

# UC Santa Barbara

## NCGIA Technical Reports

### **Title**

Accuracy of Spatial Databases: Initiative One Specialist Meeting Report (89-1)

### **Permalink**

<https://escholarship.org/uc/item/9wm428w2>

### **Author**

Goodchild, Michael F., editor

### **Publication Date**

1989

**National Center for Geographic Information and Analysis**

**Accuracy of Spatial Databases  
Initiative One Specialist Meeting Report**

Compiled by

Michael F. Goodchild

National Center for Geographic Information & Analysis  
University of California  
Santa Barbara, CA 93106

Technical Paper 89-1  
January, 1989

# NATIONAL CENTER FOR GEOGRAPHIC INFORMATION AND ANALYSIS

## RESEARCH INITIATIVE 1:

### Accuracy of Spatial Databases

#### REPORT OF SPECIALIST MEETING

Michael F. Goodchild, UCSB, 1.8.89

The specialist meeting for this first initiative of the NCGIA was held at the Casa de Maria in Montecito, CA from December 13 through 16, 1988. Participants were drawn from the three Center sites, from universities in North America and Europe, and from numerous federal agencies and companies active in the GIS field. Disciplines represented included Geography, Mathematics, Statistics and several branches of Engineering. The full list of attendees and meeting program are attached to this report.

The stated goals of the specialist meeting were to:

- o render a state-of-the-art review of issues in accuracies of GIS (spatial) databases which can be published as a workshop proceeding;
- o survey outstanding database accuracy research issues (the tall poles) that could or should be on the research agenda for NCGIA and other investigators during the next two years;
- o identify the tallest poles of highest importance as impediments to more improved and effective use of GIS for which a good level of research closure can be reached in a two-year effort and which could constitute the core of the NCGIA research effort; and to
- o develop the research strategies needed to effect the second and third items above in sufficient detail that there is assurance that these problems are indeed tractable in the two-year timeframe allotted.

The purpose of this report is to summarize the results of the meeting's discussion of research issues and strategies. A more complete record of the discussion is available in the form of the rapporteurs' reports on each session, and session recordings, and the formal presentations of the meeting will be available in due course as a proceedings volume.

The first part of the meeting consisted of presentations from participants representing three perspectives:

- o those with knowledge of the nature of error and uncertainty in spatial data, and the problems which arise from the use of inaccurate data, drawn from universities and federal agencies;
- o those with knowledge of the disciplines able to focus research on error problems, particularly spatial statistics; and
- o those with knowledge of the spatial data structures and algorithms capable of accommodating uncertainty and implementing research on error.

Towards the end of the presentations each participant was asked to list those topics considered most relevant to the development of a research agenda. These were then synthesized into a list of eleven topics which formed the basis for general discussion:

**1. Data structures and models.** Research is needed to develop models of spatial data and database structures which are sensitive to data accuracy, can represent accuracy explicitly and can support the tracking of error through spatial database operations. Many object-based models have no such sensitivity to data accuracy, resulting in numerous forms of artifacts. Specific topics discussed during the meeting included regular tessellation and hierarchical subdivision models, and the developing literature on robust methods of finite precision computational geometry.

**2. Models of error and distortion.** Research is needed on models which can successfully describe, characterize and parametrize error, both for spatial fields and for complex spatial objects. As a corollary, we currently lack adequate means to simulate error and distortion of much spatial data. The connections between error in fields or rasters, and error in objects derived or interpreted

from them, are not sufficiently understood. There may be non-statistical approaches to some of these problems, as models based on fuzzy sets and on spatial semantics may be appropriate.

**3. Error propagation.** GIS processes combine data from different sources with different levels of spatial resolution, using rules which are often complex. There is a growing interest in modeling using such diverse spatial datasets, and in many cases models are highly nonlinear, leading to error effects which are difficult to control. There is a need for research on how error propagates through each GIS process, particularly overlay. Propagation effects may be seen in terms of sensitivity, or the relationship between the product or output of a GIS process and uncertainty in the corresponding inputs. Methods used to model uncertainty propagation in reasoning systems may be relevant here. Tracking of errors is also an important issue for the user, who may need to be made aware of the origins of each type of error.

**4. Product uncertainty and sensitivity.** Uncertainty in a spatial database is ultimately reflected in uncertainty in the products of the database. Research is needed on methods for specifying acceptable levels of uncertainty in products, and for linking product uncertainty back to requirements for spatial database accuracy. In some cases the links between database error models and product confidence limits will be susceptible to analysis; in others, Monte Carlo simulation may be necessary. Extensive work of this nature will require improved ways of structuring complex GIS processes, including better taxonomies of GIS operations.

**5. Risk analysis.** criteria for acceptable levels of uncertainty in GIS products must be obtained ultimately from an analysis of the risks associated with decisions based on those products. Risk analysis is therefore an important link in the chain which stretches from error models and database concerns through to decision-making. Although analysis of risk is an accepted part of decision theory, it has not yet been applied to GIS in any systematic way. This research area forms the link between this initiative on spatial database accuracy and the forthcoming Initiative 4 on the use and value of spatial information.

**6. Accuracy concerns among users and agencies.** Moving away from basic toward applied research, there is a need to understand better the concerns of GIS users, data providers and land management agencies for data accuracy. What minimal standards for accuracy are being developed by agencies, what methods are being used to define and document data quality, and what interactions might be developed between these concerns and basic research on error modeling? There is a need to develop standard measures of quality which are compatible with error models, and can be determined and monitored at reasonable cost for standard data types. It would be useful to have standard benchmark datasets which could be used to measure the accuracy of various data entry processes.

**7. Experimentation and measurement.** There is an important area for research in the development of methods for measuring accuracy empirically. This includes methods for measuring the quality of digitizing and scanning, relative to source documents, as well as standard techniques for measuring accuracy of databases with respect to ground truth. Little is known about the design of efficient sampling schemes. Measures of accuracy developed from experiment should be designed to be easily interpreted. It would be useful to have software, which might be simple additions to standard GIS products, which could be used to monitor and measure data accuracy.

**8. Error reduction methods.** Strategies for error reduction would be useful in a number of areas. Research might be carried out on methods for minimizing error in digitizing and scanning, and in designing approaches to analysis and modeling which included reduction of error as a criterion.

**9. Interpolation and surface modeling.** The choice of data model has a significant impact on accuracy, but has more often been regarded as an issue of storage and processing efficiency. For example, the choice between DEM and TIN for storage of a topographic surface has a poorly understood impact on accuracy, as does the choice between field and object models, or raster and vector. In each of these cases the question of accuracy is linked to the nature of the spatial variation being modeled: the choice between DEM and TIN ultimately depends on the nature of the topographic surface being modeled and the erosion processes which formed it. Research is needed on the significance of accuracy versus storage and processing efficiency in choosing between different data models, and on the role of processes of spatial differentiation in this equation.

**10. Aggregation, disaggregation and modifiable areal units.** Several of the papers presented at the meeting dealt with the impact of reporting zones on the results of socio-economic modeling, and the more general question of the impact of the spatial data model itself. There is a need for further development of techniques which exploit the capabilities of GIS for investigating and controlling data model effects. This includes the effects of reporting zone aggregation and disaggregation, as well as issues of interpolation between incompatible zones. Research in these areas should be aimed both at increasing understanding of the importance of spatial data model effects, and at development of GIS-based techniques for dealing with them in practical applications.

**11. Regularity and stability.** One possible strategy for dealing with problems of scale and resolution in spatial data is to search for approaches which are comparatively free of such effects. For example, it was argued in one paper that the best models of spatial phenomena would be ones which were independent of scale, and a second presentation was concerned with ranges of scale over which certain phenomena (urban density in this case) showed a form of invariance. This area overlaps strongly with Initiative 3, which will be concerned with the representation of objects at multiple scales.

Discussion continued in two groups, one concerned with basic research and the other with applied and user-related research. The next two sections describe the topics identified by these two discussions. The first is the proposed agenda for the Center for the 18 month timespan of the initiative given its resources, and the second is a list of additional topics which were considered to be of high priority and on which significant progress might be achieved within 18 months. The Center will continue to look for additional resources and support arrangements which will allow as much as possible of the second list to be completed.

#### Center research agenda

**1 Models of error.** Work has already been initiated at Santa Barbara on models of error in fields and objects, and the relationships between them. This will include simulation, and the use of simulated datasets for modeling error in various GIS processes, and for examining data compression and storage options. We are currently investigating spatially autoregressive processes and Markov random fields, and techniques for realizing such processes over large lattices.

Deliverables: presentations and journal articles.

**2. Relative efficiency of TINs and DEMs.** Research will be initiated during 1/89 at Santa Barbara aimed at comparing surface storage models. Work is already under way on an algorithm for optimal selection of TIN vertices from a dense DEM. Surface models will be compared in terms of accuracy, as far as possible holding other criteria such as storage volume and processing efficiency constant. Of particular concern will be the relationship between accuracy and landform type: which types of landform favor TIN representation and which favor DEM? This work will be extended from topographic surfaces to other types of continuous and near-continuous spatial variation.

Deliverables: presentations and journal articles.

**3. Taxonomy of errors.** There is a need for a working taxonomy of spatial data errors, and associated methods for error handling, within each of the GIS application fields. Error might be classified by source, or by suitability for modeling or reduction, or by magnitude, across each of the types of data dealt with by spatial data handling technology. The taxonomy could include a discussion of the types of error known to be present in each major class of spatial data. It was felt that such a taxonomy could be of significant use to the GIS user community. Center work in this area will build on several existing attempts at taxonomies of spatial data and associated uncertainties.

Deliverables: Center monograph, journal article.

**4. Bibliography.** In association with the taxonomy, the Center could publish an annotated bibliography of literature on spatial data error.

Deliverables: Center publication.

**5. Short course.** As part of its program of professional courses, the Center will develop a one or two day course on the specific topic of spatial database accuracy, packaged so that it can be offered in a variety of settings, and identify a number of individuals willing and able to teach the course.

Deliverables: Course materials.

**6. Standard data sets.** The Center will work to identify, document and assist in distributing a series of standard data sets suitable for benchmarking and monitoring the accuracy of GIS processes, particularly digitizing. The data sets will represent a variety of data types and applications.

Deliverables: Center publication, documents, magnetic media.

**7. Network.** The Center will maintain a network of people interested in accuracy issues in GIS, using the participant list from the meeting as a base but adding names as they become known. Information about the initiative and related work will be circulated from time to time during the next 18 months.

Deliverables: Occasional mailings.

**8. Proceedings volume.** The Center will edit and publish the proceedings of the specialist meeting, within six to nine months, as a state-of-the-art text on accuracy issues in spatial databases.

Deliverables: Camera-ready copy to publisher.

In addition, research relevant to the accuracy of spatial databases will be undertaken as part of Initiatives 3 (Representation at Multiple Scales), 4 (Use and Value of Spatial Information), and 7 (Visualization of Uncertainty).

#### Other research topics of high desirability

**1. Text on spatial database accuracy.** This would be a compendium of current knowledge in the field. Burrough's chapter on data quality is a good base, but research in the area is currently so incomplete that a text is probably some years away. An edited volume might be developed at the end of the Center initiative period.

**2. Case studies.** It would be useful to have a set of researched and documented case studies illustrating approaches to error analysis, modeling and management in different GIS application fields.

**3. Propagation and sensitivity.** Systematic studies are needed to simulate and measure the propagation of error through the complete range of GIS processes, particularly where response to error is nonlinear. The Center does not currently have the resources to carry out such research on the scale needed. This should include an associated taxonomy of GIS operations, and might also extend to the implementation of certain types of standard error propagation analysis in software.

In addition to these, it was felt that the list of eleven research areas above contained many topics that were of fundamental importance, both as contributions to basic research and also as having potential value to the user community. It is likely that there will be active research in many of these areas in the next 18 months. While only a small part of this research can be carried out directly by the Center, we hope that the Center can be in communication with as much of this research as possible, and act as a node for exchange of information and research results.

In general, the Center-maintained network can function as a mechanism for exchange of information among the following:

- groups within the Center conducting research listed in the first part of the agenda above, and with Center support;
- individuals and groups within or connected with the Center, conducting research not supported directly by the Center, but relevant to the wider agenda of this initiative; and
- individuals and groups not connected with the Center, but conducting research relevant to the wider agenda.

While the Center has no direct influence on the latter two groups through the allocation of resources, we hope that we can nevertheless play a useful role in coordinating and catalyzing research within the broad limits of the agenda listed above.

## **Specialist Meeting on Accuracy of Spatial Databases**

**National Center for Geographic Information and Analysis  
University of California  
Santa Barbara, CA 93106**

**December 13-16, 1988**

-----

## **Abstracts**

### **Inherent and Operational Error Within a GIS: Evaluation From a User Perspective**

**Stephen Walsh**  
**University of North Carolina**

Inherent and operational errors contribute to a reduction in the accuracy of data contained within a geographic information system, and, thereby, affect the quality of the analysis and the resulting products that are generated by such systems. Inherent error is the error present in source documents. Operational error is produced through the data capture and manipulation functions of a GIS.

Every map contains inherent error based upon the nature of the map projection, construction techniques, and symbolization of the data. The amount of error is a function of the assumptions, methods, and procedures followed in the creation of the source map. Data represented as discrete or continuous, spatially and/or temporally interpolated from point to area measures, classified through various organizational approaches, and represented by spatial and/or temporal resolution considerations and resampling schemes further add to the level of inherent error within GIS base maps and documents. User requirements of the data and the relationship between the suitability of the collected information to the application and the acceptable level of inherent error of the product are important considerations.

Operational error, categorized as positional and identification errors, is introduced during the process of data entry and occurs throughout data manipulation and spatial modeling. The highest accuracy of any GIS product can only be as accurate as the least accurate data plane of information involved in the analysis. As the number of layers in an analysis increases, the number of possible opportunities for error increases. The upper and lower accuracy limits of a composite map indicate the range of accuracy probabilities when analyzing two or more thematic overlays, given the accuracies of the source map. Operational error reduces the accuracy of a GIS analysis and output product from its theoretical best.

By recognizing inherent errors within the products and the operational error created by combinations of input data, total error may be minimized. Questions regarding such types of errors and the implementation of recommended data quality standards helpful in documenting and evaluating elements of the data base is the responsibility of both the producer and the user of the data. Literature has suggested conceptual standards such as lineage, positional accuracy, attribute accuracy, logical consistency, completeness, and confidence factors, and methods of documenting and evaluating information--reliability diagrams, header information, statistical overlays of probability and confidence, numerical thresholds that relate data quality levels to specific products and to possible applications, and logical examinations of spatial co-occurrence.

Users are involved in developing research-specific thematic files and acquiring maps and digital files presumed to be appropriate to their analyses. Information regarding the quality of the thematic overlay needs to be available to the user for self-evaluation of the appropriateness of the file given certain applications and accuracy demands. Mean accuracy statements for the digital thematic files provide little use in trying to understand the spatial variability of error throughout the product suitable to a variety of different applications. Data linked to the digital file that is capable of relating specific sampling information to a minimum acceptable accuracy and the producer and user risk would add to the assessment of overlay quality by indicating the degree of quality. Statistical surfaces--probability, covariance, standard deviation--need to become a standard thematic overlay in a GIS analysis, since these measures relate to the concept of "fitness of use" decided upon by the user. User sophistication in evaluating the "fitness of use" will impact on the appropriate use of prepared overlays in GIS analyses.

Thematic overlays in a GIS can be produced in a manner that meets technical accuracy standards but still fails to transmit the needed information to the user. The choice of appropriate spatial interpolation methods, for example, must be decided upon by specific knowledge of the anticipated spatial analysis and the level of accuracy required. Improvements in data representation might be enhanced by implementing knowledge-based systems that lead the user through appropriate assumptions and decisions relating to the required accuracy levels, formats, and sampling approaches ascertainable to the general application being considered by the user.

### **GIS Product Accuracy**

**John Estes**  
**U.C. Santa Barbara**

While understanding the significance of theoretical work into the propagation of error in a GIS it is important to remember the verification of the accuracy of output products is also a critical issue in many science and applications environments. What is obvious to many working in the use of GIS technology in large area science and applications research and operational programs is that we may

be forced to use data inputs for which we have no idea of the spatial and thematic errors of at least several of the input products for a given analysis.

With this as a given, what are the most effective means of establishing the accuracy of a GIS output product? This question has both locational and thematic aspects. It involves the determination of area weighted and by-class accuracy statistics and their associated error bounds. While a variety of techniques have been employed to address this issue none have proven entirely satisfactory. The ease and efficiency of methods needs to be balanced against accuracy, costs and potential liability considerations. Methods combining field sampling with analysis of remotely sensed data appear to hold considerable promise in this area. Finally, while there is a need to understand the theory of error propagation within a GIS context, there is also the compelling and concomitant need to find the means to verify GIS product output. The spatial statistics of this issue must clearly be addressed in this initiative.

## **Real Data and Real Problems: Dealing With Large Spatial Databases**

**David Brusegard and Gary Menger**  
**The Institute for Market and Social Analysis (IMSA)**

The Institute for Market and Social Analysis (IMSA) develops and maintains large geographic databases for use in market and social research applications. This paper details a number of conceptual and practical problems which are encountered in developing consistent and accurate spatial databases. In particular, problems regarding the use of data derived from numerous sources at various levels of spatial aggregation are discussed.

## **The Small Number Problem and the Accuracy of Spatial Databases**

**Susan Kennedy**  
**U.C. Santa Barbara**

Often when the topic of the accuracy of spatial databases is introduced, discussion focuses on the location/positional accuracy of the data in question and not on other types of error which may be present in the data. The Small Number Problem occurs whenever a percentage, ratio, or rate is calculated for either a small geographic area with a small denominator or a large geographic area which is sparsely populated. In either case, small random fluctuations in the variable of interest (numerator) may cause large fluctuations in the resulting percentage, ratio or rate. This presentation will review the difficulties which random fluctuations cause.

## **The Measurement and Display of Residential Density**

**A. Stewart Fotheringham**  
**State University of New York**

**M. Batty and Paul A. Longley**  
**Department of Town Planning, University of Wales, Cardiff**

The beliefs that residential densities decline as distance from the center of a city increases and that these urban density gradients decrease over time seem to have reached the status of unquestioned tenets within urban geography, planning, economics and regional science. However, empirical support for these relationships is weakened by the ways in which residential density tends to be measured -- generally in aggregated areal units within an urban area defined by a subjectively drawn boundary. The advent of computer-generated geographic information analysis allows greater flexibility and precision in the examination of urban density relationships. In particular, we focus on the identification of appropriate continuous density estimation methods and on the removal of boundary effects by the identification of objectively-defined urban boundaries. What we term 'edge' effects will still be present in the data although these can be dealt with through the use of simple heuristics.

Much of the discussion introduces current research on the simulation of urban densities through the process of diffusion-limited aggregation and particular attention is given to identifying an urban limit within which analytical error is minimized.

## **Spatial Accuracy Induced Problems in the TIGER System**

**Jack George**  
**Geography Division, U.S. Census Bureau**



Errors in the location of features produce various geometric and topological problems in the feature network section of the TIGER data base. The topological problems in turn caused the misallocation of area attributes resulting in serious problems for continued data base development, automated map production and the assignment of population and housing information to the appropriate geography. The various types of problems and their effects are categorized.

## **Census Bureau Research Concerns for Accuracy of Spatial Data**

**Alan Saalfield**  
**U.S. Census Bureau**

The following two sets of issues relate to personal research interest and do not necessarily reflect major research goals of the Census Bureau.

Issues involving finite precision geometry: Maintenance of the TIGER database requires many geometric algorithms that must update map topology and geometry in a consistent and reliable manner. Considerable efforts have been made by researchers in the field of computational geometry to address the problems of geometric computations that must be made with limited precision arithmetic. I will present a brief survey of results and how they address specific map update problems.

Issues involving large sample surveys: The Census Bureau conducts many national sample surveys. The magnitude of the surveys dictates many "operational concessions" that must be made to ideal probabilistic and statistical strategies. GIS technology will permit the Bureau to improve its sampling methods. This issues section will focus on current practices and limitations and some of the new opportunities that TIGER and GIS technology will present to samplers at the Census Bureau.

## **Demand Point Approximations for Locational Problems**

**Rajan Batta**  
**State University of New York**

In this talk we will discuss the quality of demand point approximations for location problems. The importance of this research stems from the fact that Urban Planners often restrict facility location to be at the demand point or population centroids in their calculations. This makes intuitive sense since the demand point locations determine the eventual optimal locations of the facility, and hence if we consider restricting the set of candidate sites for facility location, the set of demand points is a natural choice. We will highlight work that has been done in this area. We will also discuss some other studies on insensitivities in Urban Planning and pinpoint some critical areas and issues for future research.

## **Impacts of Map Errors on Predictive Land Classification**

**Frank Davis**  
**U.C. Santa Barbara**

GIS-based map weighting and overlay are used routinely to predict land surface quality in terms of timber production, geomorphic stability, wildlife habitat quality, etc. The usefulness of such predictive land classification and mapping depends on both cartographic and ecological considerations, as well as complex inter-relations between the two. The cartographic considerations included the resolution, accuracy and bias of terrain maps, which depend in turn on the classification systems used to map categorical variables and on the precision at which continuous variables are measured. Ecological considerations include the strength, consistency and scale-dependence of relationships between terrain variables and predicted land surface quality.

This paper demonstrates some statistical techniques for assessing predictive misclassification due to cartographic errors, and provides two examples to illustrate the impact of such errors on predictive land classification. In the first example, information analysis of a spatial database is used to classify land surfaces in Santa Barbara County, California based on the association between mapped natural vegetation classes and categorical environmental variables. In the second example, a polychotomous logit model is developed to predictively map site potential for an oak species whose distribution has been reduced severely by historical burning and clearing. Non-random scale-dependent errors in topographic and vegetation maps lower the information content of the database and the value of resulting predictive models. Predictive models are improved significantly by masking error prone areas and smaller terrain facets.

## **Site Characterization Information Using LANDSAT Satellite and Other Remote Sensing Data -- Integration of Remote Sensing Data With Geographic Information Systems**

**William J. Campbell and Marc L. Imhoff**

**NASA, Goddard Space Flight Center**  
**Jon Robinson, Fred Gunther, Ron Boyd and Michael Anuta**  
**Computer Science Corporation**

A cooperative research project was conducted by NASA's Goddard Space Flight Center in conjunction with the Nuclear Regulatory Commission (NRC) and the Pennsylvania Power and Light company (PP&L) for evaluating the utility, accuracy and cost effectiveness of incorporating digitized aircraft and satellite remote sensing data into an operational geographic information system for facility siting and environmental impact assessments.

This research focused on the evaluation of several types of multisource, remotely sensed data representing a variety of spectral band widths and spatial resolution. High resolution aircraft photography, Landsat MSS, and 7-band Thematic Mapper Simulator (TMS) data were acquired, analyzed, and evaluated for their suitability as input to a GIS.

**Some Methods for Assessing the Accuracy of Spatial Data**

**Russell G. Congalton**  
**U.C. Berkeley**

A review of methods for assessing the accuracy of spatial data will be presented. This review will include discrete multivariate analysis techniques, sampling simulation, and spatial autocorrelation analysis. Most of the examples will incorporate remotely sensed data but will be applicable to other spatial data as well. Current work on elevation data will also be presented. Finally, suggestions about future work and ideas will be discussed.

**Observations and Comments on the Factors Contributing to Error and Variance in Digital Vector Data**

**Giulio Maffini**  
**Tydac Technologies**

Conversion of geographic information represented on paper maps is still the predominant method in the creation of digital vector data bases. What is the inherent accuracy of this transformation process? What are the factors which influence error and variance?

This presentation will review the results of in-house trials and experiments which attempted to isolate these factors. Comments and conclusions are presented on practical guidelines for inputting the accuracy of manually digitized data. Topics for further research are also identified.

**Sources of Error in Thematic Classification of Remotely Sensed Imagery**

**Jeffrey L. Star**  
**U.C. Santa Barbara**

Image processing systems for remote sensing applications almost always have functions for classification of multivariate data. Typically, quasi-continuous multispectral datasets are converted to a nominal dataset of land-use or land-cover categories, and this derived product often becomes a layer in a geographic information system. The input datasets are rarely examined to determine their underlying frequency distribution; we just assume that the data are normal enough, and that the deviations from normality are unimportant. It is not clear how deviations from a hypothetical multivariate normal might affect the power of the classification process. In a supervised classification, the number of training fields for developing a statistical description of a given class is usually arbitrary; it is unclear how small changes in the locations of the training fields affects the quality of the derived thematic information. In practical

applications of classification, one sometimes proceeds hierarchically, developing first level classifications for a small number of broad categories (perhaps in an unsupervised mode) and then further refining and subdividing only some of the initial classes.

It appears that there is a clear opportunity for a simulation study in this process. Our group has plans to identify some bounds on:

- the sensitivity of the end-to-end process to random noise in the input data channels
- the sensitivity of the process to elements of the selection of training fields
- the internal consistency of several distance measures in the classification algorithm in the face of random noise.

## **Inclusion of Accuracy Data in a Feature Based, Object Oriented Data Model**

**Stephen C. Guphill**  
U.S. Geological Survey

A digital spatial database can be considered as a multifaceted model of geographic reality. In such a model, entities are individual phenomenon in the real world; features define classes of entities; and a feature instance, one occurrence of a feature, is represented by a set of objects. One type of object, a spatial object, contains locational information. A second type, a feature object, contains non-spatial information. The objects have attributes and relationships. Attribute and relationship values can also have attributes. This recursive attribution scheme allows for the incorporation of accuracy information at any level of the geographic model.

Locational precision can be associated with each spatial object. Uncertainty measures can be assigned to any attribute value. Temporal attributes, along with uncertainty information, can also be included in the model. The mechanisms exist in data models and data structures to handle accuracy data. However the underlying questions remain: What are the uncertainty measures, how are they collected, and how are they used?

## **Error Modeling for the Map Overlay Operation**

**Howard D.F. Veregin**  
U.C. Santa Barbara

Geographic information systems permit a wide range of operations to be applied to spatial data, but too often these operations are applied with little regard for the quality of the resulting output products and the types of errors they may contain. The development of formal error models has unfortunately lagged far behind the growth of GIS technology per se and its perceived utility as a decision making tool. Few formal methods of error assessment have been developed and those which do exist are not now widely applied in practice. Consequently, information of unknown reliability continues to be used as a basis for decision making.

This paper critically examines some current approaches to error modeling for GIS operations, with particular emphasis on map overlay. A conceptual model is presented which identifies a five-level "hierarchy of needs" for GIS error modeling. Application of this conceptual model to map overlay serves to identify variables which may affect the types and levels of errors resulting from this operation and techniques which may appropriately be used in the assessment of these errors. Thus the source of the errors present in the input data (level I in the hierarchy) determines to a large extent the techniques which may be employed for modeling mechanisms of error propagation or production in map overlay (level III in the hierarchy). Successful modeling of these mechanisms is also dependent on the manner in which errors are measured and the comparability of the error measurements across different data layers (level II in the hierarchy). Strategies for managing and reducing error (levels IV and V in the hierarchy) are shown to depend in large part on the type of error propagation or error production model applied, which affects the inferences that may be made about the significance of errors in output products.

## **Accuracy and Bias Issues in Surface Representation**

**David M. Theobald**  
U.C. Santa Barbara

Surface representation is increasingly used in environmental modeling because of topography's dominant role in such processes. The accuracy of these elevation models is seldom addressed, and if it is, is usually constrained to an estimate of root-mean-square error in the vertical measure or a description of noise or striping errors. Seldom are the errors described in terms of their spatial domain or how the resolution of the model interacts with the relief variability. Additionally, when using an elevation model the research objective must define the land features that need to be captured and the precision with which they need to be represented.

These features are dependent upon the spatial frequency distribution of the terrain and how the resolution interacts with these frequencies. Thus, in defining the accuracy of a DEM, one needs to ultimately know the spatial frequency distribution of the terrain and the bias in the resolution in addition to the type of data structure used and accuracy of the source document.

A brief review of sources of DEMs, common data structures used in terrain representation, techniques used in the derivation of parameters (i.e. slope), and methods of interpolation will be given. The concept of accuracy will then be discussed in terms of the research objective, precision, resolution, and terrain variability.

## **Precision of Interpolated Maps**

**Paul Switzer**  
**Stanford University**

Statistical methods for describing and assessing interpolation precision for both qualitative and quantitative mapping variables will be outlined. These methods include model-based sampling theory approaches, Bayesian analysis, kriging, cross-validation, and resampling.

## **Frame Independent Spatial Analysis**

**Waldo Tobler**  
**U.C. Santa Barbara**

The results of an analysis of geographical data should not depend on the spatial coordinates used -- the results should be frame independent. This should also apply when areal units are used as the spatial data collection entity. Previous work has shown that some analysis procedures do not yield the same results under alternate areal aggregations, but some of these studies have used measures known to be inappropriate for spatial data, e.g., Pearsonian correlation instead of cross-spectral analysis. And there are some methods of analysis which do seem to yield frame invariant results, especially under alternate partitionings of the geographic space. In other cases it is appropriate to consider aggregations as spatial filters, with response functions which can be estimated a priori. There also exist linear spatial models which allow exact calculation of the effects of a spatial aggregation, so that consistent empirical and theoretical results can be obtained at all levels of spatial resolution. It is proposed that all methods of spatial analysis be examined for the invariance of their conclusions under alternative spatial partitionings, and that only those methods which show such invariance be allowed. Pentaminoes are suggested as one vehicle for such testing.

## **The Traditional and Modern Look at Tissot's Indicatrix**

**Piotr H. Laskowski**  
**Intergraph Corporation**

A new method of perception of Tissot's Theory of Distortions and the computation of the parameters of Tissot's Indicatrix, used to analyze map distortions of cartographic projections, is proposed. The new approach is based on the algebraic eigenvalue problem and makes use of the Singular Value Decomposition of the column-scaled Jacobian matrix of the mapping equations. The semiaxes of Tissot's Indicatrix are evaluated directly, and the usage of quadratic forms, which may cause unnecessary loss of numerical precision, is avoided. Several advantages of the new approach, versus the original Tissot's approach, are detailed.

The traditional concept of Tissot's Indicatrix is also examined for comparison. This original approach is presented here from the non-traditional point of view, as the study of a variation of the positive definite quadratic form under the action of a mapping transformation. The complete collection of the computational formulas of the original Tissot's approach (improved for numerical efficiency) is included.

## **An Experimental Field Laboratory for Testing the Accuracies of Spatial Databases**

**Alan P. Vonderohe**  
**University of Wisconsin**

During 1987 and 1988, the National Geodetic Survey performed a county-wide Global Positioning System (GPS) survey in Dane County, Wisconsin. As part of this effort, second-order, class I horizontal positions were determined for 18 permanently monumented points in a nine-section rural area. This very dense control network was the first step in the development of a field laboratory for experimental accuracy analysis. Since the GPS survey, control has been extended to an additional 88 points, including

the section and quarter section corners of two sections. This extended control formed the framework for comparison of the results of two digital cadastral mapping methods. It is expected that during 1989 actual boundary surveys will be performed for every parcel in one of the sections. This information is expected to be used in experiments which gradually degrade its accuracy until thresholds in decision-making are discovered. Eventually, the concept will be extended to resource and topographic mapping with the intent of developing high-quality information against which models of uncertainty can be tested.

## **Distance Calculations and Errors in Geographic Data Bases**

**Daniel Griffith**  
**Syracuse University**

Distance or separation measures for network link lengths or between areal unit centroids and points on a geographic surface often are used to identify land use accessibilities, for determining minimum paths through networks, to construct location-allocation solutions, or the such. To date this source of error has been all but ignored in either conceptual or analytical treatments of geographic analysis, while scholars are arguing that more analytical capabilities should be built into GISs. The purpose of this paper is to explore some elementary properties of this situation, in part building on work reported by Hillsman, Current and Schilling, and Hodgson. Expected values and variances of both additive and proportional error structures will be studied, for selected error distributions, and the presence or absence of bias will be documented. This investigation is a useful complement to those by Veregin and by Amrhein and Griffith.

## **Approaches to Determining and Correcting Areas of Unreliability in Geographic Databases**

**Peter F. Fisher**  
**Kent State University**

Errors of commission or omission are endemic to almost all maps that form the source of much of the spatial data input to GIS. Indeed, even if the data were collected in an original digital format, it is doubtful if it would necessarily be more complete. The cause of this is, of course, the need to generalize the original data to produce an acceptable map product, and is well known among the surveyors and cartographers involved in production of maps, as well as among more sophisticated users, including those many using GIS. The surveyors and cartographers preparing reports and legends to accompany the maps often attempt to compensate for the generalizing effects of the map production process. The information in these legends and reports, more often than not however, are ignored in the process of digitization of the map data prior to incorporation in a GIS. Indeed, no GIS exists that incorporates the ability to manipulate such information.

This paper will attempt to document the awareness of the surveyors and cartographers as to the generalizing process of map production, and will examine approaches that have historically been taken by them to compensate for this effect. Specifically, it will explore the use of the dasymetric map for statistical data, and the application of the same approach to other data types. The example of soil maps is used to document the means by which ancillary map information can be used to improve the reliability of mapped data. An outline is given of how the analysis can be performed within a GIS, and of the research stages needed to verify the usefulness of this approach. It is the contention of this paper that employing knowledge of experts in the domain of the phenomena being mapped (soils, population, etc.) can improve the reliability of the resulting GIS products.

## **Modeling and Mapping With Principal Components**

**Ronnie Pearson**  
**Stennis Space Center**

Traditionally most modeling and mapping has been done by regression.

In mapping the desired criteria on error is smallest error not smallest error squared. Using coefficients from the smallest principal component yields an average error comparable to regression.

In modeling one often needs to change what is referred to as the dependent variable. This requires recomputation of coefficients in regression but does not when using the smallest principal component for computing coefficients. New coefficients can be solved easily in terms of the existing coefficients since an orthogonal system is being used.

## **Confidence Limits in Geographic Analysis -- Suitability Analysis**

**Weldon Lodwick**

## **University of Colorado**

Given mapped data sets whose variations in attribute values are known and bounded, several measures to determine the variability in the resultant attribute values when these maps are used in a geographic suitability analysis. These measures are then used to develop confidence limits in the resultant map. How this could fit into a geographic information system is indicated.

### **Modelling Error in Overlaid Categorical Maps**

**Nick Chrisman**  
**University of Washington**

**Problem:** The GIS community does not need to be told of the use of map overlay as a common part of GIS software. This mechanism is used to integrate diverse data, often using nominal categorical maps. Much of the theory in analytical cartography is built around continuous surfaces and as such not directly applicable to categorical measures. Overlaid categorical maps are usually treated as exact quantities, without any consideration of error. A theory of error must be developed and implemented to temper the results proclaimed from GIS software.

A theory of error for overlaid categorical maps must depend on the nature of the particular map. Goodchild and Dubuc report on a scheme based on error in the categories deriving from a continuous phase space underlying the categorization. While this is useful for some problems, not all categorizations arise from continuous underlying variables in such a simple manner. My research recognizes the distinct character of positional and attribute error -- based on examples from empirical map accuracy experiments. These components are confounded by issues of scale. Hence the full description of error requires at least three components. My presentation will describe this view of map error, a research plan to develop the view of map error and then to apply it to problems in spatial analysis.

### **Modeling Locational Uncertainty in Hierarchical Tessellations**

**Geoffrey Dutton**  
**Prime Computer, Inc.**

Error and uncertainty information sometimes accompanies digital cartographic data; when present, it may characterize an entire dataset, a feature class or a spatial object, but rarely is accuracy data provided for individual spatial coordinates. Coordinate error information can be useful for graphic and analytic operations such as feature generalization, coalescing and overlay. It is suggested that maintaining coordinate data in a hierarchical tessellated framework can assist in documenting the precision of coordinates and in dealing with its limitations. Each point thus stored can identify the precision with which it was encoded, and can retrieve at lower degrees of precision as needed. This is an inherent aspect of quadtree and pyramid data structures, but one which the literature does not seem to address extensively, at least in the GIS arena. If the tessellation is triangular, its elements can still be organized as a quadtree hierarchy, but its geometry assures that facets will be planar and will tessellate a sphere better than rectangular tessellations. This paper explores some of the properties of hierarchical triangular tessellations, focussing on the potential for utilizing hierarchical spatial addresses as substitutes for coordinates, measures of error and similarity, and as general-purpose geocodes for both point locations and extended objects.

### **Spatially Autocorrelated Realizations of Multinomial Processes in Pixels Using Maximum Entropy and Autoregressive Methods**

**Paul B. Slater**  
**U.C. Santa Barbara**

The problem of determining a stochastic process that exhibits spatial autocorrelation effects and yields as its expectation, observed binomial (more generally, multinomial) probabilities in an  $n \times m$  array (A) of pixels, is addressed. Two approaches -- each exactly implementable in small ( $= 4 \times 4$ ) arrays, and approximately in large ( $= 100 \times 100$ ) ones -- are discussed. In one, spatial autocorrelation effects are incorporated by employing as a prior distribution over the  $2^{nm}$  possible configurations of (0, 1) outcomes, the probabilities assigned by the 2-D Ising (ferromagnetic) model to those configurations. (The "temperature" which determines the degree of spatial autocorrelation, can be exogenously assigned or endogenously determined with the computational procedure. Maximum entropy methods (iterative proportional fitting procedures) are then employed to find the distribution over the  $2^{nm}$  configurations which fits the expected pixel probabilities, and is closest in the sense of minimum discrimination information to the prior Ising distribution. The results can be represented as a dual  $n \times m$  array of probabilities. In the second approach, spatial effects are incorporated through an autoregressive model. This involves inverting (power series methods are indicated) an  $nm \times nm$  matrix (I -

$pW$ ), where  $p$  is a parameter analogous to temperature, and  $W$  is a (0,1) adjacency matrix expressing pixel contiguity relations. Multiplication of a vector of standard random normal variates by this inverse, yields a realization of a multivariate normal process. Thresholding each cell of this vector by the cumulative probability level corresponding to the observed binomial probabilities gives on of the  $2^m$  configurations (a realization of the stochastic process).

## **Modeling Reliability on Statistical Surfaces by Polygon Filtering**

**Adrian Herzog**  
**University of Zurich**

The advances in GIS technology allow us to store spatial information in increasing detail. Major human access to data stored in a GIS will remain in the form of maps. This way the accuracy of spatial databases is directly related to the quality of the visualization of their contents. It is therefore important to analyze the communication functions of maps and develop new representation models. Even very simple modeling operations on accurate data can yield results with a considerable amount of uncertainty. This is specially true in the case of fine spatial resolution. It is essential not to communicate such information accurately in a sense of giving information down to the last detail, but to convey the fundamental structure of spatial information. Generalization will be necessary to prevent the user from being disturbed by eye-catching but unimportant information. Nevertheless, cartographic generalization and data accuracy do not have to be conflicting goals: Only if we extend the claim for accuracy by the aspect of reliability we will put a GIS user into a position to take adequate decisions based on the information provided by the GIS.

The problem area as outlined will be exemplified by an analysis of the easily controllable case of statistical surfaces. In particular, we use proportion data because statistical theory is applicable in a straightforward manner and because we can work with information completely associated with uncertainty estimation. Data of zones with variable reliability will be modeled to get not only an accurate but also an appropriate cartographic product. In order not to lose substantial information we refrain from using the known routine classification and regionalization techniques. We try to adapt a quite simple concept: The statistical surfaces defined by polygons will be submitted to a low-pass filtering process. Adaptive filters -- considering both statistical reliability and cartographic principles -- will be applied iteratively to these polygon structures. As compared to traditional map procedures this unconventional map processing method leads not only to more detailed but also to more reliable maps. In addition, we can get rid of disturbing side effects caused by historically predefined zone outlines and at the same time emphasize the continuous nature of statistical surfaces. The paper presents a detailed discussion of the proposed method, an evaluation of the resulting classless and regionless maps, and a comparison of usual methods.

## **Representing Error in a Geographic Data Base**

**Scott Morehouse**  
**Environmental Systems Research Institute**

The correct definition and representation of error in a geographic data base depends on a clear understanding of the nature of geographic information.

Error is only understood in the context of a clear specification of the population being measured and the measurement system being used. In a practical sense, this means that a geographic data base must have a complete specification. This specification can be used to understand what is represented and as a standard for auditing the actual data base.

Several conceptual models for geographic information will be introduced and discussed in this context: regular point sampling (raster), categorical delineation of points, lines, and areas (layer oriented vector), and definition of landscape features (object oriented vector).

I will conclude with some practical examples on the representation of error and data quality information in an actual GIS implementation.

## **What is Special About Spatial Errors?**

**Luc Anselin**  
**U.C. Santa Barbara**

It is widely accepted that a rigorous interpretation of the degree of uncertainty which is embedded in the final product of various GIS manipulations is predicated upon a theory of spatial errors. In particular, when the GIS is used as a tool in spatial analysis, an understanding of error sources and error processes is essential. In this paper, the distinctive aspects of spatial errors are considered more closely. Different types of errors are outlined and the usefulness of spatial statistical and spatial econometric methods for developing a conceptual framework for analyzing spatial errors is assessed.

## **Learning to Live with Errors in Spatial Data**

**Stan Openshaw**  
**University of Newcastle**

Errors in spatial data are a fact of life. They have many different causes and those explicitly due to GIS based manipulations are merely a more recent source of problems. Traditionally researchers and others have simply ignored the presence of errors in both mapping, spatial analysis, and mathematical modeling exercises involving spatial data. A more responsible approach is to seek to develop a better understanding of the different sources of the problem, to investigate their propagation properties, and then to invent methods for explicitly handling the resulting uncertainties. This paper illustrates aspects of the problem with respect to: (1) errors in the spatial analysis of point data; (2) the modifiable areal unit problem involving both aggregation and ecological inference errors; (3) the incorporation of data errors into mathematical models of spatial subsystems; and (4) the handling of uncertainty in both the geo- and the demo- part of geodemographics. It is argued that in many cases appropriate solution procedures already exist. The real problems are: to widen the range of the available empirical experience, to develop operational tools from the research, and to educate users in both the problems and methods of handling them.