability information within geographic databases and some of the factors affecting its storage.

Data Reliability Encoding Based on Spatial and Thematic Aggregations

Reliability information within geographic databases exists on multiple levels. At the most detailed level, it can be expressed as feature attributes such as positional or categorical accuracy (e.g. Guptill, 1989). On the most general level, reliability characteristics can be expressed as broad attributes across a database, such as data resolution, processing tolerances, or particular processing techniques. There may also be any number of intermediate levels that relate to the database's origins or maintenance strategies. These levels can have both spatial and thematic characteristics. Figure 1 illustrates this concept. At its highest level, the database may be considered as a whole, with characteristics that are common to all aspects of the database. Within the larger database may exist regions that have special characteristics that should be noted (e.g. related to currency or update status). Regions may be further subdivided based on convenient spatial units such as tiles or collections of tiles within a library. Regions may also be based on data processing artifacts (e.g. hard copy source map borders). Ultimately, this scheme depends on how the database is structured. Thematic-based divisions can assert themselves at any level in the spatial aggregation scheme, introducing additional complexities to the overall scheme.

The most exhaustive strategy to storing data reliability is to store all known information about a feature as feature level attributes. This approach of complete geographic disaggregation has the advantage of maintaining close ties between objects and reliability characteristics that will be maintained until a user explicitly removes them from the database. The disadvantages are storage efficiency, data redundancy, and the high cost of maintaining large numbers of attributes. A more practical approach may be to identify a manageable number of discrete levels within the database at which reliability characteristics change, and concentrate efforts on maintaining this information at the highest level of aggregation available. This approach minimizes the reliability information update effort and the data redundancy problem. Spatial and thematic-based reliability changes can be managed using a reliability overlay; an idea whose analog equivalent is the reliability diagram found on many maps (Chrisman, et al., 1987). The approach also has clear disadvantages. Once information is divorced from the feature or from a commonly used collection of features (such as a tile), it can be quickly lost. One is also faced with the problem of identifying the means of storing information at awkward intermediate levels within the database. Digital reliability diagrams could be difficult to maintain, especially when the added dimension of thematic-based reliability is introduced.

Figure 1: Spatial and Thematic Aggregation Levels



Problem Statement

The DCW source material are positive film separates reproduced from original negatives used in the ONC lithographic production process. The separates are scanned and the data are subject to a raster to vector conversion process. Afterward, the graphic and attribute portions of the database are processed in ARC/INFO and translated into their final form in preparation for encoding on CD-ROM media.

The capture and storage of reliability information within the DCW is shaped by a number of conditions. An important factor related to the availability of reliability information is that the database is derived from a single source, limiting reliability indices to information represented on the source and to data inconsistencies observed during the database development effort. Information on the source charts is depicted in a variety of ways, including line and point feature symbology, textual descriptions related to definite regions and features, and sheet-based marginalia. Data inconsistencies on the source are also apparent, with instances of poor registration between positive source separates and incorrect feature attributes being common.

Computer storage requirements are dictated by the sheer data volume involved (in excess of two gigabytes), necessitating that a comprehensive data storage strategy be developed. This has taken the form of a multi-tiered library structure devised within the Vector Product Format (VPF), a data storage structure developed in conjunction with the contract in a parallel effort (DOD 199 1). Finally, and perhaps most significantly, the labor costs associated with encoding this information must be considered, being in some ways an adjunct to the task of creating a database with a high degree of fidelity to the source materials.

The DCW Approach to Storing Reliability Information

The DCW will exist as a multi-tiered database within VPF. Reliability information will be stored at three distinct spatial levels within the database, depending upon information availability, ease of encoding, and spatial extent of the information. These levels are: 1) library, 2) source map processing module, and 3) feature.

The library level is used to store the most general information about the data. Two libraries exist within the DCW; one containing the primary 1: 1,000,000 scale data, and the other containing highly generalized international boundary data to be used for certain display functions with the DCW applications/display software. The primary mechanism for storing reliability information at this level is the VPF data structure. The VPF standard nominally follows the recommendations of the National Committee for Digital Cartographic Data Standards Data Quality Working Group (NCDCDS, 1988) by assigning specific fields within the data structure for positional accuracy, attribute accuracy, and logical consistency; although the category for lineage is superseded by a series of more specific fields, and "completeness" has been separated into graphic and attribute components (see Appendix 1).

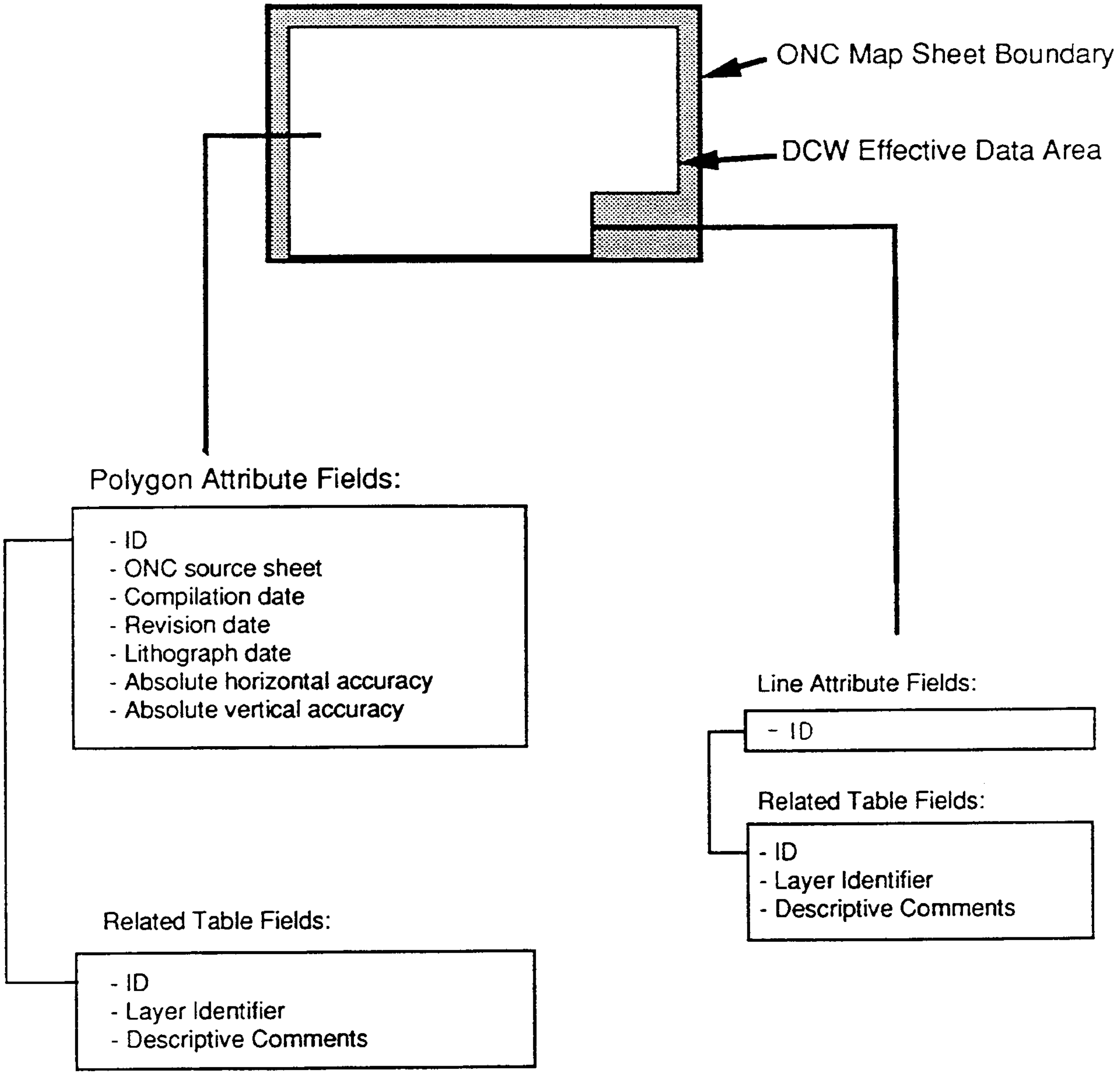
VPF is flexible enough to incorporate lineage information as a related file on any level. In the DCW implementation of VPF, all of the standard fields will be implemented to the fullest extent possible on the library level; although alternate methods are being used on lower levels for reasons that will be discussed later. In addition, information will be provided as to standard library characteristics with respect to data resolution, processing tolerances, and decision rules enacted during the database development process.

The second highest spatial data reliability level within DCW is the processing module level. Although the concept of the processing module quickly disappears in a tiled database, particularly one with cross-tile topological links as are present in the DCW implementation of VPF, data origins remain a key reliability characteristic. Processing modules correspond to the effective data areas being extracted from the individual ONC sheets. Considering the amount of data variability that exists on a sheet-by-sheet basis, it is a natural that the sheet areas be tracked as separate database components. To accommodate this, the DCW design team created a separate database layer containing former processing module borders (the Data Quality layer). When merged into the final database, this layer will essentially become a single large data reliability diagram. The layer will have both polygonal and line attributes (Figure 2). Polygon attributes will include: source sheet identifier, original compilation data, revision date, lithograph date, and absolute horizontal and vertical accuracy. Additionally, the sheet boundary polygons will have a related table for miscellaneous comments with an additional field keyed to DCW layer name. Line attributes will be comments in a related table also keyed to layer name; in this case, dedicated to documenting edgematching problems between source sheets. After tiling, some tiles will have fallen wholly within processing modules, while others will contain multiple former processing module boundaries, effectively transferring reliability area and boundary features to individual tile coverages. The Data Quality layer also provides the database user with an interpretive tool to account for changes in observed characteristics between former map sheets. For example, a user may observe an unusual number of data inconsistencies (such as feature density) within a well-defined zone of the database. In this case, the user could view the data of interest in conjunction with the former processing module boundaries in the Data Quality layer to determine whether they can be attributed to differences in source map compilation strategies.

The lowest level within the DCW data reliability scheme is the feature level. This information, which is derived solely from symbology on the source maps, will be stored explicitly in the DCW database as feature attributes. For example, primary attributes for the DCW roads layer will include a status item with four of the valid values reserved for reliability type information, including "existence doubtful", and three types of ESRI-compiled connectors (Figure 3). Criteria for selecting secondary attribute codes include: 1) presence of explicit symbology on the source maps, 2) frequency of occurrence of the feature, and 3) data source. In the example presented above, the "existence doubtful" attribute was clearly symbolized and appeared with sufficient frequency as to reserve a separate attribute value for it. The ESRI-compiled data types represented either interpreted or arbitrary information added to maintain connectivity of the road network and clearly are data from a source other than the ONCs. Another candidate reliability attribute was It approximate alignment." In this case, the decision was made not to reserve a separate attribute value for it due to its infrequent occurrence on the ONCs. Rather than ignore this information altogether, it is being stored as an annotation text string placed in its original position relative to the feature on the source maps. The decision to use this approach was based upon the desire to minimize the number of attributes to be encoded, thereby helping to streamline the database development effort. In most cases, all ESRI-added information will be explicitly coded as to its interpretive status.

Conceptually, a sub-processing module level exists within the spatial hierarchy described above between levels 2 and 3 that must be treated separately since no explicit mechanism has been created to handle it. The ONC source maps contain definite and indefinite regions with reliability characteristics distinct from surrounding areas. A good example of the former case is the DCW elevation layer, which contains a significant number of "no data" areas in regions of the world where the information is not available (e.g. the Amazon Basin). Gaps in elevation data are explicitly stored during automation on the processing module level, yet these are phenomena that do not depend on artificial map sheet boundaries. As a result, "no data" polygons are a library-wide phenomena. The second sub-processing module level relates to indefinite regions. These are typically associated with problems found on the source materials during data production. A good example of an indefinite region having differing reliability characteristics relates to the registration of data between layers. ESRI has often discovered off-sets between source separates; apparently as a result of inconsistent alignment between ONC compilation and/or update sessions. It is difficult to identify the extent of regions affected by this problem, since its occurrence is sporadic and its effects are variable within any given area. In the DCW, problems of this kind will be addressed through the use of annotation text strings and approximate descriptions of the problem’s spatial extent in the Data Quality layer.

Figure 2: Data Quality Layer Design



Polygon Attribute Fields:

- ID
- ONC source sheet
- Compilation date
- Revision date
- Lithograph date
- Absolute horizontal accuracy
- Absolute vertical accuracy

Related Table Fields:

- ID
- Layer Identifier
- Descriptive Comments

Line Attribute Fields:

- ID

Related Table Fields:

- ID
- Layer Identifier
- Descriptive Comments

Figure 3: Physical Design of the DCW Roads Layer

Coverage Name: RDLINE

Line Attribute Table

Variable	Field Name
Feature ID	COVER-ID
Road Type	RDLNTYPE
Road Status	RDLNSTAT

RDLNTYPE

1 =	Dual lane (divided) highways
2 =	Primary or secondary roads or highways
3 =	Tracks, trails, or footpaths
8 =	Connector

RDLNSTAT

1 =	Functional
2 =	Under construction
3 =	Existence doubtful or reported
4 =	Compiled road connector
5 =	Data compiled from adjacent, more recent chart
6 =	Compiled, under construction
9 =	None

Annotation Requirements for Roads Layer

Annotation will be used when there is a note along a road that is not covered by the coding scheme.

Integration Requirement for Roads Layer

Roads on levees and dams will be coincident but not integrated

Discussion

Within the context of the NCGIA agenda, ESRI's DCW experience relates most strongly to the Data Models and Database Issues theme. The effort has unearthed several concepts associated with this theme, and has suggested a number of possible solutions. The first issue relates to the nature of reliability information within geographic databases. Data reliability variation can manifest itself on a variety of levels from the global to the feature specific. That being the case, the question arises as to how to best store the information. From a database development and maintenance perspective, it is advantageous to store the information in a manner that minimizes effort without losing information content. In general, this means creating well-defined "repositories" within a database where the analyst can alter or enhance information as needed. An example of this is the DCW Data Quality layer, which is a repository for all processing module-based reliability information. In the case of the DCW, the Data Quality layer could form the basis for a more complex digital reliability diagram, as users enhance the data for their own special purposes. Feature attributes can also serve as information repositories but with significantly more effort required for encoding and update.

An important factor that has not been discussed here is database applications and the nature of the user base. The DCW is expected to have wide distribution, with many of the users concentrating on specific database components (such as elevation or hydrography) outside of the VPF environment. This scenario favors the use of feature based and perhaps layer based data reliability storage schemes. VPF allows for the storage of standard reliability fields (described above) at any level within the database, including the coverage level. At the same time, information stored explicitly within the data storage structure will likely be lost in data conversion processes if no equivalent information fields exist within the target format. Regardless of the manner in which the information is stored, it is unclear how the user community would utilize it.

With respect to the NCGIA research agenda, it is clear that a comprehensive approach to the storage of data quality information must be developed that accounts for user needs, technological and practical constraints, and data structures. The user community must be educated regarding information reliability and how it affects their spatial analyses.

Standard approaches should be developed such that users become accustomed to seeing reliability storage mechanisms in their data in recognizable forms.

References

Department of Defense, 1991, Vector Product Format, Interim Military Standard, MILSTD-600006, Defense Mapping Agency.

Chnisman, N.R., Gurda, R., and Beard, M.K., 1987, National Quality Charting Standards, Land Information and Computer Graphics Facility, University of Wisconsin, Madison, Wisconsin.

Guptill, S.C., 1989, "Inclusion of accuracy data in a feature-based, object-oriented data model", Accuracy of Spatial Databases, M.F. Goodchild (ed.), pp. 91-91, Taylor and Francis, London.

National Committee for Digital Cartographic Data Standards (NCDCDS), 1988, The proposed standard for digital cartographic data. The American Cartographer, 15.

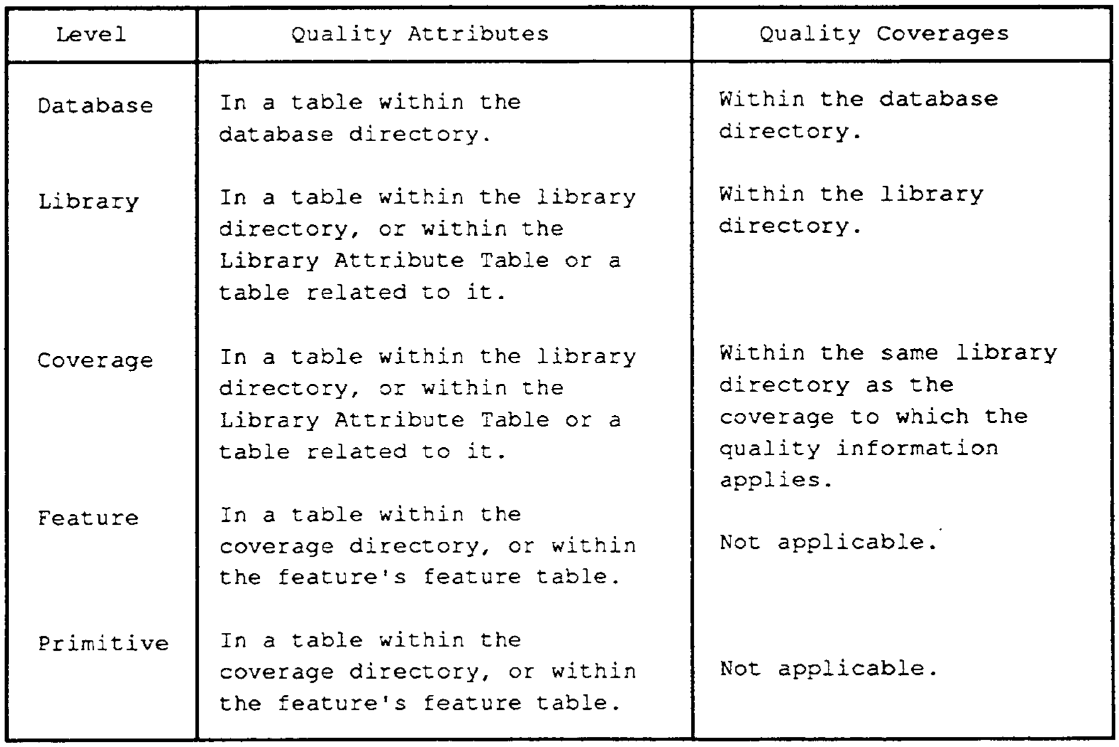
Appendix 1: VPF Data Quality Descriptive Information

5.4 Data quality. VPF contains data quality information so that users may evaluate the utility of the data for a particular application. While the exact form of data quality information supplied for a data set is specified by a product specification, certain general rules apply in the following cases:

5.4.1 Data quality hierarchy. Data quality information at the database level applies to all libraries of the database, except where those libraries contain their own data quality information of the same kind. Similarly, data quality information at the library level (which may have been inherited from the database) applies to all coverages within the library except those that contain their own data quality information of the same sort. Coverage level data quality information applies in the same manner to features. Feature level data quality information likewise applies to both the spatial primitives and attributes that compose them.

5.4.2 Data quality encoding. Data quality information is represented as attributes or as a coverage. If as attributes, it may be either added to an existing VPF table, or as an independent table residing at the appropriate level. If a coverage, it shall be a coverage of topology level 3 (polygon planar graph) whose area or complex features designate areas with uniform data quality information of specified types. TABLE 10 depicts the attribute and coverage location of data quality information through the database.

TABLE 10. VPF data quality information.



Level	Quality Attributes	Quality Coverages
Database	In a table within the database directory.	Within the database directory.
Library	In a table within the library directory, or within the Library Attribute Table or a table related to it.	Within the library directory.
Coverage	In a table within the library directory, or within the Library Attribute Table or a table related to it.	Within the same library directory as the coverage to which the quality information applies.
Feature	In a table within the coverage directory, or within the feature's feature table.	Not applicable.
Primitive	In a table within the coverage directory, or within the feature's feature table.	Not applicable.

5.4.3 Types of data quality information. There are seven types of data quality information: Source, Positional Accuracy, Attribute Accuracy, Currency, Logical Consistency, Feature Completeness, and Attribute Completeness. These are VPF's standard types of quality data. However, a product specification may call for additional types of data quality information as well. The following paragraphs define these types.

a. Source. Source describes the origin or derivation of a single feature, primitive or attribute. This includes any processing techniques applied to the data, as well as the data source.

b. Positional accuracy. Positional accuracy provides an upper bound on the deviation of coordinates in VPF from the position of the real world entity being modeled. Positional accuracy must be specified without relation to scale and shall contain all errors introduced by source documents, data capture, and processing.

c. Attribute accuracy. Attribute accuracy describes the accuracy or reliability of attribute data.

d. Currency. Currency represents the date at which the data was introduced or modified in the database. This date of entry is used as a proof of modification for a single data element, permitting statistical interpretation of groups of data elements-

e. Logical consistency. Logical consistency describes the fidelity of relationships encoded in a VPF data set. Logical consistency requires that all topological foreign keys match the appropriate primitive, that all attribute foreign keys match the appropriate primitive or features, and that all tables described in feature class scheme tables do indeed have the relationships described.

f. Feature completeness. Feature completeness indicates the degree to which all feature of a type for the area of the data set have been included.

g. Attribute completeness. Attribute completeness indicates the degree to which all attributes of a feature have been included for that feature. Actually, since this information can be derived from the feature itself, simply by counting null values, this particular form of data quality information should not need to be explicitly included.

A View on Visualizing Spatial Data Quality

Matthew McGranaghan

NCGIA	Geography Department
University of Maine	University of Hawaii
Orono, ME 04469	Honolulu, HI 96822
207/581-2117	808/956-8465
matt@ thrush.umesve.maine.edu	matt@uhunix.uhcc.hawaii.edu

Introduction

Visualizing spatial data quality requires that we develop appropriate techniques to graphically display the "goodness" of data. An obvious problem is defining data quality. For our purposes, we can assume that this has been done elsewhere, and merely note that data quality is multidimensional and dependent on intended use. Symbolizing quality, like symbolizing location and other attributes, posses no special conceptual problems. It is subject to the same constraints and needs for generalization as any other graphic communication. Once data quality is understood, it can be displayed.

One reason that we are here is that cartography often displays too much quality. In commenting on the map as a communication graphic, J. K Wright (1942, p. 528) noted that "indifference to the truth may also show itself in failure to counteract, where it would be feasible and desirable to do so, the exaggerated impression of accuracy often due to the clean-cut appearance of a map." This "exaggerated impression of accuracy" often stuns cartography students who notice the apparent authority of their work. They also notice, as did Wright (p. 542), that "an ugly map, with crude colors, careless line work, and disagreeable, poorly arranged lettering may be intrinsically as accurate as a beautiful map, but it is less likely to inspire confidence." From these observations, it follows that in order to indicate less than perfect quality in our data and to inspire appropriately less confidence in it, we need only make maps which are purposefully ugly, using crude colors, careless linework, and disagreeable, poorly arranged lettering. Ironically, there is a practical advantage to this approach to visualizing data quality: given the graphic quality of many current systems, we need not intentionally degrade the output (we only have to teach people to recognize it for what it is).

But the issue at hand involves more than simply decreasing, unwarranted confidence in maps and geographic data. We need methods to map the various aspects of data quality into graphic variables such that one might visually assess other aspects of the data in light of their quality. This requires that we consider the nature of the things about which we have data, the types and nature of the quality of that data, the graphic variables we have available to represent them, and the interactions among these. We have a vast array of means for putting these together. The challenge we face is to create ways to put these together to meet any given communication objective.

The Graphic Design Space

Designing a visual display of the quality of spatial data requires striking a balance within a design space defined by the spatial things and nature of the data describing them, our objectives in communicating about them, the visual cues available in the presentation media we use, and the limits imposed by presentation media and the visual and cognitive systems of the display user. In the following, I expand on these constraints on the design.

Spatial Data

In “Crossbreeding" Geographical Quantities" J. K. Wright (1955, p. 61) noted 11, 232 distinct classes of geographic objects, and he had not even begun to consider the number that could result if we "crossed" these with the types of uncertainty we might have about the quality of data describing those objects. It seems counter productive to extensively cross categorize geographic data. Rather, it may be sufficient to characterize a datum as to its aspect, form, level of measurement, and derivation. Aspect refers to distinctions between positional, temporal or attribute data. Form refers to distinctions between punctiform, linear or areal features (or aggregates of them, aggregates of areal samples for instance giving rise to fields). Level of measurement refers to the distinctions between nominal, ordinal, interval, or ratio scales to record data values (Stevens 1947). The notion of derivation is to suggest that there is something fundamentally different between values that we can count or measure directly and those which we can derive from counts and measurements.

Communication Objectives

Given measures of data quality to communicate, the objectives in that communication are important. What aspect(s) of data quality is the user to "visualize"? Should the display only indicate which, if any, information is "unsure"? Should it show the uses for which the quality is adequate and those for which it is inadequate? Should it indicate the amount of uncertainty about the data? Qualitative symbology can indicate that some data are suspect, or that some data are measurements while others are derived. Quantitative symbols are needed to indicate how unsure data is or the degree of goodness in a derivation.

To whom is the information to be communicated? The intended audience for the visualization is also important in designing the communication. Experts looking for patterns in the data quality, experts looking for patterns in the data where quality might influence their understanding of the data, or more casual users looking at the data have very different quality visualization needs. The idea that the data may be inaccurate may never have occurred to the very casual user, and it may or may not be appropriate for the display to slyly seduce this user into questioning the quality of the data, or into at least realizing that all data are not equally good.

The style of use is another concern. Should the display be part of a passive presentation of information, an animated one, or perhaps an interactive one in which the user takes a role in creating and defining, rather than merely reading, the display. More interaction may produce a deeper more lasting understanding of the data's error structure but may be anathema to someone quickly screening a large amount of data.

Graphic Codes

All of the communication objectives must pass through the filter of graphic symbols available to the display designer. (I am assuming that we are limited to visual presentation because of the name of this initiative. I have no objection to, and considerable interest in, expanding our charge to include all of the senses, though I confess to some apprehension about calling an initiative "sensualization of spatial data quality".) A number of authors have advanced lists of useful graphical variables and indicated the kinds of meaning they most readily carry (Bertin 1983, McCleary 1983), Tufte 1983, Robinson et al. 1984, Dent 1985, and many many others!). Differences in these visual variables are used to represent differences in the data. The list below is presented as a summary and extension of these. It is structured with a group of the most primitive marks, followed by a group of variables that are more complex, and finally a set of very complex, emergent variables that are built upon the more primitive variables.

Presence- the existential qualifier. A mark (or even the lack there of) is understood to mean something. It is very difficult to show presence without also using location, shape, size and color.

Location- where the symbol is placed. With spatial data, this is often determined by geography, and most often used to encode spatial relations.

Time- when the symbol is shown. This is often disregarded, perhaps because most graphics are static and permanently show a state at a particular time - a "time slice". When time is treated as a variable, it often represents time itself, though questions of scaling it are largely unresolved. Time may carry either qualitative or quantitative information.

Size- the extent of a symbol is used to show quantitative differences. In planar constructions it often conveys differences in magnitude. In other displays it is used as a depth cue.

Shape- the geometric form of a symbol. Most often shape is used to show qualitative differences. It also is used to show the form of a geographic entity and hence to identify geographic objects by their forms.

Orientation- of some linear symbols can be used to show quantitative (naturally, directional) differences.

Color is in fact a melange of three separable variables. *Hue*- is the aspect of color which is determined by its dominant spectral component, and gives rise to distinctions between, for instance, red, blue and green. It is important aesthetically and usually represents qualitative differences. *Value*- the lightness or darkness of a color is the most important visual dimension of color. Dark areas tend to draw the eye. Value differences usually represent quantitative differences. Tradition and experimental studies suggest that darker symbols should represent greater amounts. *Saturation*- is the purity of a color, the degree to which it is dominated by a single spectral component. It has generally not been used in cartography as it is very subtle. The visual system is less sensitive to saturation than to hue or value.

Pattern- is the use of repeated elements to build up a complex symbol over and area or along a line. It has several components. *Elements*- are the basic graphic marks which are repeated to build a pattern. They are usually not used as a symbol of themselves but could convey qualitative information. *Texture*- often specified as lines per inch (lpi) is a measure of how frequently the elements of the pattern are repeated in space. It has been used to convey both qualitative and quantitative information, and can be a strong depth cue. *Value*- with pattern, as with color, is a measure of lightness or darkness. It is often specified as the percentage of the area that is actually covered by pattern elements. It conveys quantitative. *Arrangement*- is the lattice Structure which determines the relative positions of the repeated elements. It could carry qualitative information but usually is not used by itself. *Orientation*- refers to the directional bias of the pattern. It is often used to carry qualitative information.

Focus- one of the complex emergent variables. I believe MacEachren was the first one I heard/read explicitly mention focus as a visual variable, though it has long been used as a depth cue (atmospheric haze) in landscape painting. It refers to the sharpness of a symbol and is related to the notions of focus in photography and to the notion of smudged linework. It can carry at least qualitative, and probably quantitative information. I know of no cur-rent implementations using it.

Realism- is another emergent visual variable, referring to how realistic the image appears. It is a computationally complex variable, involving reflectance and surface texture modeling. It could be used to carry a great deal of information and is being explored in the context of virtual reality (VR) research.

These graphic variables are clearly not orthogonal to one another. Variation in size and shape presupposes existence and location. Pattern and realism emerge from complexes of the more primitive variables. For example, the appearance of a shiny brass ball emerges from the proper placement of a set of correctly colored pixels on a CRT screen. Taken together, these graphic variables Constitute a lexicon that can be used to construct meaningful graphic representations.

Media and Perceptual Limitations

The use of these graphic variables is subject to the graphic limits of the presentation media and to the limits of the human perceptual and cognitive systems. These limits determine how subtle a distinction might be presented to the user. The presentation limits vary between media (e.g., printing or CRT screens). The perceptual limits vary among individual display users (anomalous color vision, reading style, and experience are examples of these differences). Here, we need be concerned that sufficiently many distinctions may be shown and that these distinctions will be large enough to be noticed.

In the short run, the principal output device for visualization of data quality will be the color CRT screen. (I recognize that the vast majority of visual displays are printed, and that this will continue, but the challenges and potentials for innovative visualization work are tied to computer graphics on CRTs.) In this environment, the graphic designer's obvious concerns are: enough pixels and adequate spatial resolution to allow sufficiently fine placement of graphic details, and a color gamut large enough and color resolution fine enough to realize a graphic. Screen update speed is such that smooth animation of fairly complex graphics is possible, but it is still possible to envision (and specify) displays that are impossible to compute and display in real-time.

As used in most current workstations, the constraints on CRT spatial resolution are more binding than are those on color resolution. A large CRT may be 50 cm by 50 cm, and be capable of displaying an array of 2048 by 2048 pixels. A more typical display is 18 cm by 23 cm and 640 by 480 pixels, or perhaps 30 cm by 30 cm and 1024 pixels on a side. The pixel spacing of the typical high resolution CRT is about .22 mm. The smallest spatial distinction that can be shown on such devices, viewed at normal reading distances, is considerably larger than the useable threshold for visual acuity. In color space, on the other hand, *the defacto* standard 24-bits of color (eight each for red, green and blue) results in a relatively dense color space, where the minimum step is below the practical limits of visual discrimination.

Cognitive Limitations

Another limit on the design space for data quality visualization is imposed by the human cognitive system. There are limits to the amount of image complexity people can handle. Models of mental imagery Such as Kosslyn's (1983) indicate that there are limits on the detail in mental images, and imply that there is a limit to the resolution with which one might perform "mental overlay" of a quality map with another map. This argues in favor of pushing complexity onto the visual system (and an external store) rather than expecting detailed image memory for interpreting quality displays.

Cognitive limits appear also when we consider the amount of information that the user is expected to maintain in mind to interpret the display. Miller's (1956) magic number seven +/-two probably is optimistic if applied to numbers of different symbol dimensions employed to represent orthogonal data attributes. Presumably, it would be very difficult to interpret nine separate aspects of a single graphic symbol, even if they were made graphically distinct. Limits on such processing argue for considerably simpler symbology. For instance, it may not be feasible to show four aspects of the quality of a data point and still show its location. We should not be put off by this, but should rather be mindful to not fixate too strongly on the idea that visualizing spatial data quality requires maps. If the spatial aspect of the data can be forgone, then more other variables can be shown.

Representing Uncertainty

Given that one has characterized the aspect(s) of data quality (the model of data quality) that is to be shown, graphic codes must be employed to convey this information. Generally, the symbology will have to allow identification of the objects being shown and carry the data quality as an additional meaninig. This is more complex than simply identifying what is at a place, but is not impossible nor particularly daunting. The codes should be chosen such that the desired information is carried and unintended communication is minimized (see Cuff 1973 for a discussion of unintended messages being conveyed by maps). For instance, a trail with uncertain position should not appear to be a wide highway or a braided stream.

There seem to be several strategies for displaying data quality. One is to explicitly encode the quality, attached to objects, points, lines, areas or fields; the other is to use graphic ambiguity to create visual and cognitive ambiguity. The first of these is direct, and assumes that the data quality should be prominent in the graphic. It presupposes that data quality is specified, measured, and can be mapped like any other attribute data. The second approach, using graphic ambiguity, is more subtle. It assumes that the goal is more to communicate the notion that the data are known to contain problems the exact nature of which are unknown. The two strategies are not mutually exclusive.

Explicit uncertainty codes

When data quality is known, it can be treated as simply another attribute to be displayed. Point, line, and area symbols (including fields) can have data quality graphically encoded as part of the symbol. The quality might be represented at a nominal, ordinal, or interval/ratio scale using a number of existing symbol conventions.

One approach is to directly label the quality of the data. In a rather extreme version of directly labeling the quality of data, J. K Wright (1942, p. 528) suggests that codes such as: broken contours, "P.D." (position doubtful), or "E.D." (existence doubtful) might be, employed, as might a margin-map "showing the character of the surveys and other sources on which a map is based – a 'relative reliability' diagram." (These were used on several American Geographical Society maps and on the International Map of the World - all in the early 1940s.)

To allow direct symbolization, the collector is assumed to have passed data quality data on to the cartographer. In an automated system, data fields would be needed but not much else (provided we knew what constituted quality). Then symbols for the objects could be designed such that they reflect the data quality. One might find objects for which the data are known to be unreliable shown in red while more reliable objects are yellow, and those with the greatest reliability are shown in green. Similarly, a symbol like Tissot's Indicatrix might be used to show the uncertainty in a point's location, even while color shows uncertainty about its classification.

Users may need a little coaching on how to interpret these symbols so that "green" is not taken for "pasture", rather than high confidence. Highly prominent legends may be needed until people get used to having this type of display available. A symbol for blank space (making no claim as to what exists in an area, but being absolutely sure of that) to distinguish it from an area that is "normal" (and to which the reader might with assurance ascribe the default attributes found over an area) may be needed.

Zones on the map for which data are known to be of greater or lesser reliability could be shown as a set of polygons overlaid on the rest of the map, where each of these polygons is symbolized as having some level of quality. Perhaps red and yellow would be used to suggest danger and caution. This would be much like the marginal map of reliability, but would superimpose the reliability on the graphic.

Individual points, lines and area symbols could be portrayed varying one or more of their graphic dimensions to encode the nature, rank, or amount of data quality associated with them. Shape might be used for point features such that triangles mean very high confidence in their existence and geometric figures with greater numbers of sides mean there is relatively less confidence in the existence of the feature. A round dot, having an infinite number of sides, would then be a much less certain feature than would a Pentagon, or a square. The order of these symbols may seem more natural if it were reversed, a question for experimentalists.

It is possible to take advantage of the animation potential of dynamic mapping. One might use something like blinking point line or area symbols, with the frequency of blinking, or the amount of time "on", being proportional to the confidence in, or quality of, the data. Or a symbol might oscillate between two colors. One might be a function of the feature's meaning (stream) and the other a function of the confidence that the meaning is correct. A blue stream may oscillate between blue and green if there is confidence in its existence and between blue and red if there is not. An alternative use of oscillation might be to display the range of values that a feature might have (absolute or within confidence intervals). The blue of the steam might fluctuate through a range of values indicating the range of its discharge.

Graphic Ambiguity to Create Visual Ambiguity

Locational ambiguity may be symbolized using positional ambiguity. For instance an uncertainly located item may be, shown in multiple positions. The multiplicity might be color, pattern, size or otherwise coded to reflect the probability of each location being correct. Or none may be indicated as more likely. A line might be shown as a braid of several or many lines. A point as a cloud of points, and an area as a set of (likely overlapping) areas. The same type of symbols could be used only to indicate "unsureity", with none of the graphic representations meant to be taken as an actual, probable, or even possible location.

The absence of "hard" lines or edges where data are uncertain can create the impression of being "unsure" about positions and other attributes. The edges between soil polygons for instance might be depicted with broad fuzzy lines. The value or saturation of the line could constant (it's in there somewhere) or could vary across the line's width (perhaps as a Gaussian normal distribution, or in discrete bands of confidence, or as the probability of the line's true position.

Blending zones between overlapping /adjacent areas where there are mixes of phenomena might be shown by mixing colors over the area. Colors or patterns that blend psychologically such as the two ended color schemes (red-purple-blue or blue-aqua-green) might well serve for this. Smooth transitions might be shown (perhaps requiring very subtle color distinctions and large numbers of colors), or perhaps dithered patterns of more patchy or blotchy mixes (perhaps with lower color resolution) might be used to show different degrees of homogeneity suspected (or known) to be in the data.

The same type of effect might be used to blur the edges of point and line symbols so that their degree of fuzziness can be used to determine the quality of the data. This might be considered a change in the "focus" of a symbol.

Animation might also be used to convey ambiguity. A point or line that has uncertain position could be animated so that it moves over the range of its possible positions (perhaps occupying each for a period proportional to the probability that it is the correct location). Thus, the movement of the object gives the sense that it is the area but that it is hard to pin down exactly where. (This might be called the Heisenburg technique). The effect could be achieved through sprite, VLT , or frame animation.

Animation allows one to take advantage of tile Visual systems' tendency to understand input in terms of the experimental three-dimensional world. According to J. J. Gibson's view of perception as a process embedded in the world, many visual cues should be interpreted as making sense not in the image plane bill in three-dimensional space. Thus, differential velocities of pairs of an object projected oil the retina do not mean that the object is being oddly stretched in a plane, but that the object out there rotating. This type of manipulation in a graphic could encode spatial relations in an animated image when their inclusion would overload a static image.

Another animation idea with an experiential basis is to shroud uncertain data in fog. A "mist" or "fog" could drift in and out, over and around uncertain objects or areas, partially obscuring them, much as fog obscures objects in reality. Things that are sure may not fade from view at all while things that are uncertain may fluctuate between invisibility and being barely visible. The fading could be achieved by varying, primarily, the saturation of the background and symbols oil an otherwise normal display.

Interaction

Interactive exploration of data and data quality is very attractive for several reasons. It gives better, more direct access to the data so the user develops a better "feel" for it. It also promotes more rehearsal in learning the data through manipulating it. Various graphic interface environments provide tools (both software and hardware) for interacting with graphic images (moving them, rotating them independently on each axis, zooming, panning, changing the viewpoint etc). Access to these tools is increasing and standardized application of the various widgets to the manipulation of data quality data would make it easier for users to transfer learning to new systems. We should find the most natural ways to interact with quality data. One obvious example is given below.

A map could appear with essentially any normal (or abnormal) symbology and with a "Quality Slider". The slider is manipulated to set a data quality threshold that controls the selection of map features for display. It could be used as a select-for-display mechanism where only super (or sub, or near) threshold objects are displayed, or it could be used to alter the appearance of those objects such that while all objects are shown, those selected by the operation of the slider are symbolically distinguished from the others. This type of display would facilitate and encourage "What if..." types of analyses in considering data quality in a spatial context.

Conclusions

Given adequate models of data quality, there art! no logical impediments to creating visualizations of it. The technology we have now for graphic communication is very powerful and will become more so. The graphic equivalent of a lexicon or perhaps of phonemes is large and may become slightly larger. Human information processing capabilities are a relatively inflexible constraint, though within bounds we can reprogram (teach) people to interpret the visual cues they are given. Constraints on the ability to separate and use different visual dimensions of symbols in a display limit the number of separate aspects of data that can be portrayed simultaneously. Overloading the graphic will result in a break-down in communication. There are interesting, challenges ahead to imagine and create the needed displays.

References

David Cuff, 1973, "Shading on Choropleth Maps: Some Suspicions Confirmed", Proceedings of the Association of American Geographers, v. 5, pp. 50-54.

Jacques Bertin, 1983, Semiology of Graphics: Diagrams Networks Maps, Translated by William J. Berg, University of Wisconsin Press.

Borden D. Dent, 1985, Principles of Thematic Map Design, Addison-Wesley, Reading, Mass.

J. J. Gibson, 1979, The Ecological Approach to Visual Perception, Boston, Houghton Mifflin.

Stephen M. Kosslyn, 1983, Ghosts in the Mind's Machine: Creating and Using Images in the Brain, Norton.

George F. McCleary, 1983, An Effective Graphic Vocabulary, IEEE Computer Graphics and Applications, v. 3, n. 2, pp. 46-53.

G. A. Miller, 1956, The Magical Number Seven Plus or Minus Two: Some Limits on Our Capacity for Processing Information, Psychological Review, v. 63, pp. 8 1-97.

Arthur H. Robinson, Randall D. Sale, Joel L. Morrison, and Phillip C. Muehrcke, 1984, Elements of Cartography (5th ed), John Wiley & Sons, New York.

S. S. Stevens, 1946, On the Theory of Scales of Measurement, Science, v. 103, pp. 677680.

S. S. Stevens, 1951, Mathematics, Measurement and Psychophysics, in S. S. Stevens (editor) Handbook of Experimental Psychology, John Wiley & Sons, New York.

Edward R. Tufte, 1983, The Visual Disphiy Of QUantit,,itive Information, Graphics Press.

John K. Wright, 1942, Map Makers Are Human: Comments on the Subjective in Maps, The Geographical Review, v. 32, n. 4, pp. 527-544

John K. Wright, 1955, "Crossbreeding" Geographical Quantities, The Geographical Review, v. 46, pp. 52-65.

Key Issues in the Use of Experiential Graphics for Exploring Data Quality

Mark Monmonier
Department of Geography
Syracuse University
Syracuse, New York 1324-4-1160
mon2ier@sunrise.acs.syr.edu

In this short statement I identify six key issues in the visualization of data quality from a perspective that might be called 'dynamic statistical graphic," with a geographic slant'. This viewpoint reflects more the union than the intersection of statistical graphics and cartography; the former field has had little to do with maps and geographic data, whereas the latter has tended to neither appreciate nor exploit the integration of maps, statistical charts, and other windows of information such as text blocks. The adjective 'dynamic' indicates a concern with display systems that can generate sequences of graphics, change the display sufficiently rapidly to warrant use of the term 'real time' and respond readily to the analyst's requests to interrupt and modify the presentation.

'Experiential graphics' is a new term that seems worth promoting. It describes interactive graphics that allow the user to 'experience' patterns and trends in the data rather than merely look at them. Although this concept is an article of faith in both electronic technology and human visual-information processing, its goal is improved insight rather than increased efficiency. Removing the wait and frustration of contemporary interactive graphics will most certainly promote a fuller, more serious exploitation of the enormous information-processing capabilities of the eye-brain system. And even if this fuller involvement often demonstrates just how shallow our data really are, such a realization should itself be a valuable stimulant for better geographic scholarship and the more effective use of GIS technology.

The four research questions I present relate to two of the four themes of this meeting, Representational Issues and the Evaluation of User Needs. The types of graphic solutions I propose require careful attention to graphic symbolization techniques and graphic theory as well as an assessment of how effectively analysts employ such tools and what improvements are needed.

Key Issues in the Visualization of Data Quality

The following key issues address not only deficiencies in the data but deficiencies in current methods of examining and using data. Data quality is a problematic concept not easily divorced from the quality of the analytical methods we use to examine, process, and extract meaning from our data.

1. **The need for machine-supported vigilance**. Although visualization can be important for exploring geographic data and assessing their usefulness, the assumption that the analyst will reliably and consistently identify blunders in data development or potentially troublesome areas seems naive, particularly for large and complex data bases. The human analyst might not scan the data thoroughly, in a systematic or otherwise rational search pattern, and might not always see obvious mistakes or trouble spots, however readily apparent on the screen. After all, looking for deficiencies in data can be a tedious and tiring task, subject to vigilance error. Why not program or train a geographic visualization support system to carry out this important but computationally demanding task of trawling for blunders?

Systematically scanning for deficiencies in the data is akin to automatically screening the data for important or potentially interesting or meaningful patterns. Development of this aspect of a geographic visualization support system might benefit from somewhat parallel developments in the use of supercomputing for automated geographic analysis and "trawling for meaning" (Openshaw, 1987; Openshaw et al., 1987) and in the automated screening for geographic correlation (Monmonier, 1990; Openshaw, Cross and Charlton, 1990).

2. **The need to measure and identify blunders, trouble spots, and other geographically meaningful patterns and relationships**. Any attempt to develop machine-supported vigilance for geographic data begs the question of how most effectively to define and identify interesting and potentially meaningful geographic patterns that an alert and competent human analyst might be expected to notice. But in order to design a system that hunts for error we must first develop an effective means of describing such blunders and trouble spots.

Inferential statistics (particularly tests concerned with residuals from regression) has given geographic analysts a convenient method for identifying relatively extreme values, or outliers, which might reflect measurement error. But the types of data used in GIS, particularly qualitative data, and the variety of GIS applications call for a wider range of approaches. Because a rule-base strategy might point out relatively more bizarre situations that deviate from customary relationships among geographic features, expert systems and artificial intelligence (Robinson and Frank, 1987) offer a promising approach to automated blunder detection. Geometric pattern-recognition strategies (e.g., Hudson, 1969) might also prove useful within a rule-base framework.

3. **The need to offer an efficient and informative introduction to users not acquainted with a geographic data base, and thereby orient them to both the richness and the limitations of the data**. To be efficient, such an introduction might consist, at least in part, of an animated sequence of graphics that goes well beyond the staid and static format of conventional user documentation. This animated sequence could exploit the recognized potential of real-time cartographic animation (Gersmehl, 1990; Moellering, 1980) and graphic sequencing (Slocum et al., 1988; Taylor, 1982) as well as the automated troublespot-detection strategy mentioned above. Ideally, this graphic overview should be tailored to both the data and the prospective user, who should be neither bored with unnecessary details nor inundated with sketchy facts unrelated to prior knowledge or experience.

A graphic orientation to a data base is but one type of graphic script, a term I use to describe a sequence of maps, statistical graphics, text blocks, and other exhibits designed to describe or explore a region, a correlation, or an entire data base (Monmonier, 1989b). These programmed graphic sequences can address a variety of themes and functions. As examples, graphic scripts could describe the growth of settlement of the United States, the 120-year pattern of Chicago municipal elections, or the range of sites the state of New Hampshire might consider 'suitable' for a low-level radioactive waste dump. Graphic scripts can be composed slowly and deliberately by a human author, or devised automatically by a geographic visualization support system to present the salient aspects of a geographic data set or phenomenon in a meaningful and intelligible sequence to a general audience or a specific viewer. The sequence might involve more than one window, and even include a voiceover recorded by a human narrator for playback or generated by a speech-synthesizer in real time. The interactive viewer can interrupt or reverse the sequence of graphics, and accelerate or slow down its pace.

Another term, the *graphic phrase*, describes an inherently shorter sequence of graphics with a narrower objective. A graphic phrase might consist, for example, of a sequence of graphics that examines the major geographic trend of a single variable and the principal departures (if any) from that trend. An interactive system might provide a menu of graphic phrases, and several graphic phrases might be linked to produce a graphic script.

Among the facets of a data base a graphic script might examine are the degree of covariation among the variables, the tendency of extreme values to occur in similar parts of the study region, and the relative homogeneity of the data within a widely accepted set of geographic regions. Additional information, for example, the set of attributes and thresholds that makes a site 'acceptable' for a landfill, can add meaning to raw data and support development of a more relevant graphic script. Providing a stored profile of the user's interests or allowing the user to indicate directly his or her priorities and objectives is essential for the successful automated generation of a tailored graphic script.

4. **The viewer's need not only to interrupt and alter the graphic script (in particular to refine queries and request further detail) but to use hypermedia. or other pointing and linking tools to explore the data more freely**. The limitations of human memory can be an enormous impediment to the efficient and accurate analysis of geographic data. Analysts need time to absorb what they see, to relate it to what they know, and to avoid the notorious information-processing bottleneck of short-term memory (Kosslyn, 1985). And when long-term memory fails to supply the needed information, the analyst could benefit greatly from a system that readily retrieves relevant facts. Being able to point to a feature or a region with a mouse or touchsensitive screen and then to display information about the entity so selected can be highly useful to the researcher with an imperfect memory or no prior experience with the data set. This strategy is also an ideal method of exploiting the carefully coded annotations of a well-documented data base.

A device for pointing to features on the screen might also be used to activate a variety of icons or keywords representing nodes in a hypermedia system (Conklin, 1987; Irven et al., 1988). By clicking on these nodes the analyst might highlight features based upon the same source materials-or similar in level of uncertainty. Hypermedia technology also provides a: convenient mechanism for integrating a GIS with an expert advisory system for dealing with uncertain and imprecise data (Gaines and Linster, 1990).

A point-and-click strategy might be particularly useful in dealing with features that vary in positional accuracy. Geographically imprecise features include coverages extracted from unrectified aerial photographs and composite coverages based upon the overlay of features with horizontally uncertain representations. The analyst selecting such a feature might temporarily replace its crisp line or boundary symbol with a fuzzy band that fosters a greater understanding of the true nature of the data as well as a more sensible interpretation of site analyses and other solutions based on spread functions and map overlay.

5**. The need to measure and identify complementary representations, especially the complementarity of maps and statistical charts**. GIS and cartography must eventually abandon the traditional but in many ways dishonest single-map solution to problems of graphic representation. Yet some selectivity is essential because the number of unique graphics possible for even a modest-size data base is enormous, and no one would want to look at them all. With reliable and meaningful measures of visual and informational complementarity, we can design visualization support systems that point out features and relationships the user might otherwise ignore yet avoid an overwhelming graphic bombardment.

There is a pressing need to examine the complementarity of maps and charts, and to seek a bridge between GIS and exploratory data analysis (Tukey, 1977). Although geographic data are commonly analyzed numerically in both geographic space and multivariate space, current strategies for their visual analysis are overwhelmingly cartographic (Monmonier, 1988). EDA can provide useful insights by portraying important relationships in attribute space and by adjusting the data for obvious effects, such as those of age or urbanization on mortality, so that the analyst can focus on more revealing patterns. EDA techniques that have proven useful in guiding or influencing the selection of attribute-space views of the data might also be useful in automating the selection of statistical graphics displayed in a graphic script (Buja and Asimov, 1986; Donoho et al., 1986; Tukey and Tukey 1981).

Brushing, a tool developed in statistical graphics for the interactive analysis of multivariate data displayed in a scatterplot matrix (Becker and Cleveland, 1987; Carr et al., 1986), is an EDA technique with considerable promise for experiential cartography. In scatterplot brushing, when the analyst uses a rectangular 'brush' to highlight a group of observations in one cell of a scatterplot array, the points representing these observations are highlighted in all other cells as well. In geographic brushing, places represented by the selected points are also highlighted on the map in a 'geographic window' (Haslett, Wills and Unwin, 1990; Monmonier, 1989a). In addition, the user could choose observations for highlighting by pointing to or circling places on the map or by selecting one or more regions listed on a pull-down menu of regions identified by name. A promising tool for the interactive exploration of covariance relationships, brushing might also be used to explore contiguity effects by relating boundaries on the map to pairs of observations in a scatterplot array.

6. **Cognitive issues**. Adding this highly general and somewhat vague topic to my list of issues reflects a reverent and somewhat mystical sense of the importance of cognition, particularly to a well-grounded development of experiential cartography and its use in assessing data quality. It would be good to replace the intuitive and possibly naive notions currently guiding this development with a more solid empirical and theoretical base-with knowing rather than merely feeling or reasoning. But I detect little progress of a type actually useful in telling us what we should or should not do in addressing data quality. Published studies seem both too reductionist and too general to be of much relevance to our mission, yet the deep optimism of the positivist in me suggests that this issue might merit some discussion, although not a great deal. Please note that I distinguish cognitive issues from human-factors issues, which a number of workers in the GIS arena are addressing with promise and satisfaction.

Important Research Questions Concerned with Representational Issues and the Evaluation of User Needs

The following four research questions relate directly to the issues presented in the previous section and require little further discussion.

1. **What graphic scripts are most efficient and effective in introducing the user to a new geographic data base?** This question calls for a systematic assessment of what users would like to see in such an introduction and how they would use this information. Promising qualitative data-gathering strategies useful in assessing user needs and evaluating prototype graphic scripts include the focus-group interview (Greenbaum, 1988; Krueger, 1988) and the sensemaking debriefing interview (Dervin, Nilan and Jacobson, 1982; Nilan, Peek and Snyder, 1988).

2. **How might the concept of the user profile be used to tailor an introductory graphic script to the type of data base and the interests and experience of the user?** A user profile is a representation of an analyst's information needs, priorities, knowledge, and experience (Daniels, 1986). An expert-advisor system might be useful in constructing profiles by querying users about either their interests or their satisfaction with the information presented by previous graphic scripts. User profiles offer a promising strategy for circumventing the usually difficult task of designing a system to accommodate a range of user needs.

3. **What types of information about the data might the user acquire with interactive brushing? What modifications of brushing strategies are needed to make this information particularly useful for assessing data quality?** As with the development and assessment of graphic scripts, this research question calls for a series of development cycles based on user assessment and the incremental refinement of a prototype system.

4. **How effectively do users employ other interactive pointing strategies for exploratory data base interrogation? How do they act on information obtained in this way? What additional information might they find useful? What role might graphic phrases play in the interactive interrogation of a geographic data base?** As with the overview graphics script and geographic brushing, focus-group interviews and sensemaking debriefing can guide the iterative development of a working system.

Concluding Remarks

To conclude, I add one further research question, namely:

5. **What technical and institutional impediments are blocking a wider adoption of experiential approaches to the visual assessment of data quality?** Although outside the purview of Initiative 7, this is an important research question. Unless users see the possibilities and value of better systems and developers perceive a significant competitive advantage in improving their products, the embarrassing gap between the state-of-research in the academic sector of GIS and the state-of-practice in the commercial software sector seems likely to grow wider. A careful, convincing and widely publicized analysis of both the technical and the institutional constraints on the incorporation of efficient visualization tools in GIS software might lead to timely implementations that enhance the social return on Initiative 7 and individual research efforts with similar aoals.

Literature Cited

Becker, R. A., and W. S. Cleveland. 1987. Brushing scatterplots, Technometrics 29: 127-142.

Buja, A., and D. Asimov. 1986. Grand tour methods: an outline, Computer science and statistics: the interface, ed. D. M. Allen, Elsevier Science Publishers, New York, pp. 63-67.

Carr, D. B., et al. 1986. Interactive color display methods for multivariate data, Statistical image processing and graphics, ed. E. J. Wegman and D. J. DePriest, Marcel Dekker, New York, pp. 215-250.

Conklin, J. 1987. Hypertext: an introduction and survey, IEEE Computer 20 (no. 9): 17-41.

Daniels, P. J. 1986. Cognitive models in information retrieval-an evaluative review, Journal of Documentation 42: 272-304.

Dervin, B., M. S. Nilan, and T. L. Jacobson. 1982. Improving predictions of information use: a comparative predictor of types in a health communication setting, Communication Yearbook, Vol. 5, ed. M. Burgoon, Transaction Books, New York, pp. 807-830.

Donoho, D. L., et al. 1986. The man-machine graphics interface for statistical data analysis, Statistical image processing and graphics, ed. E. J. Wegman and D. J. DePriest, Marcel Dekker, New York, pp. 203-213.

Gaines, B. R., and M. Linster. 1990. Integrating a knowledge acquisition tool, an expert system shell, and a hypermedia system, International Journal of Expert Systems 3: 105-129.

Gersmehl, P. J. 1990. Choosing tools: nine metaphors for map animation, Cartographic Perspectives no. 5 (Spring issue): 3-17.

Greenbaum, T. L. 1988. The practical handbook and guide to focus group research, D. C. Heath, Lexington, Massachusetts.

Haslett, J., G. Wills, and A. Unwin. 1990. SPIDER-an interactive statistical tool for the analysis of spatially distributed data, International Journal of Geographical Information Systems 4: 285-296.

Hudson, J. C. 1969. Pattern recognition in empirical map analysis, Journal of Regional Science 9: 189-198.

Irven, J. H., et al. 1988. Multi-media information services: a laboratory study, IEEE Communications Magazine 26 (no. 6): 27-44.

Kosslyn, S. M. 1985. Graphics and human information processing, Journal of the American Statistical Association 80: 499-512.

Krueger, R. A. 1988. Focus groups: a practical guide for applied research, Sage Publications, Beverly Hills, California.

Moellering, H. 1980. The real-time animation of three-dimensional maps, American Cartographer 7: 67-75.

Monmonier, M. 1988. Geographical representations in statistical graphics: a conceptual framework, 1988 Proceedings of the Section on Statistical Graphics, American Statistical Association, pp. 1-10.

__________. 1989a. Geographic brushing: enhancing exploratory analysis of the scatterplot matrix, Geographical Analysis 21: 81-84.

__________. 1989b. Graphic scripts for the sequenced visualization of geographic data, Proceedings of GIS/LIS '89, Orlando, Florida, Nov. 26-30, 1989, pp. 381-389.

__________. 1990. Strategies for the interactive exploration of geographic correlation, Proceedings of the 4th International Symposium on Spatial Data Handling, Zurich, Switzerland, 1990, pp. 512-521.

Nilan, M. S., R. P. Peek, and H. W. Snyder. 1988. A methodology for tapping user evaluation behaviors: an exploration of users' strategy, source and information evaluating, Proceedings of the 51st Annual Meeting of the American Society for Information Science 25: 152-159.

Openshaw, S. 1987. An automated geographical analysis system, Environment and Planning A 19: 431-4306.

Openshaw, S., A. Cross, and M. Charlton. 1990. Building a prototype Geographical Correlates Exploration Machine, International Journal of Geographical Information Systems 4: 297-311.

Openshaw, S., et al. 1987. A Mark 1 Geographical Analysis Machine for the automated analysis of point data sets, International Journal of Geographical Information Systems 1: 335-358.

Robinson, V. B., and A. U. Frank. 1987. Expert systems for geographic information systems, Photogrammetric Engineering and Remote Sensing 53: 1435-1441.

Slocum, T. A., et al. 1988. Developing an information system for choropleth maps, Proceedings of the Third International Symposium on Spatial Data Handling, August 17-19, 1988, Sydney, Australia, pp. 293-305.

Tukey, J. W. 1977. Exploratory data analysis, Addison-Wesley, Reading, Mass.

Taylor, D. R. F. 1982. The cartographic potential of Telidon, Cartographica 19 (nos. 3/4): 18-30.

Tukey, P. A., and J. W. Tukey. 1981. Preparation; prechosen sequences of views, Interpreting multivariate data, ed. Vic Barnett, John Wiley and Sons, Chichester, pp. 189-213.

m2--13mar91

Representing Error in GIS Modeling[1]

James F. Palmer
SUNY College of Environmental Science and Forestry
Syracuse, New York, 13210-2787

Problem

I would like to suggest that as specialists in the area of geographic analysis, we work in a virtual reality, even though our work and its purpose has a corresponding physical reality. In and of itself, that is not a problem until we must make decisions from virtual information that will take physical form.

Burrough (1986:112) describes the problem of faults in the virtual representation of the environment this way:

> Most procedures commonly used in geographical information processing assume implicitly that (a) the source data are uniform, (b) digitizing procedures are infallible, (c) map overlay is merely a question of intersecting boundaries and reconnecting a line network, (d) boundaries can be sharply defined and drawn, (e) all algorithms can be assumed to operate in a fully deterministic way, and (f) class intervals defined for one or other 'natural' reason necessarily are the best for all mapping attributes. ... rarely have these problems been looked at as a consequence of the way in which the various aspects of the world have been perceived, recorded, and mapped.

These assumptions have serious effect when the electrical impulses are Census data used to redraw the districts of our legislative representatives, or when they represent attributes used to identify a synthesize area thought to correspond to wetlands in which development is restricted, or the resistance values in the shortest path analysis used by Domino's to deliver its pizzas on time. This paper describes one approach to represent possible error in data used for a GIS analysis.

Example of Data Error Representation Visibility Analysis.

The identification of 'seen areas' or visibility analysis, could arguably be considered the most objective form of scenic analysis conducted by landscape architects--since it is a simple matter of geometry--with little variation among the findings of different professionals (Palmer 1983). This expectation of objectivity is only enhanced with the reliance on computerized geographic informations systems (GIS) to perform the calculations. However, computers give us a false sense of confidence and are particularly subject to uncertainty errors related to the phenomenon known colloquially as "garbage-in-garbage-out" or GIGO.

A typical visibility map is shown in figure 1. What do you see? What information can it provide someone needing to make a decision about siting a facility? What is left out or is even misleading? A decision-maker interested in the bottom line would reasonably draw conclusions about whether some location in the landscape is seen or not seen from the selected viewpoint. In their reviews of approximately 100 major planning and project impact reports, Felleman (1982) and Griffin (1989)found maps that looked very much like figure 1, with little or no documentation of methods or parameters used. For instance, figure 1 was calculated using a 30-meter digital elevation model (DEM) derived from a USGS 7.5 minute quadrangle topography map. The elevation of the viewer's eye is two meters, the visible points are at ground level, and there is no consideration made for land cover. The program used is MAP II (Pazner 1989), which does not document the algorithm used from among the many available (Sutherland et al. 1974).

[1] Research in this paper was funded in part by Cooperative Agreement No. 23-88-27 from the U.S. Forest Service, North Central Experiment Station. John Felleman was the principal investigator.

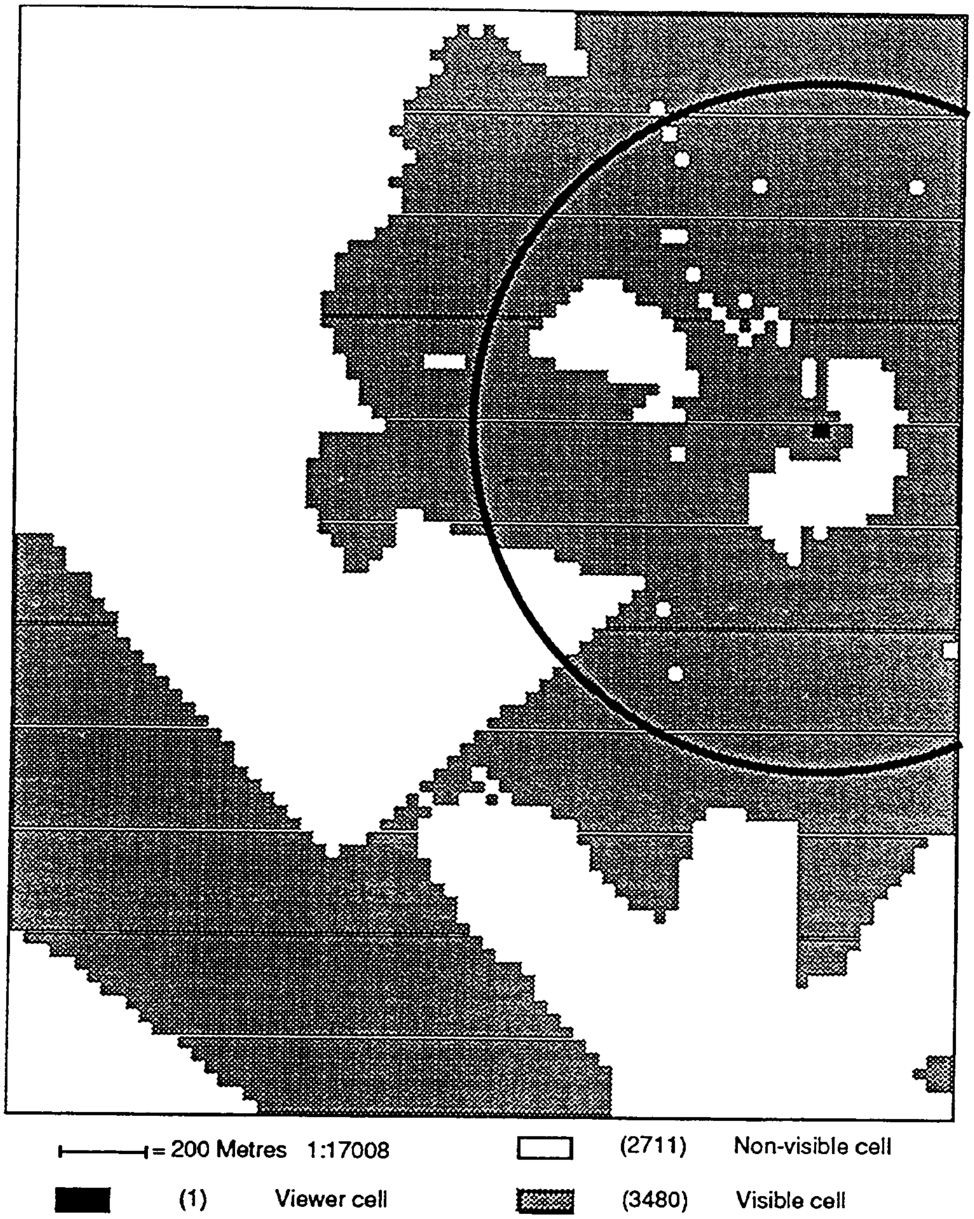


Figure 1. A typical visibility map showing areas seen and not seen from a viewpoint with an arc indicating the boundary between foreground from the middle ground.

Evaluating Visibility Accuracy. The U.S. Geological Survey is the source of most topographic data used for visibility analyses. Most of their products conform to the National Map Accuracy Standard "that no more than 10 percent of the points tested shall be in error by more than a certain tolerance" (Thompson, 1988). By assuming that the error at any point is independent of the error at any other point and that these errors are normally distributed, this standard can be implemented statistically using the standard error or root-mean-square-error (RMSE). Thompson (1988) shows that the allowable tolerance in the elevation contours (in feet) for a 1:24,000-scale map with a horizontal tolerance of 40 feet on the ground for 90 percent of the horizontal test points and a vertical tolerance of one-half contour for 90 percent of the vertical test points is:

$$\text{allowable RMSE} = 0.3\ \text{CI} + 24t$$

where CI = contour interval, and t = tangent of slope angle.

Because it is related to the normal distribution which is one of the foundations of most parametric statistics, the RMSE provides a convenient method for evaluating map accuracy in statistical terms.

Method for Modeling Probabilistic Visibility. A Monte Carlo approach is used to evaluate the effect of possible map error on the results of a visibility analysis. Monte Carlo methods provide approximate solutions to complex problems by investigating a series of models based on the random sampling of simulated data. The topographic database used for this paper is of Howe Hill near Worcester, Massachusetts. It comes bundled with IDRISI, a GIS for MS-DOS PCs distributed by the Department of Geography at Clark University. It was manually digitized from a 7.5 quadrangle map, but is in the format of a USGS DEM for the quadrangle series. The data base is 86 rows by 72 columns with a cell resolution of 30 meters. The elevation has been convened from ten foot contour intervals to the nearest meter. It ranges from 294 to 360 meters, with a mean of 330 and a standard deviation of 16.0 meters.

A group of fifty separate DEMs for Howe Hill were created, with each one introducing a different set of random normal perturbation to the topography based on the RMSE for each cell. A viewpoint was chosen near the crest of a hill of moderate elevation within the site.

The data for the fifty elevation maps with random normal perturbations were prepared in Wingz (Informix, 1988) using the NORMAL(*standard deviation*) function. The RSME as described by Thompson (1988) for a 1:24,000 series topographic map and adjusted for the change in scale from feet to meters was used as the standard deviation in NORMAL function. The "tangent of the angle of slope" was calculated in MAP II using the GRADIENT operation with the maximum option. This gives a percent slope map, and was divided by 100 to arrive at the tangent of the angle. The random error was added to the original control elevation of each cell. The 3-by-3 cell area surrounding the viewer cell was reset to the original control elevation on the assumption that the error in the immediate foreground relative to the elevation of the viewpoint would be marginal or absent. A visibility map is created for each of the fifty randomly perturbed DEMs using the RADIATE command in MAP II.[2]

The Monte Carlo approach used here adds several of these visibility maps together. The resulting probabilistic visibility map indicates the number of times each cell was seen from the viewpoint. To facilitate interpretation, the boundary between the foreground (0 to .5 mile) and middle ground (.5 to 3.5 miles) is indicated- The distances used are appropriate for the Northeastern region where the site is located (Felleman 1982). The probabilistic visibility map produced from all fifty Monte Carlo trials is shown in figure 2.

Results. The total seen area of the control visibility map in figure I is 3480 cells. Only one of the fifty Monte Carlo simulations had a greater seen area: trial 23 is 9.1 percent larger at 3861 cells. The size of the seen area of the other 49 trials ran-es between 2889 and 3295 cells with a mean of 3114 cells. Therefore, the Monte Carlo approach indicates that the control visibility map in figure 1 may over-estimate seen area by five, to fifteen percent.

[2] The command used was "Radiate <<VIEW POINT>> To 7620 At 2 Over <<RANDOM ELEVATIONS>>" where the 3x3 cell area surrounding the view point in RANDOM ELEVATIONS was reset to its unperterbed control elevations.

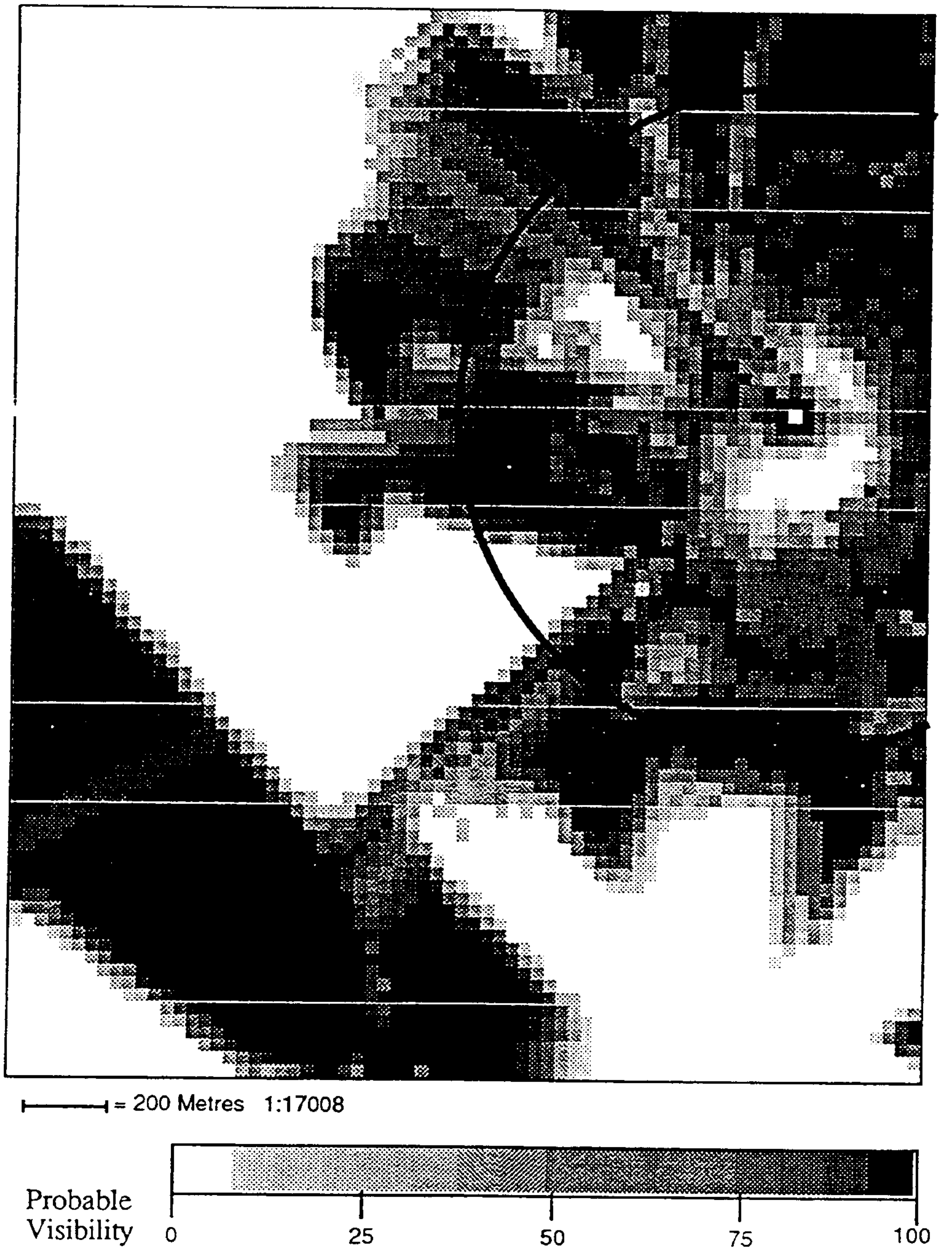


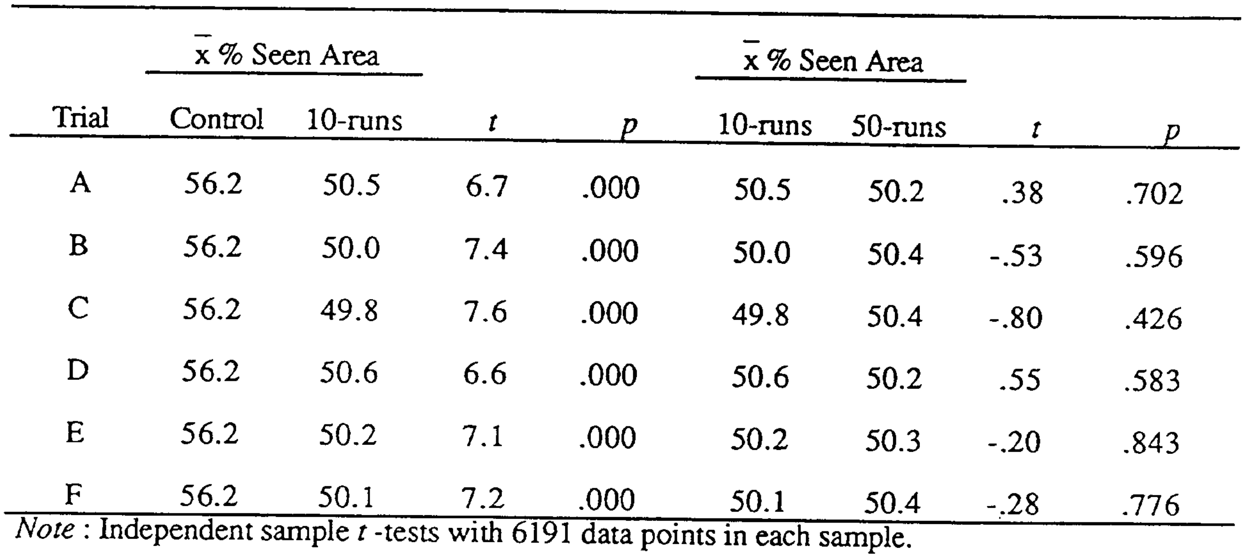
Figure 2. This visibility map shows the probability of a point being seen from the view point based on a summation of fifty Monte Carlo simulations. The boundary between the foreground and middle ground is indicated by the arc.

While approximately 43.8 percent of the control map is never visible, this percentage is reduced to 33.3 in the 50-run probabilistic visibility map in figure 2. The results are even more dramatic for the always seen areas which drop from 56.2 to 17.3 percent. In other words, approximately half of the total map in figure 2 is in a "gray" zone of less than certain visibility or invisibility.

Even when the Monte Carlo trials are grouped into five bundles of 10, there is significant variation among the different probability profiles of the resulting visibility maps. There is very high agreement about the number of cells with no probability of being seen. For instance, these include areas on the backside of the larger hills, particularly those in the middle ground. There is also consider-able agreement about the number of cells that are always seen. For instance, these include slopes on the opposite side of the valley.

A comparison of the percent area seen in the control visibility map with the mean probabilistic visibility from the 10-run and 50-run probabilistic visibility maps is shown in table 1. In order to make the comparisons using independent data sets, it was necessary to create a sixth set of ten randomly perturbed elevations. The t - tests in table I are calculated by comparing the probabilistic visibility of corresponding cells for the whole map. The results on the left half of table I indicate that the means obtained for the control and 10-run visibility maps are significantly different in each of the six trials. However, the means for the 10-run and 50-run maps do not give significantly different results. The implication for this view point and topography is that the fuzziness added by the 10-run probabilistic visibility maps is a significant addition to the information contained in the control visibility map. However, the 50-run simulations do not seem to add significantly to the information contained in the 10-run probabilistic visibility maps.

Table 1
Comparison of Mean Percent Seen Areas for the Control Visibility Map,
10-run and 50-run Probabilistic visibility Maps



	$\bar{x}$ % Seen Area				$\bar{x}$ % Seen Area			
Trial	Control	10-runs	*t*	*p*	10-runs	50-runs	*t*	*p*
A	56.2	50.5	6.7	.000	50.5	50.2	.38	.702
B	56.2	50.0	7.4	.000	50.0	50.4	-.53	.596
C	56.2	49.8	7.6	.000	49.8	50.4	-.80	.426
D	56.2	50.6	6.6	.000	50.6	50.2	.55	.583
E	56.2	50.2	7.1	.000	50.2	50.3	-.20	.843
F	56.2	50.1	7.2	.000	50.1	50.4	-.28	.776

Note : Independent sample *t* -tests with 6191 data points in each sample.

Conclusion

Would the professional seeking to site a facility make an improved, or even different decision using the approach to represent error is illustrated here? Without further research, there is no way to know.

I would suggest that there are three interrelated areas of research:

1. To develop computer programs to organize and analyze large amounts of data, and allow the viewer to dynamically interact with it.

2. To develop computer hardware that allows viewers to use all of their senses as usefully as possible.

3. To improve our understanding of how viewers approach the problems that are the object of the analysis, how they perceive the GIS displays, and the extent to which one is facilitated by the other.

These issues need to be investigated together in the same laboratory, or at least in related cooperative laboratories. I envision a studio superficially much like those planners or landscape architects have traditionally used. Practicing professionals would be invited to use the lab to work on real world problems. However, this studio would be outfitted with advanced hardware and software, as well as human guides to orient the cooperating professionals to the new medium. We could use this opportunity to study how environmental professionals approach problems, to understand how they assimilate the and use the analyses available, to test alternative approaches to improving the quality and efficiency of their effort.

References

Felleman, J. and C. Griffin. 1990. The role of error in GIS-based viewshed determination: A problem analysis. IEPP Report No. 90-2. Syracuse, NY: SUNY College of Environmental Science & Forestry.

Felleman, J.P. 1982. Visibility mapping in New York's coastal zone: A case study of alternative methods- Coastal Zone Management Journal 9(3/4):

Griffin, C. 1989. Assessing current practice in viewshed mapping. (unpublished technical paper) Syracuse, NY: SUNY College of Environmental Science & Forestry.

Informix. 1988. Wingz: Reference. Lenexa, Kansas: Informix Software, Inc.

Palmer, J.F. 1983. Visual quality and visual impact assessment. In K. Finsterbusch, L. LlewIlyn & C.P. Wolf (eds.) Social Impact Assessment Methods. Sage Publications. pp. 263-284.

Pazner, M., K.C. Kirby and N. Thies. 1989. MAP II Map Processor: A Geographic Information System for the Macintosh. New York: John Wiley & Sons.

Sutherland, I.E., R.F. Sproull and R.A. Scumaker. 1974. A characterization of ten hidden-surface algorithms. Computing Surveys 6:1-55

Thompson, M. M. 1988. Maps for America, Third edition. Reston, Virginia: USGS.

Issues on Visualization of Spatial Data Quality

Alan Saalfeld
Bureau of the Census[*]
Washington, DC 20233
saalfeld@cs.umd.edu

April 18, 1991

Abstract

Visualization is a very glamorous new technology. It has the potential to present data in a very attractive package. We are just beginning to explore uses of visualization in geographic and cartographic applications. I believe that we should focus on the "easy" applications first until we master the technology. There are many straightforward and simple applications. We must be careful not to be seduced by the high level of the technology nor by the dazzling appearance of the products of visualization.

1 A Broad View of Visualization

Let's start with a very general definition of visualization: *Computer graphics tools and technology for looking at data*. This view encompasses all of the less sophisticated applications; and it includes everything that is currently being done to make maps by computer. With this view, GIS users and computer cartographers are already old hands at visualization. They are just not *old hands* at visualizing error or spatial quality. As a first step, these old hands should try applying familiar tools to the problem of error and quality: for instance, they should make maps of error distributions, if possible.

2 The Easy Applications

We do not need to attack the most sophisticated problems first--especially if we are new at using the tools. Making maps of error distributions (when we can assign each error to a place) and overlaying maps to compare differences (visually, not quantitatively) are two very reasonable first steps.

2.1 Exploratory Data Analysis

The first principle of EDA is: *Graph it and look at the data!* A corollary to the first principle is that we can often see more than we can quantify (even after looking at the data), although seeing first helps guide us toward the proper quantification scheme and toward a logical explanation. EDA will not build our reputations as high-powered statisticians, but EDA will often get a job done.

2.2 Tools for Comparing Maps

Maps cannot be compared against absolute truth, but two maps can be compared to each other; and even subjective comparison can be valuable. In the absence of absolute measures of quality, we can often say something significant about relative quality-one map has straighter lines, one map has more features, one map shows a street that the other map does not have. Windowing, highlighting, and spatial searching all facilitate map comparisons by computer. Some of the most revealing comparisons were accomplished during the development of the Bureau of the Census' prototype Conflation System.

2.3 Conflation Research

Merging two maps requires first comparing them for sameness and differences. The Bureau of the Census built several computer graphics tools for facilitating feature matching of two maps of the same region.

2.3.1 Overlay Operations

Overlaying equivalent representations of the same region reveals a great deal about similarities and differences. Distortion between two maps is usually not uniform; however, rubber-sheeting can remove non-topological errors and permit an operator of the Conflation System to focus on topological differences of two map representations. Rubber-sheeting can be done iteratively in real time. This valuable visualization tool can help pinpoint map errors by finding map differences-if two maps disagree about the

[*] The views expressed herein are those of the author and do not reflect the views or official policy of the Bureau of the Census

existence of a feature, at least one of them is wrong. This visualization technique cannot correct errors, but it can focus expensive field operations on known problems.

2.3.2 Quantifying Visual Entities

In order to automate feature matching through overlay, we had to assess our own visual skills and tools to quantify the processes. We needed to discover how well we could correctly match features and how well we could program the computer to correctly recognize and make matches. In experimenting with rules for matches, we discovered tolerances within which we could operate effectively in matching features. In some sense, we found empirically legitimate error levels and distortions that would not impinge on automated procedures for linking two map representations.

3 Feasibility-Driven Research

We often tend to do what we can do readily rather than what is hard, but needs to be done. By recommending working on the easy applications first, I am not advocating that we ignore the difficult research. I suggest we merely postpone it. What I would like to caution, however, is that we do not do research most intensively in areas where we have clean theory merely because we have clean theory.

3.1 Sampling Error Syndrome

The Census Bureau is guilty of avoiding *dirty* theory as much as anyone. Census Bureau statisticians spend 95% of their time on what I suspect is 5% of the error in sample surveys-sampling error. Because we know how to design for and estimate sampling error for any number of complex surveys, we spend a great deal of time wrestling with formulas and generating sample variances. We deal with non-sampling error by a footnote in our publications-a disclaimer stating that there probably is some! I suspect that spatial error may also have certain hard to measure, hard to quantify, but very real, components.

3.2 Simplifying Assumptions

Assumptions such as independence, stationarity, and the like make computations easier, but they don't model the real world very well. I don't know what the corresponding *dirty* model even looks like, but we must attempt to seek it out.

4 Technology-Driven Research

"If your only tool is a hammer, you tend to see every problem as a nail." Visualization technology has advanced considerably in other fields-medicine, aerodynamics, etc.; and tools developed for those fields may be easier to adapt than to replace by new tools specifically developed for spatial issues of geography and cartography. We must make an effort not to be unduly influenced by the existing technology or the directions it has taken.

4.1 Parallel Architectures

Hypercube and grid architectures permit easy modeling of grid-like arrays of spatial objects. Processor output may be piped, one processor per pixel, directly onto a frame buffer for extremely fast raster displays. Processors communicate with their four neighbors in the cardinal directions, encouraging the analyst to model interaction of spatial processes in terms of contiguous raster elements (at least such a model has a facilitated implementation.)

Vector-based theory, on the other hand, does not map as easily onto the connectivity of parallel machines. It is probably worth the effort to develop vector-based applications for parallel machines.

4.2 Color Overkill

The technology to use millions of colors resulted in people doing just that. DIDS used a full color palette to completely saturate sensory perception; and the ultimate consequence was that the display package communicated very little to the users. We could not discern or differentiate everything we had hoped to visualize. The lesson here is that visualization technology may be able to provide far more than we can digest. So, again, the word of warning is to proceed cautiously and not too ambitiously, and to test effectiveness along the way.

5 Conclusions

Visualization offers many opportunities and many pitfalls. Geographers, cartographers, and spatial scientists should proceed carefully to develop tools for examining spatial quality. Tools that provide a picture, but not a completely clean supporting theory, are very valuable in this early development stage and should not be disparaged. Tools that match a clean, but not very realistic, theory should not dominate research at this exploratory stage. Existing technology, in and of itself, should not drive major research efforts, but instead such technology should be examined and adapted if necessary to fit spatial problems of geography, cartography, and spatial science.

Position paper prepared for NCGIA Specialist Meeting on
"Visualizing the Quality of Spatial Data" (Castine, Maine, 8-12 June, 1991).

USER ISSUES IN A GIS ENVIRONMENT

Peter Stringer,
Northern Ireland Regional Research Laboratory

Carol McGuinness and Anneke van Wersch
School of Psychology,
The Queen's University of Belfast

We are engaged in a research programme on psychological aspects of GIS use. Our research is confined to how users respond to the graphical displays generated within a GIS environment - how these displays are created, understood, interpreted and analysed. Our primary focus is on the cognitive processes of different types of user and we are developing different methodologies for examining user-GIS interactions.

WHO ARE GIS USERS?

Before any attempt can be made to evaluate user needs, a population of users must be identified and characterised. In how many ways can we identify and characterise users? A first distinction might be between "organisations as users" and "Individuals as users". Within the GIS context the discussion of user needs often refers to organisational needs and constraints - the tasks which the organisation needs to accomplish, whether they have a general or specific GIS need, the resources which are available for investment, the personnel training requirements, the means of introducing the new technology, and so on? (Burrough, 1986). An organisation might want to answer these questions in order to obtain the "best" GIS. A commercial supplier might attempt to answer them by market research. Organisations as users might then be classified as having general/specific GIS needs, high/low resources, high/low skilled personnel, positive/negative attitudes to new technology, etc.

But if we adopt a cognitive perspective and think of a GIS user as the person who sits in front of a computer to complete a complex information processing task, then different characterisations may be, more appropriate. Who are the individual users of GIS? The first generation of users were primarily environmental scientists and agencies. They tended to be academics, researchers - and their students -, government mapping agencies, utilities companies, etc. They were often not only the first users but also the early software developers. They may have had knowledge of several types of GIS or a specific understanding of one type of application. In terms of their knowledge, they tended to have a high degree of expertise in both spatial issues and computer skills.

While the application of GIS is at present restricted, it is argued that in a short time they will make available map-making and spatial analysis facilities to a much wider community of users - a community who may not be as adept at spatial reasoning (or computer use) as environmental scientists, surveyors, cartographers, etc. GIS may increasingly be used to educate and communicate with planners, politicians and the public. Openshaw et al. (1990) have commented that "...GIS is quite different from many other types of computer activity in that it is essentially end-user technology and the end-users are increasingly not computer specialists". The needs of this new generation of users may make demands above all on interface design and making the current functionality of the software more accessible. For example, Raper and Bundock's (1990) purpose in designing a "layered user interface" is to allow access to the functionality of the software to multiple groups of users -infrequent or job-limited users, the spatially aware professional, the database administrator, as well as the system developer.

Characterising the user in terms of prior knowledge or expertise is a common method for identifying the information-processing demands of complex computerised tasks. Most analysts of human-computer interactions (Card, Moran & Newell, 1983; Shneiderman, 1987; Norman & Draper, 1986; Wearn, 1989) recognise that expertise in computerised tasks can vary along two dimensions - expertise in the task to be completed and expertise in computer use. Consequently, users may be high on task knowledge and on computer knowledge, high on one but not the other, or low on both types of knowledge. And, of course, there is a wide range of middle positions. It is not surprising that human-computer interaction theorists refer to a community of users, characterised by diversity rather than homogeneity (e.g. Frase, Keenan & Dever, 1980; Holynski, 1988; van Muylwijk, van der Veer & Waern, 1983; Visvalingam, 1988).

What might this mean for GIS? It is usual for the first generation of users of a computer technology to be experts in both task and computer skills - and, as already mentioned, this is probably true for GIS users as well. Thereafter, the user community normally widens to include those who have considerable expertise and insight into the data at issue (in this case, cartographers, geoscientists, planners, demographers, etc), but whose computer experience is limited. If people are experienced in the task domain they bring with them prior knowledge from their previous mode of operation. Prior knowledge can have mixed effects. It can facilitate transfer and make learning easier, or it can result in failure to exploit the new opportunities which the computerised version of the task presents. It can even result in frustration, if the functionality of the software does not meet the expectations of the expert in the task domain. The main challenge for this type of user is managing the computer: gaining general conceptual knowledge about how the computer works (developing a mental model of the system), learning what commands to use, what keys to press, and learning the interface language (command driven, menu driven, direct manipulation).

On the other hand, many users will come to GIS with substantial experience of computer software - word processing, business graphics, spreadsheets, databases, statistics packages. Their mental models of computers may be well developed and they may know about different types of interface interactions; but their knowledge of the structure of spatial databases, mapping conventions, and spatial analysis may be scanty. Many new users of GIS applications may fall into this category.

Other users will be true novices, whose first experience of computers is a GIS. They come to GIS to deal with spatial issues, e.g. geography or social science students. Although many references are made to the special needs of novice users of GIS, we have very little systematic information about what they are. What is missing in current discussions about GIS is any substantial and systematic evidence about how a novice learns a GIS, what difficulties are encountered, how expertise is acquired, the influence of prior knowledge, and so on.

A further distinction can be drawn between interactive GIS users and those who only read GIS printouts. Printed versions of GIS map displays may be used to communicate data to fellow professionals, to present arguments to policy-makers about the allocation of resources, and to illustrate textbooks in a variety of disciplines. The importance of this type of user should not be underestimated, even though some may argue that they are not 'true' GIS users.

This listing of GIS users, and the distinctions which we have drawn, are purely speculative. The lack of systematic evidence about them makes it difficult to predict (except by analogy to the spread of other technologies) who the next generation of users are likely to be, and what their training needs will be. If this is a fair assessment of the state of our knowledge, certain interesting questions are raised. What is the driving force behind GIS developments? In what sense (if any) are the technological developments userdriven? What model of the user (albeit implicit) is in the developer's mind?

VISUALIZATION - ANOTHER USER VARIABLE?

Within the context of information technology the term visualization has many different meanings. Sometimes, it is used to refer to anything that appears on a VDU screen: or it is used to characterise a particular type of interface design - graphical user interface (GUI, WIMP, WYSIWYG, direct manipulation). It can also be used in a more precise way, to refer to techniques (and algorithms) which produce map-like displays: 2d plots, remote sensing images, terrain elevation models, etc. Because spatial concepts, and references to visualization, penetrate discussions of GIS at many different levels, it is important to clarify the meaning in use at any one time.

From the cognitive point of view, visualization techniques are successful only in so far as they achieve communication. Because of the intrinsic spatiality of GIS, and the need for screen visualizations to make the processes and outcomes of GIS transparent to the user, GIS places heavy information-processing demands on powers of spatial reasoning. To expand this point, we would like to draw attention to some experimental studies in cognition which demonstrate how user differences can affect the interpretation of other types of complex graphical displays. These studies also show how user responses can be assessed.

One of the most striking and consistent findings in cognitive psychology over the past few years is the importance of expertise in complex information processing, particularly in the interpretation of visual and graphical displays. Why is expertise important? It may seem an obvious point that experts are better at a given task than novices - they know more. But how does this knowledge affect their performance? Do they "see" the same information in a display as novices do? Do they extract the same information from a display and follow the same steps in solutions to problems as do novices or less experienced people? In cognitive processing terms, we ask whether experts' mental representation of a display and/or solution processes are similar to that of novices. (It should not be forgotten that when we contrast experts and novices we are sampling the extremes of a dimension of expertise. It is equally important to examine intermediate states of knowledge if we are to develop a better understanding of how expertise is acquired and how it can be trained.)

How can we quantify the performances of experts and others in order to make systematic comparisons? A number of experimental techniques are commonly used. Subjects are asked to exercise their skill in the normal way. They are given a mini-problem to solve which is very similar to the types of problem which they encounter as part of their everyday practice. Their responses are timed and a record is made of their solution attempts. But because much of what is interesting about their cognitions is going on inside their heads, so we shall want to externalise their thought patterns. This can be done by the method of "thinking aloud" or verbal protocol analysis. Subjects are asked to give a running commentary of their thoughts and decisions and these are recorded verbatim and then analysed. Verbal protocol analysis (Ericsson and Simon, 1984) is clearly very time-consuming but it does provide in-depth qualitative information on cognitive processes. It can be very helpful in finding out how experts and novices spend their time on a graphical display and why they generate certain solutions.

The classic expert studies were reported by de Groot (1965,1966) in the midsixties when he compared master chess players with less experienced players. Chess is a good example of a dynamic visual display, in the sense that the board positions are constantly changing and must be interpreted and re-interpreted as both players make their moves. (It differs from other types of display as control of the display is not in the hands of a single player - but that need not concern us here). From thinking-aloud protocols, de Groot reported surprisingly few differences between his master and less experienced players, in terms of time and numbers of moves considered. What distinguished the experts was the "quality" of their moves, particularly of the first move. This led de Groot to develop a technique to examine the knowledge structures of the players which he thought were determining the nature of the moves: memory for chess positions. He asked the players to view briefly (5 secs) chess-board positions in mid-game and then to reconstruct the board positions from memory. The master players' memory for the chess positions was far superior to that of the novice players; the superior memory could not be attributed to visual memory alone because, when random board positions were used, recall was equally poor for both masters and novices. Subsequent studies by Chase and Simon (1973) have confirmed that what characterises expert chess playing is pattern recognition ability: the experts have the ability to "chunk" the pieces on the board into meaningful wholes. In other words, they do see a different board than the novices do. This chunking sustains their memory performance in the memory task and, ultimately, determines the moves which they make in the course of a chess game.

Using these methods, De Groot's findings about the superior pattern recognition abilities of master chess players have now been replicated for many different types of experts. Egan and Schwartz (1979) reported that experienced electronic technicians could memorise and redraw symbolic drawings (electrical circuit diagrams) in ways which indicated that they were chunking the wires in the drawing into functional units or layers. For example, the skilled technicians knew that a power supply is likely to include a source, a rectifier, a filter, a regulator, etc. This conceptual category resulted in experts grouping and chunking the units which allowed them to search the drawings more systematically. When recalling building plans, experienced architects (Akin, 1980) produced a hierarchy of patterns - local patterns such as wall segments and doors were first recalled, then rooms and other spaces, and finally clusters of spaces. In a thinking-aloud study, Lesgold (1984) reported that, compared to medical students, experienced radiologists showed different cognitive processes when making a diagnosis from an X-ray film. Radiologists "see" a patient when they look at a film, not just a complex visual stimulus. They zoom in on target features of the films, while t novices are preoccupied with properties of the X-ray itself

If we turn to the study of hardcopy maps, a similar story emerges. While most cartographic evaluation research has concentrated on the design features of maps and maplike stimuli, a small number of studies have compared map users with different levels of geographical experience (e.g. Chang et al. 1985; Gilhooly, Wood, Kinnear and Green, 1988; Williamson and McGuinness, 1990). All these studies report differences between novice and expert map-users. For example, Williamson and McGuinness (1990) simply asked their subjects to describe portions of Ordnance Survey maps. The less experienced geography students (and non-geographers) concentrated on the surface details of maps - on colours, names of places - and just listed the names of map features without further elaboration. In contrast, the more experienced geographers evaluated the map features in terms of their quality, incidence and distribution, they located the map features within a spatial frame of reference, integrated and interrelated discrete features. Like the experienced radiologists, the expert geographers were able to "see" the reality behind the map.

What these studies show is that pattern recognition ability has a major impact on how a graphical display is interpreted and searched, and, ultimately, on the problem solutions generated from that display. Pattern recognition ability is deeply embedded in the knowledge structures of experts.

GIS map displays differ from these other examples in important ways. They are not static single-event graphical displays. They are the graphics capability of a spatial information system. They allow the user to interact with the database and to create maps; not just to learn facts about spatial data, but to hypothesise, analyse and synthesise. In so doing, however, they rely heavily on the pattern-making and pattern recognition abilities of the user. Users can become map-makers, which implies some (even rudimentary) knowledge of cartographic design principles. Users can become spatial analysts, which demands pattern recognition of the incidence and distribution of features as a result of a query; spatial inference and hypothesising about the next database query; spatial memory in a sequence of queries; complex pattern recovery as a consequence of overlaying, and so on. It is probably fair to say at this stage that we are not at all clear about the spatial information processing demands of even the simplest GIS tasks.

THE RESEARCH PROGRAMME

Our research programme on how users respond to map displays in a GIS environment was born of these concerns and we have responded to them in the following general ways.

1. We adopt a firmly cognitive perspective of the user - a model of "GIS user as thinker".
2. What distinguishes GIS from other information systems is spatial data, spatial display, spatial analysis; consequently, we have chosen to focus on that part of the human GISinteraction which exemplifies the spatial nature of the interaction - the map display.
3. When a GIS user is construed as a thinker, who responds to a map display in terms of his or her own cognitive structures, considerable variability can be expected in response to the same map.
4. Because of the widening community who are likely to become GIS interactive users or to be presented with GIS printouts, our research is studying users with different levels of expertise - expert GIS users, advanced geography students, novice geography students, and non-specialist graduates.
5. Psychological research is often accused of designing experimental tasks to study which are so trivial that they reveal very little about how real users perform in real situations (Stringer, 1974). In our research we have tried to develop tasks which mimic as closely as possible the real use of a GIS.
6. With the advent of GIS, it will be increasingly possible for non-cartographers to produce their own maps. The research will begin to examine not only how users interpret maps, but also how they create them.
7. Our methodology is designed to examine the knowledge structures of developing expertise, as well as performance measures.

The research consists of two stages: a development stage and an experimental stage. The goal of the development stage is to prepare problem 'scenarios', experimental map displays, and suitable (ecologically valid) map evaluation tasks for use in the experimental stage. The experimental stage tests the map displays on two types of user (experts and novices) in two contexts (interactive use and use of printed hardcopy). Two types of spatial databases are used to generate problem scenarios for study with different types of users - census data for Northern Ireland and an 'urban' database (accessed through the Ordnance Survey Northern Ireland).

Six experiments are planned. The purpose of the first two experiments on the census maps and urban maps (the Interactive Create-Map studies) is to collect detailed, indepth information (verbal protocols) on the cognitive processes of a small group of experts and novices (N=7/8 of each). Experts in these studies are defined as very experienced users of the system, and novices are less experienced users who have sufficient knowledge to interact with the system without extensive training. From the protocol analyses preliminary comparisons will be made between expert and novices in terms of their search strategies on the map, the content and organisation of their map descriptions, and the critical dimensions they use to compare and contrast map displays. These data will contribute to the formulation of a model of "map user as thinker". Also from these studies a sample of expert-generated and novice-generated maps will be available for subsequent experiments.

The next four experiments are designed to extend the findings of the protocol studies to a larger sample, and they focus more directly on the effects of alternative map displays. In the first two studies (the Interactive Map Display studies), experts (advanced undergraduate/postgraduate geography students) and novices (1st year social science undergraduates) will be presented with one of three alternative map displays (which will vary in the number of features/attributes and will be chosen from those generated in the protocol studies). One hundred and twenty subjects will be individually tested, 20 in each condition (2 user groups x 3 map displays). After a short period of pre-training on how to interact with the map display, the subjects will be required to complete search tasks, map descriptions, and to compare the target map with the other displays in the experimental set. The same group of students will participate in both the urban map and census map experiments. From the Interactive Map Display Studies, conclusions can be drawn about the relative efficiency of alternative map displays; the sources of map misunderstanding (e.g. complexity, clutter) can be identified for different types of user. The preliminary conclusions from the protocol studies about expert/novice search strategies and maprelevant cognitive organisation can be assessed on a large sample of subjects and generalisations, and qualifications, can be made.

The last two experiments (the Printed Map Display studies) will compare the effects of alternative printed map displays on groups of users who might encounter printed versions of computer-generated maps in textbooks, journals, and reports. Experts (advanced undergraduate/postgraduate geography students) and two groups of novices (A-level geography students, and non-geography graduate civil servants) will be presented with one of four alternative map displays (one of which will be printed in black/white). They will complete comprehension questions, describe their target maps, and compare their target maps with the other displays in the experimental set. One hundred and eighty subjects will be tested in small groups, 15 subjects in each condition (4 maps x 3 user groups). The same subjects will participate in both the urban maps and census maps studies. From the Printed Map Display Studies, statements can be made about the relative efficiency of printed map displays for different types of user and the sources of misunderstanding can be considered. Particular attention will be given colour/black-andwhite comparisons. From the results of the map comprehension, map description and map comparison tasks, conclusions can be drawn about the map-relevant cognitive organisation of the three different types of user studied in these experiments. The conclusions from these studies can be readily compared with the current literature on hardcopy map evaluation.

Comparisons can also be made across the studies. (1) A sample of advanced geography students will have participated in all six studies. As well as being able to compare their performance in the different experimental paradigms, substantial data will be available for this group to begin to formulate a theory of GIS spatial understanding. (2) Data will be available from a wide range of different users in the six studies - GIS experts, geography students at different stages in their education, social science students, and nonspecialist graduates. It will be possible to trace the pattern of acquiring expertise in map understanding and map use and to make associated statements about education and training. (3) Comparisons will also be made between the two problem scenarios - the census maps and the urban maps. Urban maps show concrete geographical features, while census maps display quantitative information. Conclusions can be drawn about user responses not only in terms of display formats, but also in terms of the relative concreteness/abstractness of the displayed information.

REFERENCES

Akin, O. (1980), Models of architectural models, London: Pion.

Burrough, P.A. (1986), Principles of geographical information systems for land resources assessment. Oxford: Clarendon Press.

Card, S.K., Moran, T.P. & Newell, A. (1983), The psychology of human -compute interaction. Hillsdale: Erlbaum.

Chang, K., Antes, J., Lenzen, T. (1985), The effects of experience on reading topographic relief information: analysis of performance and eye movements. The Cartographic Journal, 22, pp.88-94.

Chase, W. & Simon, A. (1973), Perception in chess. Cognitive Psychology 4, pp.55-81.

De Groot, A. (1965), Thought and choice in chess. The Hague: Morton & Co.

De Groot, A. (1966), Perception and memory versus thought: Some old ideas and recent findings. In: Kleinmuntz, B. (Ed.). Problem solving. New York: Wiley.

Egan, D. E. & Schwartz, B. J. (1979), Chunking in recall of symbolic drawings. Memory and Cognition, 7(2), pp. 149-158.

Ericsson, K.A. & Simon, I-LA. (1984), Protocol analysis. Cambridge, Mass.: MIT Press. Frase, L.T., Keenan, SA. & Dever, LJ. (1980), Human performance in computer-aided writing and documentation. In: Kolers, P.A., Wrolstad, M.E. & Bourna, H. (Eds.) Processing of visible language 2. London: Plenum Press.

Gilhooly, K.J., Wood, M., Kinnear, P.R.& Green, C. (1988), Skill in map reading and memory for maps. Quarterly Journal of Experimental Psychology, 40, pp. 87-107.

Holynski, M. (1988), User-adaptive computer graphics. International Journal of ManMachine Studies, 29, pp. 539-548.

Lesgold, A. (1984), Acquiring expertise. In: Anderson, J.R. & Kosslyn, S.M. (Eds.), Tutorials in learning and memory. San Fransisco: Freeman.

Muylwijk, B.van, Veer, G. van der, & Waern, Y. (1983), On the implications of user variability in open systems. An overview of the little we know and the lot we have to find out. Behaviour and Information Technology, 2(4), pp. 313-326.

Norman, D.A. & Draper, S.W. (1986), User-centered system design. New perspectives on human-computer interaction. Hillsdale: Erlbaum.

Openshaw, S., Cross, A., Charlton, M. & Brundson, C. (1990), Lessons learnt from a post mortem of a failed GIS. Proceedings of the Second National Conference of the Association for Geogriphic Information: ’GIS-The key to managing information’, Brighton.

Raper, J.F. & Bundock, M.S. (1990), GIS user interfaces: a window on the future Proceedings of the Second National Conference of the Association for Geographic Information:’GIS-The key to managing information’. Brighton.

Shneiderman, B. (1987), Designing the user interface. Strategies for effective humancomputer interaction. Worldrigham: Addison-Wesley.

Stringer, P. (1974), A use of repertory grid measures for evaluating map formats. British Journal of Psychology, 65(l), 23-34.

Visvalingam, M. (1988), User interface design: differing requirements of novice, occasional and expert users. University Computing, 10, pp. 80-85.

Wearn, Y. (1989), Cognitive asspects of computer supported tasks. New York: Wiley.

Williamson, J. & McGuinness, C. (1990), The role of schemata in the comprehension of maps. In: Gilhooly, K.J., Keane, M.T.G. & Erdos, G. (Eds.) Lines of thinking, volume 2. New York: Wiley.

Visualizing Quality Begins with Epistemology

(A discussion paper for NCGIA Initiative 7 - Visualization of spatial data quality)

Denis White
METI, US EPA Environmental Research Lab
200 SW 35th St.
Corvallis, Oregon 97333

Visualizing spatial or map data quality starts with the nature of phenomena and processes that are to be visualized. A preliminary attempt at understanding the relationship between phenomena and their cartographic visualization is formulated in the outline below (but was originally "visualized" as a tree):

- Nature of phenomenon
  - Continuous surface
    - Sampling pattern
      - Irregular
        - Regularizing technique or surface prediction method
          - Interpolation
        - Trend surface
          - Binomial smoothing
        - Graphic technique
      - Regular
        - Graphic technique
  - Discrete units
    - Type of unit
      - Point
        - Graphic technique
      - Line
        - Graphic technique
      - Area
        - Area definition
          - Proximal (i.e., Thiessen, Voronoi)
            - Graphic technique
          - Conformant (i.e., choropleth)
            - Graphic technique

The original diagram included a short list of typical graphic techniques at the appropriate points in the tree. Techniques were followed by names of Harvard Laboratory for Computer Graphics programs that performed them.

The distinction between conceptualizing phenomena as continuous variation or as discrete objects seems more basic to visualization than the dimensions of symbolization developed in cartographic literature. This distinction then precedes the two-way classification of symbols by point, line, and area versus measurement class in Robinson et al (78); by point, line, and area versus image or graphic variable in Bertin (81); or by spot, band, and field versus extent, darkness, and count in Fisher (82).

Making the continuous-discrete distinction more prominent in cartographic visualization echoes other conceptual debates about spatial structure. For example, in vegetation ecology the concepts of gradient (e.g., McIntosh 67) and community (e.g.,Daubenmire 68) contrast in a similar way. The imagery debate in psychology (Block 81) addresses mental constructs of spatial images -whether a pictorial representation is somehow manipulated in the brain.

Some arguments in the geography and geographic information systems literature have debated a similar distinction but from quite different points of view. Chrisman (78) helps place the cartographic data structure debate in the context of broader issues in the philosophy of science. The continuous-discrete terminology is reversed, however, because the focus is on the medium rather than the message. Continuous space is associated with discrete object models and discrete space with atomized objects rather than with a discretized continuous field of phenomenon, thus ignoring the epistemology of object. The subsequent argument on behalf of the topological object-oriented model begs the question of what is an object.

Peuquet (88) introduces computer vision concepts of space to the debate on models of geographic space and their cartographic data structures. The argument for a dual basis in representation, by location and by object, seems weakened or misplaced, however, by starting with the computer science entity-relationship model and associating entities with locations or geometric objects, respectively. The spatial character of the phenomena being modeled is little considered even though it provides the strongest justification for alternate models.

Regionalization and classification debate in geography has recognized the issue of unit of analysis or representation. Hartshorne (39) judges the geographic "individual" to be best represented as a farm, slighting many other conceptions of geographic objects and fields. Harvey (69) and Grigg (65) also discuss the issue.

The development of error models for geographic analysis is sharpening some of these debates. Goodchild (89) argues for the primacy of field models of error for some phenomena by explicitly invoking an assertion about the spatial nature of these phenomena. One part of the argument suggests that cartography tends to favor discrete object models of phenomena because of its technological history.

Spatial statistics models and methods reveal similar contrasts. Cressie (88), in a comprehensive review of spatial prediction methods, starts with a data model that is continuous both in geographic and attribute measurement space. Griffith (88) is representative of models and methods based on partitions of space into discrete objects.

In probability sampling, Overton et al (90) explicitly recognize a distinction between discrete and extensive natural resources. Finite population estimation for discrete resources can be complemented in the continuous case by translation of the continuous universe of an extensive resource into a discrete structure by spatial partitioning or by model-based estimation derived from point sampling.

These many instantiations of a similar conceptual distinction reinforce its fundamental nature. Visualizations of the quality of data should start with such a distinction.

References:

Bertin, J. 1981. Graphics and graphic information-processing. New York: Walter de Gruyter.

Block, N.J., editor. 1981. Imagery. Cambridge, MA: The MIT Press.

Chrisman, N.R. 1978. Concepts of space as a guide to cartographic data structures. Proceedings of the First International Symposium on Topological Data Structures for Geographic Information Systems, G.H. Dutton, ed. Laboratory for Computer Graphics and Spatial Analysis, Graduate School of Design, Harvard University.

Cressie, N. 1988. The many faces of spatial prediction. Preprint 88-13, Statistical Laboratory, Iowa State University.

Daubenmire, R. 1968. Plant communities. New York: Harper and Row.

Fisher, H.T. 1982. Mapping information: the graphic display of quantitative information. Cambridge, MA: Abt Books.

Goodchild, M.F. 1989. Modeling error in objects and fields. Accuracy of spatial databases, M.F. Goodchild and S. Gopal, ed. London: Taylor & Francis.

Griffith, D.A. 1988. Advanced spatial statistics. Boston: Kluwer.

Grigg, D.B. 1965. The logic of regional systems. Annals,Association of American Geographers, 55:465-491.

Hartshorne, R. 1939. The nature of geography. Lancaster, PA: Association of American Geographers.

Harvey, D. 1969. Explanation in geography. New York: St Martin’s Press.

McIntosh, R.P. 1967. The continuum concept of vegetation. Botanical Review, 33:130-187.

Overton, W.S., White, D., Stevens, D.L. 1990. Design report for EMAP, Environmental Monitoring and Assessment Program. Department of Statistics, Oregon State University, Corvallis.

Peuquet, D.J. 1988. Representation of geographic space: toward a conceptual synthesis. Annals, Association of American Geographers, 78(3): 375-394.

Robinson, A.H., Sale, R.D., Morrison, J.L. 1978. Elements of Cartography, 4th edition. New York: John Wiley and Sons.

COMMENTS ON THE VISUALIZATION OP THE QUALITY OF SPATIAL DATA

by

Joel Z. Yan[1]
Geography Division
Statistics Canada
Ottawa, K1A0T6
FAX: 613-951-0569

Submitted to Research Initiative 7
of the National Centre for Geographic Information and Analysis

1. BACKGROUND

Data quality measurement and reporting is vitally important to a national statistical agency such as Statistics Canada. As a result, a number of policies and procedures related to data quality implementation have been developed. Rather than prepare a formal position paper I have chosen to submit two items:

1) very brief and practical comments vis-a-vis research priorities in the area of spatial data quality reporting and visualization;

2) an annotated bibliography describing some of the relevant documents produced at Statistics Canada which could be of some value to other agencies, particularly data producing agencies. Comments would be welcome.

II. RECOMMENDATIONS REGARDING RESEARCH PRIORITIES

There seems to be quite a large gap between the data quality information recommended (as part of the Spatial Data Transfer Standard, for example) and the data quality information actually being supplied by many spatial data suppliers. Not too much practical research work seems to have been published in terms of checking with data users on exactly what information they would like to see regarding the quality of the spatial data before they decide to use the data.

In my opinion, some issues which would need further research:

- what aspects of data quality information is most critical to potential data users? is there a common set or does it vary completely with the application and the data type?

- what is the best way to present this information to potential data users? narrative lineage reports? tables describing positional accuracies with various measures? graphic representations of the accuracy? For each type what are the best (or good) examples?
- how to efficiently maintain lineage and accuracy records for data sets which undergo regular revision, perhaps using a number of update sources at various locations?

One specific proposal is for the NCGIA, with participation from others who attended the specialist meeting, to assemble in a single or two documents a compendium of existing standards, recommended guidelines, and good examples for:

1) narratives on spatial data quality ; and
2) visualizations of spatial data quality.

This would both support the NCGIA research initiative objectives and also move the state of the implementations forward. It could serve as an excellent and well used reference document for those involved with presenting or reporting spatial data quality.

III. REFERENCES

We have found the references below to be useful in our work at Statistics Canada. The references are organized into four categories:

[1] The views expressed herein are those of the author and do not necessarily reflect the views or official policy of Statistics Canada.

a: general policy documents related to Data Quality
b) specific methods for reporting spatial data quality;
c) statements of data quality and methodology produced in the Geography Division of Statistics Canada (STC); and
d) additional references regarding data quality of geographic products.

Several of the references are annotated as to their content in parentheses after the citation.

Most of these documents are available for reference from the Chairperson of the Geography Task Force on Standards and Quality, Geography Division, Statistics Canada, Ottawa, K1A OT6.

A. GENERAL STATISTICS CANADA DOCUMENTS RELATED TO DATA QUALITY

Fellegi, I. [1986] Policy on Informing Users of Data Quality and Methodology, Statistics Canada, April 11, 1986. (the general departmental policy and guidelines which indicates all data products should he accompanied by statements which inform users of the methodology used and indicators of the data quality)

Statistics Canada. [1987].
Quality Guidelines- Statistics Canada, April. (recommended concepts and techniques for implementing quality assurance and quality control in general in a sample survey operations)

B. DEVELOPING METHODS FOR REPORTING SPATIAL DATA QUALITY AT STC

Geography Task Force on Quality Reporting [1988] Final Report, Statistics Canada, Geography Division, working document, April 19, 1988 (the initial report which proposed methods for implementing spatial data quality measurement and reporting in Geography Division)

Lundin B., Yan J., and Parker, J.P. [1989] "Data Quality Reporting Methods for Digital Geographical Products at Statistics Canada, Proceedings of the First National GIS Conference, Ottawa, pp. 236-251.

Statistics Canada. [1991]. "Directive Number 2: Statements of Data Quality and Methodology for Geographic and Cartographic Data Disseminated by the Geography Division", 7 pages, Geography Division Policy and Procedures Manual. Volume One. June. (the revised policy for producing quality statements for products from within Geography Division)

C. STATEMENTS OF DATA QUALITY FOR SPECIFIC GEOGRAPHY STC PRODUCTS

Geography Division [1990a] Detailed User Guide, Postal Code Conversion File, January 1990 .Version, June 1990

Geography Division [1990b] Draft Statement of Data Quality and Methodology for the 1986 CSD CARTLIB Digital Boundary File, May 1990.

Geography Division [1989] Detailed User Guide, Postal Code Conversion File, January 1989 Version, June 1989

D. OTHER SPATIAL DATA QUALITY REFERENCES WE HAVE USED

Canadian Council on Surveying and Mapping (CCSM), [1984] Standard for the Exchange of Digital Topographic Data, Volume I, Data Classification, Quality Evaluation, and EDP File Format, Energy, Mines and Resources, Topographical Survey Division, Surveys and Mapping Branch.

Moellering, Harold, et al., [1987] A draft proposed Standard for Digital Cartographic Data, Issues in Digital Cartographic Data Standards, Report No. 8, July 1987.

National Committee for Digital Cartographic Data Standards (NCDCDS)
[1988] The Proposed Standards for Digital Cartographic Data, The American Cartographer, Journal of American Congress on Surveying and Mapping, Volume 15:1, January 1, 1988

Office of Population Censuses and Surveys. [1990]. Quality Assurance Standards. Project Management Unit. Census Operations Support Branch. September.

U.S. Geological Survey [1988] Digital Cartographic Data Standards, Procedure Manual (Draft/RDR/9-88), Department of the Interior, National mapping Program, Technical Instructions