

CLOSING REPORT

NCGIA RESEARCH INITIATIVE 1 ACCURACY OF SPATIAL DATABASES

Michael F. Goodchild

National Center for Geographic Information and Analysis
University of California, Santa Barbara, CA 93106-4060

January, 1992

ABSTRACT

This report describes the results of NCGIA Initiative 1 on the Accuracy of Spatial Databases. It begins with a discussion of objectives and the process of developing a research agenda. Each of the seven major areas of the research agenda is discussed, and research activities of the center during the active period of the initiative (1988-90) and since its completion are described. The report ends with an assessment of the initiative against five criteria.

BACKGROUND

GIS processes operate with high precision, but frequently must process data of strictly limited accuracy. On the other hand conventional, manual methods of spatial data analysis operate with a precision which is compatible with the inherent accuracy of the data. In essence, GIS features such as scale change, and the ability to overlay data from different sources and of different quality are often regarded as the technology's most significant advantages, but in practice may represent significant problems. The importance of accuracy and related issues of data error and uncertainty have been recognized for a long time in the GIS literature, but systematic efforts to come to grips with the problem are much more recent. Burrough (1986) devoted an entire chapter to "Data Quality, Errors and Natural Variation" in his groundbreaking text on GIS, arguing that the high costs of building spatial databases and associated software could only be justified if the data in them were of sufficient quality. During the writing of the Santa Barbara/Buffalo/Maine consortium proposal for NCGIA in 1987, data errors and uncertainties were identified as one of the most important impediments to the successful implementation of GIS (NCGIA, 1989). We lacked not only the means to describe the uncertainty known to be present in spatial data of all kinds, but methods for propagating uncertainties through GIS operations and expressing their effects as confidence limits on products. Few standards of data quality existed beyond those used for traditional mapping, and few public agencies had established policies regarding quality assurance and control. The topic of spatial database accuracy was given highest priority as the first Center initiative.

The specialist meeting for this first research initiative of the NCGIA was held at La Casa de Maria in Montecito, CA from December 13 through 16, 1988, to lay out the initiative's specific research agenda. 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. Efforts were made to achieve a representation from academics already known to be working on issues of accuracy, with an established publication record; representatives of agencies known to be concerned about data accuracy, particularly agencies providing spatial data in digital form; and development engineers employed by GIS software vendors, who might provide useful channels to actual implementation of results.

Because this was the first NCGIA specialist meeting, the program had to be designed with virtually no experience. Should the meeting be structured, allowing each participant some minimal amount of time to present ideas, or should it consist of largely unstructured discussion? The well-defined nature of the topic of Initiative 1 suggested a more structured approach, although with hindsight, the less structured option seems to have worked best as a general model for specialist meetings. Each participant was asked to prepare a paper for presentation, and the program was organized as a progression from general discussions of issues to basic description and modeling, and ending with policy. The meeting was designed for three full days, including evening sessions on most days, to accommodate the almost 30 speakers.

Revised versions of the position papers of the meeting were collected and edited in the months following the meeting, and delivered to Taylor and Francis for publication as a book in the late spring of 1989 (Goodchild and Gopal, 1989). The book appeared in early October, and reviews have been published in many journals. The following sections discuss each of the research issues identified at the meeting and the research progress achieved during the active period of the initiative. The initiative was formally completed in November 1990 by a series of presentations in three special sessions of GIS/LIS 90, but research is still continuing, and

much remains to be published. The existence of the initiative, and particularly the specialist meeting, has stimulated work on these problems in many other institutions worldwide, and Center research continues under Initiative 7, with an emphasis on the visualization issues of data quality, and under Initiative 12 with specific reference to the accuracy of remote sensing inputs to GIS.

This report presents a review of the current state of the art in spatial database accuracy, and ends with an assessment of the contribution of NCGIA Initiative 1. The first sections look at the various items of the research agenda identified at the specialist meeting, and recent work in each area inside the center. Continuing work is also discussed, although the initiative is now formally completed. This is followed by a summary evaluation, based on the five standard criteria devised by the NCGIA Board of Directors for assessment of initiatives.

Since the topics of NCGIA research initiatives tend to be open-ended, the completion of an initiative should not be taken as implying that the problem has been solved in any sense, or that no more work is necessary, or that the topic has been exhausted. Research almost always leads to more questions, and to the need for more research. Completion implies one or more of the following:

- the research agenda established at the specialist meeting needs revision based on subsequent findings, so it is time to hold a new specialist meeting and reconsider research objectives; initiatives will likely reach this state after approximately two years;
- research has led in unexpected directions, and a new specialist meeting should be held to define a more appropriate research agenda;
- in view of the limited resources available to the Center, it appears that the benefits of further research on the topic no longer outweigh the costs;
- the key personnel involved in research on the topic within the Center are no longer available;
- research has led to a body of significant results, and it now seems more appropriate to put effort into their practical implementation than to continue with basic research.

In the case of Initiative 1, there are elements of the first, second and fifth reasons in the decision to complete the initiative.

PROGRESS ON THE RESEARCH AGENDA

The research agenda developed at Casa de Maria was published as a center Technical Paper (Goodchild, 1989), and is also summarized in the preface to the proceedings volume (Goodchild and Gopal, 1989). Early in the research period, Veregin developed a working bibliography of the existing literature on spatial database accuracy, and this was published as an annotated bibliography and associated review (Veregin, 1989a,b) in the center Technical Papers series.

The research agenda identified seven major themes:

- *data structures and models* - research on the representation of uncertain data through appropriate data models and data structures within a digital environment;
- *models of error and distortion* - research on statistical models of error, the estimation of their parameters, and their use in practical applications;
- *error propagation, product uncertainty and sensitivity* - techniques for predicting and tracking the propagation of error from uncertain databases through to uncertain information products, including issues of non-linearity and the definition of confidence limits for GIS products;
- *accuracy and risk* - the relationship between uncertainty and policy, and the development of quality assurance and control standards and procedures in agencies engaged in handling spatial data;
- *experimentation and measurement* - techniques for measuring errors in practical situations, and for minimizing them;
- *interpolation and surface modeling* - the uncertainties introduced in interpolating values between data points, such as those inherent in the various data models for topographic surfaces;
- *aggregation, disaggregation and modifiable areal units* - the uncertainties due to the reporting zones used for much social and economic data, and introduced by attempting to transfer data from one set of reporting zones to another, incompatible set.

Data Structures and Models

This research theme has concerned the development of 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. Two papers at the Specialist Meeting directly addressed this issue by recommending a finite resolution approach in which the precision of computation is adjusted to the accuracy of the data. Geoffrey Dutton in particular discussed the importance of this approach for global data, where accuracy is highly variable, and where no suitable finite element schemes existed.

Yang Shiren and Michael Goodchild began research on global tessellation schemes in January 1989 at Santa Barbara. Working with a modified version of Dutton's proposal, they succeeded in developing a finite element scheme for the globe based on projection onto an octahedron, followed by recursive subdivision of triangular facets. The scheme was included in EPA's review of possible sampling schemes for the EMAP project, as it provides a method of achieving approximately uniform sampling over a spherical surface, but rejected in favor of a scheme based on projection onto the faces of a truncated icosahedron, since the entire

conterminous US can be contained in one appropriately positioned hexagonal face of this solid, leading to scale factors between 1.00 and 1.01, whereas the scale factors in the simpler octahedral scheme range from 1.00 to 1.57. Research to date at Santa Barbara has determined the mathematics of the necessary projections, and algorithms for such basic functions as conversion to and from latitude/longitude, chain coding of lines, dilation and region filling. The algorithms have been moved to the IBM RS/6000 workstation, implemented in 3D display technology, and used as the foundation of a prototype global GIS. Work on this project is continuing. The basic details of the scheme were published as two Center reports (Goodchild and Yang, 1989; Goodchild, Yang and Dutton, 1991) and a revised version of the first has been published by *Computer Vision, Graphics and Image Processing*; the second is under review.

GIS provides a number of ways of converting real geographical variation into finite, discrete digital representations. This issue of data modeling is critical from an accuracy perspective, as any model of uncertainty must be specific to data model, and thus can inform the choice. Since December 1989 the Santa Barbara site has held regular seminars on the conceptual design of the "Next Generation GIS", in part through its joint study agreement with IBM. A paper by Goodchild, developing a perspective on GIS data modeling consistent with the findings of Initiative 1, has been presented at a seminar on Data Modeling sponsored by the Canadian Inter Agency Committee on Geomatics in Ottawa in March 1990, and at the GIS Design Models Conference sponsored by the Midlands Regional Research Laboratory in Leicester, UK, also in March 1990, and will appear in *Computers and Geosciences* (Goodchild, in press). A paper in the Initiative 1 sessions at GIS/LIS '90 (Goodchild, 1990) enlarged on aspects of this theme.

Models of error and distortion

In this area we have concentrated on developing models which can successfully describe, characterize and parametrize error, both for spatial fields and for complex spatial objects. Early ideas on this theme were presented at the Third International Symposium on Spatial Data Handling in Sydney in 1988 (Goodchild and Wang, 1988) and at AutoCarto 9 in 1989 (Goodchild and Wang, 1989), and were developed further in Chapter 10 of Goodchild and Gopal (1989). Several methods of simulation have been developed for modeling error in different types of spatial data. In essence, each assumes that uncertainty is described by a raster in which each pixel carries not a single deterministic class, but a vector of probabilities of belonging to each of a set of classes. We have explored methods for generating a population of distorted "maps" from such input, under two constraints. First, the proportion of occurrences of each class in a pixel across realizations must be as input; second, the outcome within one realization must be spatially dependent (it is easy to ensure the first condition, but more difficult to deal with the second; the consequences of disregarding the second are clearly evident in the simulations of Fisher, 1991b). A method based on a spatially autoregressive process has been programmed by Sun Guoqing at Santa Barbara, and implemented using a grant of computing time from IBM on the 3090 facility at Palo Alto Research Center. More recently a much more efficient method of obtaining realizations of the process has been used, based on an iterative technique due to Heuvelink (in press). A further alternative method based on swapping values between realizations was implemented by Fang Chiuwen, but rejected. We have worked out the conceptual design of a GIS capable of encoding the level of uncertainty in each of its layers, and tracking the propagation of uncertainty through GIS operations into confidence limits on GIS products. The results of this work have been described in several presentations, and will appear in the *International Journal of Geographical Information Systems* (Goodchild, Sun and Yang, in press). An implementation of the approach is being developed using the GRASS GIS and will be made available over the GRASS user network. Work is continuing at Santa Barbara under Initiative 7 on the use of this model in visualizing uncertainty, and particularly in accessing the results of fuzzy classification of remotely sensed scenes.

These methods are defined entirely within the field domain. We have concluded based on this work that in general it is feasible to build models of uncertainty in the field domain, but much more difficult in the object domain. Models of uncertainty for objects are more readily constructed by distorting underlying fields from which the objects have been derived, and we have devised methods for doing this. This serves to point up an interesting and fundamental distinction between Initiatives 1 and 2. Whereas the field domain is more significant in I1, I2 has dealt almost entirely with cognitive views of a space populated by objects. Indeed, our vocabularies for describing continuous variation in fields are quite limited when compared to those for describing relationships between objects. But whereas the object view is clearly implied by human perception and cognition, the field view is preferred for describing uncertainty and for modeling many physical processes.

Moreover, the object view is clearly preferred when dealing with GIS applications in facilities management and land records, where objects have real identity, rather than being artifacts of the process of discretization. This dichotomy emerged very strongly in a meeting on Accuracy of Spatial Databases held in Melbourne, Australia in June 1991, and stimulated by Initiative 1, and it also permeates other debates on spatial data accuracy. Work at Maine on survey adjustment (Buyong and Kuhn, 1990a,b; Buyong, 1989; Buyong and Frank, 1989; Buyong, Kuhn and Frank, in press) falls into this area. On the other hand, David Mark and Ferenc Csillag have investigated the nature of boundaries on area-class maps, and have argued in a paper published in *Cartographica* (Mark and Csillag, 1989) that the process of generalization of such boundaries requires a continuous view of space, using surfaces representing probabilities of class membership.

At Maine, David Pullar is completing a doctoral thesis on the computation of the geometrical intersection of two or more polygon-nets having imprecise numerical data. Imprecision in the numerical data reflects uncertainty in the geometry of objects, and it is measured by a variance parameter. A variance is incorporated into the geometrical intersection procedure to prevent slivers, gaps, and other artifacts. The results from completed work are described in three papers (Pullar, 1990, 1991; Pullar and Beard, 1990). First, a comparative study was carried out on the complexity and performance of algorithms for reporting geometrical intersections. A fast pragmatic algorithm is proposed under the conditions that line segments are randomly distributed over the subject space. Second, an algorithm was developed to decide if two line

segments with imprecise numeral data, cross, nearly cross, or are separate. It is referred to as a "fuzzy-intersection" algorithm. Third, the previous work is extended to treat intersections among arrangements of polylines. An algorithm for computing the geometrical intersection of a number of polylines, with associated variance parameters, is proposed and is called "fuzzy-chain-intersection". The algorithm uses clustering techniques to: i) resolve positional ambiguities, and ii) produce a valid planar topology within specified tolerance bounds. Further work has described the way geometrical variance propagates through the process of fuzzy-chain-intersection.

Error Propagation, Product Uncertainty and Sensitivity

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. Many physical processes are known to transfer energy between scales, thus allowing the effects of errors at one scale to propagate to other scales. The "butterfly effect" argues that the minute atmospheric disturbance generated by the flapping of a butterfly's wings can potentially propagate into a major cyclonic disturbance. Initiative research in this area has concentrated on understanding how error propagates through each GIS process, particularly overlay. Lodwick described existing work on this topic at the specialist meeting and has continued this line of research, in part with funding provided by the Center (Lodwick, 1989; Lodwick, Monson and Svoboda, 1990). Funding was obtained from USGS to support work on this topic at Santa Barbara in the Fall of 1989 by Giuseppe Arbia (University of Rome) and Robert Haining (University of Sheffield). Arbia and Haining have developed a general linear model which includes numerous forms of spatial data error as special cases, and provides a rich framework for analysis and simulation. The first set of results has been described in a paper currently under revision for a statistics journal (Arbia and Haining, in press).

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. At the Buffalo site, Rajan Batta has addressed the relationship between data error and the results of spatial decision-making, particularly in the context of location. One way to represent uncertainty is through the use of a stochastic location model, and objective functions which explicitly recognize uncertainty. Another is to look at the database as uncertain, and to examine the propagation of database uncertainty into the results of optimization. The references to Batta's work below include both of these approaches (Carson and Batta, 1990; Berman *et al.*, 1990). Also at Buffalo, Rogerson has researched the link between region size and shape, and the migration process, particularly in the estimation of migration distances (Rogerson, 1990a,b).

Another facet of the propagation of errors in spatial data is their effect on methods of spatial data analysis. At Santa Barbara Luc Anselin and Serge Rey continued to investigate this issue in terms of the performance of various statistical test for the presence of spatial dependence in regression analysis. In a large number of Monte Carlo simulations, they compared tests for spatial error autocorrelation and for substantive spatial dependence and found that Lagrange Multiplier tests were superior. The results of their research were presented at the North American RSA and Annual AAG Meetings and appeared in *Geographical Analysis* and as a NCGIA Technical Paper (Anselin and Rey, 1991; Anselin, 1989, 1990a,b). The impact of spatial errors and other issues related to spatial data analysis in modeling international relations is further investigated in collaborative research (separately funded by NSF and the University of California Institute on Global Conflict and Cooperation) by Luc Anselin and John O'Loughlin (University of Colorado, Boulder). Initial results will appear as two chapters in a book on *The New Geopolitics*, edited by political scientist Michael D. Ward.

The error model for categorical data developed by Goodchild, Sun and Yang (in press) has been applied to the propagation of errors through overlay, and the effects of data uncertainty on estimates of area from a GIS. Suppose, for example, that a patch on a land cover map is known to be a mixture of 80% class A and 20% class B. A presumably unbiased estimate of the area of the patch that is truly A would be 80% of the patch's total area, and the same would be true similarly of any part of the patch. But the uncertainty in this estimate depends on the sizes of the inclusions of B, and simulations using the model of Goodchild, Sun and Yang (1991) have shown that the standard error of estimate rises rapidly and in non-linear fashion with inclusion size, even though the percentage of B is held constant. Similar simulations were also run on estimates of area after topological overlay.

In the Fall of 1989 David Lanter proposed that his Lineage Information Program offered the possibility of adaptation for storing and propagating data accuracy indices to parallel the propagation of data in spatial analytic GIS applications. Lanter and Howard Veregin automated a series of error propagation functions to model the proportion correctly classified (PCC) index. To date they have automated functions that propagate PCC through the UNION, INTERSECTION, and RECLASSIFY GIS operations. A function to model the propagation of PCC through BUFFER operations has been designed and will be tested using a Cray Supercomputer. Preliminary findings of this research were presented at GIS/LIS '90 (Lanter and Veregin, 1990).

Subsequent work resulted in the formulation of a research paradigm of error propagation research for layer based GIS. The research paradigm focuses attention on the role of assumptions about source error measures for characterizing whole cartographic themes, the spatial distribution of this error in derived product layers, and the need for error propagation models each matched to an individual error index and GIS function. This is presented in "A Research Paradigm for Error Propagation in Layer Based GIS", in press with *Photogrammetric Engineering and Remote Sensing* (Lanter and Veregin, in press).

Additional research with Howard Veregin focuses on the application of inverse error propagation functions in lineage analysis. Preliminary results indicate that it is possible to determine the relative sensitivity sources have with respect to the accuracy

of a derived product layer. As a result a desired product accuracy can be 'locked in' to determine the minimum accuracy requirements of a specific data source. This has the potential of serving as a tool for prioritizing the re-survey of existing digital source base maps. Research findings were presented in the paper "A Lineage Information Program for Exploring Error Propagation in GIS Applications" at the ICA meeting in Bournemouth, England in September 1991 and appeared in the proceedings (Lanter and Veregin, 1991).

Accuracy and Risk

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 a link between this initiative on spatial database accuracy and Initiative 4 on the use and value of spatial information. We hope that the error models developed as part of this research effort will ultimately lead to a better-informed approach to risk. 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.

Partly through Initiative 1, the center has become involved in various efforts to establish quality standards for spatial data, and procedures for quality assurance and control. Work by Amrhein at the Buffalo site has focused on the specification of accuracy requirements for Statistics Canada's geographical products (Amrhein and Schut, 1990). Goodchild was a member of the ICA's Science Advisory Board during its efforts to develop quality standards for international spatial databases, and this work is continuing under the ICA Commission on Spatial Data Quality. In the US, the proposed federal Spatial Data Transfer Standard includes a detailed section on data quality, and this is having a significant influence on international debate on spatial data quality. The Environmental Protection Agency's GIS research and development group in Las Vegas is in the process of developing a quality assurance/quality control policy for GIS applications within the agency, in consultation with the center. Finally, the center has provided advice to several other federal and state agencies seeking to develop data quality policies, including Oak Ridge National Laboratory (Department of Energy).

Two reports sponsored by the Center and aimed at increasing awareness of accuracy issues have been written by Howard Veregin and published in the Center's technical report series: a bibliography of accuracy literature and an associated taxonomy of errors (Veregin, 1989a,b).

Once risk is measurable, it is necessary to develop strategies which deal with it appropriately, and incorporate it in decision-making. Work at the Buffalo site by Batta has centered on location models which explicitly recognize risk, and distribute it equitably over a dispersed population. These have been applied to the specific cases of liquified-gas transportation, and the location of mobile ambulance units. The effects of risk on environmental policies developed using GIS was the subject of a session at the First International Conference/Workshop on Integrating GIS and Environmental Modeling organized and sponsored by NCGIA and several federal agencies in Boulder, CO in September 1991.

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. The Santa Barbara site of the Center is involved in a continuing project with the California Department of Forestry to measure accuracies in GIS data used to estimate the State's forest resources. Estimates are currently made from a variety of sources including vegetation maps, air photos and ground truth. A model is being built to optimize the use of these sources, and a demonstration data set has been built using ARC/Info as part of the project. The final report for the first phase of this project was delivered in March 1991 and has been published as a center Technical Paper (Goodchild, Davis and Stoms, 1991), and NCGIA is currently organizing an Accuracy Assessment Task Force for forest land mapping on behalf of CDF.

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. Center research in this area has focussed on developing efficient methods of constructing TINs, and comparisons between the methods of DEM generation used in public agencies. David Theobald completed his Master's thesis on modeling hydrology using TINs in March 1990 (Theobald, 1990) and made a presentation on this research at GIS/LIS 90 (Theobald and Goodchild, 1990), and a paper based on this work is under revision for a journal. Mark Kumler, a PhD student at Santa Barbara, worked at USGS Menlo Park in the summers of 1989 and 1990 on the potential of TIN models for terrain modeling, and is currently completing this work as a PhD dissertation. Preliminary results have already appeared (Kumler, 1990; Kumler and Goodchild, 1991).

Aggregation, Disaggregation and Modifiable Areal Units

Several of the papers presented at the specialist 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. Research in the late 1970s by Openshaw and others (Openshaw and Taylor, 1979) had demonstrated that reporting zones can have a substantial, biasing influence on the results of analysis, and that these effects (the "modifiable areal unit problem") pervade the spatial analysis of social and economic data. 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. As part of the initiative effort we have developed a general framework for transferring socioeconomic data from one set of spatial objects to another (reported by Deichmann, Goodchild and Anselin at the North American RSA meetings, 1989, and the Conference on Advanced Computing in the Social Sciences, Williamsburg, VA, 1990), and currently under review by a journal. The problem is seen as one of estimating one or more underlying density surfaces under various assumptions. An initial application of the approach and an evaluation of its effect on the interpretation of the results of integrated multiregional models was carried out for a model of the State of California (reported by Anselin, Rey and Deichmann at the European RSA meetings, 1989 and subsequently published as a book chapter). The primary innovation of the research is in showing that greatly improved estimates of the attributes of incompatible zones can be obtained by exploiting the imprecise and largely intuitive knowledge of density surfaces that is often available to the analyst.

Work by Amrhein at the Buffalo site of NCGIA has been focussed on the analysis of aggregation effects in migration modeling. Although social processes are best defined at the individual level, most data are available only at aggregate levels in order to protect privacy. Amrhein examines the extent to which scale is important in the calibration of migration models, and finds a surprising degree of independence. Also at Buffalo, Rogerson has examined the effects of aggregation in the time domain on migration modeling (Rogerson, 1990b), arguing that migration data collected for periods of differing lengths will yield inconsistent population forecasts. Goodchild and Noronha have developed a method for creating functional regions from interaction tables, which may allow a more rational approach to aggregation in such cases (Goodchild and Noronha, in press).

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, Tobler argued at the specialist meeting that the best models of spatial phenomena would be ones which were independent of scale, and Fotheringham's 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 is concerned with the representation of objects at multiple scales. Tobler presented a paper on design criteria for a resel-based processing system at Zurich in July, 1990, and a revised version will appear in the *International Journal of Geographical Information Systems* (Tobler, in press). Fractals provide one theoretical framework for dealing with scale effects, and work at the Buffalo site by Stewart Fotheringham, in collaboration with Michael Batty and Paul Longley of the University of Wales, has explored fractal models of uncertainty in socioeconomic data under Initiative 3. Fractals have also been the focus of work at Santa Barbara, and a paper evaluating various methods of fractal measurement in the context of landform processes will be published in *Earth Surface Processes and Landforms* (Klinkenberg and Goodchild, in press).

ASSESSMENT

The previous section reviewed activities under Initiative 1 in each of the seven major areas of the research agenda. Some parts of the agenda received more attention than others, either because they were perceived to have higher priority, or because there was more interest in them within the center. Although the initiative lasted only from December 1988 to November 1990, and is now formally completed, many of the key personnel were involved in accuracy research before the start of the initiative and have continued to be active in research since its completion. In addition substantial research in this area has been under way outside the center, in many cases with strong links to the center.

This section presents a scientific assessment of the initiative, organized according to the five criteria for initiative assessments established by the NCGIA Board of Directors.

1. RESEARCH ACCOMPLISHED. What do we know now that we did not know before about the questions addressed by the initiative?

The results accumulated during the initiative represent substantial progress in each of the following areas:

- how to build data structures on the globe to handle data of varying quality and spatial resolution, for research into global human and environmental systems;

- how to model error in fields of categorical variables - area class maps; specifically, how to deal with boundaries that are actually transition zones, and polygons that are not homogeneous;
- how to propagate error in area class maps through GIS operations and estimate confidence bands on related products;
- how to collect data during land cover mapping exercises in ways that are understandable to the user, minimally disruptive, and yet useful in modeling error in the resulting database;
- how to build a system that tracks the uncertainty in each data layer, and provides the user with a visual interface to this information;
- the relative efficiencies of TINs, contours and DEMs as alternative ways of capturing digital terrain data, and the usefulness of the TIN model as a basis for determining hydrology;
- how to build a system that allows the user to input knowledge of the variation of population density, in order to make more accurate estimates of the populations of incompatible reporting zones;
- the severity of the modifiable areal unit effect in multivariate analyses of spatial data.

This list is necessarily incomplete, and addresses only the more obvious of the initiative's outcomes. More detail is provided in the next section.

2. RESEARCH AGENDA DEVELOPMENT. How has the research agenda been affected by this initiative?

The NCGIA research plan, written in late 1987, had the following to say about data accuracy and its treatment in GIS:

"Research in this area should be directed first at developing appropriate statistical models of complex spatial objects and the errors present in them. Methods are needed for estimating the parameters of these models and for designing numerical means of measuring accuracy and proportioning error among the contributing sources. Finally, each of the common GIS functions, such as area measurement and polygon overlay, should be analyzed to develop appropriate confidence limits based on the underlying models" (NCGIA, 1989 p.120).

The specialist meeting in 1988 refined and extended this statement, and identified seven major areas of research.

In the first area, data structures and models, the principal contribution of the initiative has been the development of a finite element scheme for the globe with potential applications in the visualization, storage and analysis of global databases. As a hierarchical scheme it allows data to be expressed at variable levels of accuracy, which is essential in handling data for research into the human dimensions of global change, since all international social and economic data will vary enormously in quality. Many other aspects of data quality, such as the parameters of the error models described in this report, can be handled by storing attributes of appropriate spatial objects or classes of objects.

The center's most significant contribution has been in the second area, models of error and distortion. The theory of random fields already provided a means of describing and modeling error for fields of continuous variables, and work at Utrecht by Burrough and others (*e.g.* Heuvelink, Burrough and Stein, 1989) exploits Taylor series expansions and Monte Carlo simulation to propagate errors through GIS processes. Fisher (1991a) has used simulation to show how estimates of a viewshed derived from a GIS are sensitive to errors in the underlying DEM. However these methods deal only with fields of continuous variables, and are not useful as models of error in fields of categorical variables - so-called area class maps. The model described by Goodchild, Sun and Yang (in press) is the first to provide a solution for categorical data, which is found frequently in GIS in the form of soil maps, land use/land cover maps, maps of forest inventory, zoning maps *etc.* In fact the polygon data model inherent in area class maps was the only model supported by early versions of ARC/INFO, and its raster equivalent is the dominant data model in most raster GIS. This new model can be readily incorporated in such raster GIS as GRASS, to provide confidence limits on the products of a wide variety of GIS processes.

In summary, we now have adequate data models for both classes of field data models - continuous and categorical variables - and the basis on which to build a comprehensive error-handling GIS. For the object data models, points create no problem. Problems still exist, however, for line and area data, except where these can be regarded as derivatives of fields. For example, uncertainty in the objects shown on contour maps can be modeled by assuming the contours to be derived from a field of a continuous variable, elevation.

In classical error analysis, which deals with scalar measurements, the Gaussian model provides a comprehensive theoretical framework for error modeling. It describes what is expected from a large number of random disturbances with additive effects. In some circumstances, information may be available on the actual processes causing error, and models of these may differ substantially from the Gaussian norm. Similarly in spatial data, the two error models described above represent expectations in the absence of any specific knowledge of error-generating processes. If sufficient information is available, it may be possible to model processes such as digitizing error explicitly. Although there is already a substantial literature on models of digitizing error, the magnitude of this error source is often minimal compared to the errors introduced by other, less well understood processes, such as the replacement of transition zones by sharp boundaries, or the assumption of homogeneity within polygons on area class maps. Now that we have the

equivalent of the Gaussian distribution for the most important types of spatial data, future research should concentrate on models of specific sources of error, such as raster/vector conversion, and on improved methods of calibration for the basic models.

In the third area, error propagation, we are substantially further ahead as a result of the initiative. Lanter's work has provided a framework for managing information on propagation, and the error model developed under the initiative has been shown to be readily amenable to supporting error propagation analysis. The research agenda should now shift to an emphasis on implementation, and particularly the development of visual methods and interfaces that avoid the problems caused by the conceptual sophistication of spatial statistics. We need ways of describing the magnitudes of errors that are meaningful to users without advanced knowledge, and yet are readily translatable into the appropriate statistical parameters. Goodchild, Sun and Yang (in press) argue that the answer must lie in visual pattern-matching, and work on this is continuing. But unfortunately the key parameters of spatial dependence and autocorrelation have little intuitive meaning. We also need well developed examples of the effects of spatial data errors on the performance of large models, and Walker is investigating error effects in a mesoscale atmospheric model with this in mind.

In the fourth and fifth areas, which deal with risk and the measurement of accuracy, the most significant contribution of the initiative has been to draw attention to the effects of error in a variety of media and forums. Although it is easy to convince people that spatial data contains errors and uncertainties, and that these propagate in GIS, it is much more difficult to convince people that they should take significant action - too often, errors are swept under the carpet and ignored. The 1991 Statistics Canada Symposium focused on Spatial Issues in Statistics, and featured several presentations by center personnel on aspects of data quality and spatial analysis. Goodchild gave the keynote address; Amrhein discussed the use of small area data in demographic analysis; Fotheringham and Wong presented a paper on the modifiable areal unit problem in multivariate statistical analysis; and Deichmann reviewed the results of Initiative 1 in a paper coauthored with Goodchild and Anselin. In June, 1991, Goodchild made the keynote presentation at a conference on Accuracy of Spatial Databases at the University of Melbourne. In February, 1992, he presented a review of Initiative 1 at a symposium on data uncertainty organized by Energy, Mines and Resources Canada in Ottawa. Workshops on spatial data accuracy have been presented at the AAG Annual Meetings in 1991 in Miami, and the ESRI Users Conference in Palm Springs, and a further workshop is planned at the Fifth International Symposium on Spatial Data Handling in 1992. Numerous presentations on Initiative 1 have been made to universities and conferences. The book *Accuracy of Spatial Databases* has also drawn attention to the issue, and has been widely reviewed. Initiative 7, which deals with visualizing the quality of spatial data, will provide opportunities to continue this outreach process.

In the sixth area, interpolation and surface modeling, the major contribution has been the work of Kumler and Theobald on terrain representation. Kumler's work concludes that for a wide range of different types of terrain, the DEM is not only the most efficient sampling scheme in the absence of prior terrain knowledge, but is also the most accurate for a given volume of digital storage. This reverses a longstanding assumption in the GIS field, and weakens the argument for the TIN model significantly. Theobald's work identifies substantial problems in using TINs for hydrologic modeling.

In the seventh area, modifiable areal units, significant progress has been made in developing general methods for interpolation of data between incompatible reporting zones. Fotheringham's work on the impact of modifiable areal units on analysis is continuing under Initiatives 3, 6 and 14. However it will likely be many years before the basic message of this work has much impact on the many disciplines that analyze data for reporting zones, or that traditional methods of data collection, analysis and display are replaced by more appropriate ones.

In summary, the research agenda has changed significantly as a result of this initiative, with a greater emphasis on visual interfaces, applications and implementation now that the most important basic problems have been solved.

3. CONTRIBUTION TO GIS EDUCATION. How has the education of GIS scientists been enhanced by the initiative?

The initiative is at least in part responsible for the fact that data quality is now a significant subfield within GIS, with sessions at conferences, chapters in texts and papers in journals. Four units in the NCGIA Core Curriculum are devoted to it, and reference has already been made to conference workshops. Two conferences, one in Australia and one in Canada, have addressed data quality in the past year, and other activities described in the fifth section below are also relevant to an educational mission. Under Initiative 7, the center is developing a hypercard stack illustrating issues of data quality, in the expectation that this will help to draw attention to the quality problem using an attractive visual medium. Without a constant concern for data quality in a field in which all data is inherently uncertain to some degree, we run the risk of investing millions in building sophisticated and elegant systems to analyze rubbish. After all, a cynic once described a land cover map as "lines that do not exist surrounding places that have nothing in common".

4. SCIENCE POLICY. What recommendations would NCGIA make in this area?

Lists of recommendations could be developed from many areas of Initiative 1 research, but perhaps the most important of these concern the development of data quality standards and policies. These ultimately reflect the expression of the results of this initiative and similar research in the form of policies and procedures.

Data quality can mean many different things, but the approach taken in this initiative has been that data quality is determined by the relationship between the contents of a database and the reality which it purports to represent. Several assumptions are implicit in that definition that deserve explanation. First, the definition assumes that reality can be defined, a position that receives more sympathy in the physical and environmental sciences than in contemporary social science. In practice, reality in areas like soil

mapping is in part subjective, and it is accepted that different observers may report different views of the same reality. To deal with this, we assume that reality can be defined, but that the definition may contain uncertainty, which will combine with other sources of uncertainty. Second, the definition implies that quality is not measured by the degree to which the database accurately captures the contents of input documents and maps, since those documents will themselves contain errors. Map quality standards are not necessarily relevant to the GIS data quality question, since they are concerned with the accuracy of map contents, not with the relationship between database and reality. We see the data quality standards adopted by many mapping agencies as falling into this category.

For example, the data quality standard for contour maps is highly relevant to the US Geological Survey's mission. But a GIS user concerned with the accuracy of elevation estimates at arbitrarily chosen points, or the accuracy of measures derived from elevations, such as slopes, aspects or viewsheds, is not able to use a contour map accuracy standard to estimate uncertainties. The accuracy of the positions of map features, such as contours or soil map boundaries, is not helpful in determining the accuracy of GIS products. Thus we distinguish between map accuracy standards and GIS database standards.

- There is a pressing need to develop accuracy standards relevant to users of GIS databases; map accuracy standards address a separate (but important) set of issues.

- The data quality section of the proposed federal Spatial Data Transfer Standard is primarily concerned with the accuracy of features derived from maps, and falls short of meeting the needs of GIS users and error-handling GIS.

Although GIS represents a major shift in perspective on spatial data, the methods of spatial data collection remain largely the traditional ones. Land cover mapping, for example, still proceeds largely by dividing areas into more or less homogeneous patches of single classes, and with sharp boundary lines between them. While this approach is compatible with mapping techniques, and is easily digitized, it fails to capture the mapmaker's full knowledge of spatial variation, or to take advantage of the power of modern database technology. A forester or soil scientist often knows where boundaries are broad transition zones rather than sharp breaks; where inclusions of different class occur within patches, but below the minimum mapping unit size; and where patches are characterized by mixtures rather than single classes. It is straightforward to capture this knowledge in the form of additional attributes of boundary arcs and polygons, in order to provide better information to an error-handling GIS.

- We need to promote simple enhancements to mapmaking and data collection in order to provide more information on known uncertainty, and to make better use of the knowledge available to the mapmaker but traditionally rejected.

- We need to develop better methods of mapping that can communicate knowledge of uncertainty to the map reader. This is the topic of Initiative 7, which began in June 1991.

Classification methods used in remote sensing - whether supervised or unsupervised - commonly result in a single class being assigned to each pixel in the scene. This is convenient, but like many mapping techniques it creates a false picture of certainty to the user and the analyst. Methods of fuzzy classification, reflecting the expectation that many pixels will be mixed, and others cannot be classified with certainty, have been available for some time but are rarely used. It is also simple to modify a likelihood classifier to produce several class probabilities per pixel rather than one. Fuzzy classification is more reflective of the knowledge of the scientist, and provides more accurate estimates of products such as area when analyzed in a GIS. The error model developed in this initiative is designed to accept fuzzy classifications and to estimate the uncertainties associated with derived products.

- Fuzzy classification of remotely sensed scenes would provide a more appropriate input to GIS, and would be better suited to error analysis than the conventional single-class classification.

- We need to develop methods for visualizing fuzzy classifications (this is the subject of active current research under Initiative 7).

The modifiable areal unit problem is a serious source of uncertainty in the many disciplines that analyze and model social and economic data, including geography, sociology, economics, regional science and demography. The practice of reporting on the basis of discrete, arbitrarily defined objects also affects the development of public policy in arbitrary and poorly understood ways. Its effects on the political system, in the form of gerrymandering, are well known. The accepted way to minimize its effects is to work always at the lowest possible level of aggregation, and GIS provides an indispensable toolkit for working with large volumes of disaggregate data. But this fundamental change in technology has had relatively little effect thus far on the agencies that collect data, and distribute it for analysis.

- To minimize the impact of the modifiable areal unit problem, we need to develop better ways of integrating data sources with GIS, and of managing, analyzing and displaying them at the lowest possible level of aggregation.

5. THE RESEARCH INITIATIVE PROCESS. What were the strengths and weaknesses of the research initiative process in facilitating the research in the initiative?

Initiative 1 was the first to be organized by the center. Its specialist meeting broke new ground, but proved that the concept could achieve its stated goals of formulating a research agenda. In subsequent specialist meetings the center has placed more emphasis on open discussion rather than formal presentations, but while this has led to better debate, it misses the opportunity for a proceedings volume of formal papers; in the eyes of many observers of NCGIA, *Accuracy of Spatial Databases* remains the most tangible product of Initiative 1 to date. Subsequent meetings have also downplayed the role of the specialist meeting in developing a research agenda for the center, and have emphasized the role that the meeting plays in stimulating interest in the topic and promoting

research in the entire community. Participants at more recent meetings seem to have had a much clearer idea of the meeting's objectives, and to have been much happier with them.

The center also learned from I1 and other early initiatives the importance of creating a general level of anticipation and awareness in the GIS community before the initiative began. More recent initiatives have held small preparatory meetings, announced plans much earlier, and in the case of Initiative 9 (Institutions Sharing Spatial Information) have advertized widely, soliciting abstracts and reviewing them before issuing invitations.

During the active period of the initiative a newsletter was circulated to specialist meeting participants and other people expressing interest in the research. This worked well, and has been repeated in several other initiatives, often with editors outside the center.

Perhaps the most important lesson of Initiative 1 was the value of a narrowly defined, tangible topic with immediate significance to the user community. It is relatively easy to interest people in the relevance of accuracy, and the immediate value of research results. Although the subject matter is very basic, and centers in a relatively obscure area of statistics, its results are graphic and immediately meaningful, and are relevant to many GIS activities, from data collection and interpretation through digitizing to report generation and mapping.

ANNOTATED LIST OF NCGIA PUBLICATIONS FROM INITIATIVE 1

The following section includes the references and abstracts to the 52 papers resulting from Initiative 1 to date. Additional references cited in the text are listed in a separate section at the end - many of these are by authors who contributed to the specialist meeting, and may have been stimulated or influenced by it in their own research. Authors affiliated with the Center are listed in bold.

Amrhein, C.G. and Schut, P. (1990) Data quality standards and Geographic Information Systems. *Proceedings: GIS for the 90s, Ottawa*. (Also presented at the Western Regional Science Association meetings, February 1990)

The rapid expansion of GIS applications has spawned a parallel increase in demands for information regarding the quality of spatial databases, particularly from government data collection agencies. But while positional accuracy is important, it is not a sufficient measure of accuracy for all applications. This paper reviews a range of accuracy measures for digital cartographic products in the context of one agency, Statistics Canada.

Anselin, L. (1989) What is special about spatial data? Alternative perspectives on spatial data analysis. *Technical Paper 89-4*. National Center for Geographic Information and Analysis, Santa Barbara, CA.

This background paper prepared for the Initiative 1 Specialist Meeting describes the unique statistical characteristics of spatial data, and some of the techniques developed for dealing with them. Also published by IMAGE (Anselin, L. In D.A. Griffith, editor, *Spatial Statistics, Past, Present and Future*. Monograph Series, Institute for Mathematical Geography, Ann Arbor, Michigan).

Anselin, L. (1990a) Spatial dependence and spatial structural instability in applied regression analysis. *Journal of Regional Science* 30: 185-207.

The stability of regression coefficients over the observation set ("regional homogeneity") is typically assessed by means of a Chow test or within a seemingly unrelated regression (SUR) framework. When spatial error autocorrelation is present in cross-sectional equations the traditional tests are no longer applicable. This is evaluated both in formal terms as well as empirically. A taxonomy of spatial effects in models for structural instability is introduced and its implication for testing is discussed. The performance of traditional tests, robust approaches, maximum-likelihood procedures and pretest techniques is compared by means of a series of simple Monte Carlo experiments.

Anselin, L. (1990b) Some robust approaches to testing and estimation in spatial econometrics. *Regional Science and Urban Economics* 20: 141-163.

Some robust approaches are outlined that form a basis for a more realistic statistical inference in spatial econometric models. Three specific issues are addressed: significance tests on coefficients in the spatial expansion method that are robust to the presence of heteroskedasticity of unknown form; heteroskedasticity-robust specification tests for spatial dependence; and bootstrap estimation in spatial autoregressive models. The techniques are presented in formal terms and their application to spatial analysis is illustrated in a number of simple empirical examples.

Anselin, L. and **Rey, S.** (1991) Properties of tests for spatial dependence in linear regression models. *Geographical Analysis* 23(2): 112-131.

The paper compares the properties of several tests for spatial dependence, based on a large number of Monte Carlo simulations on a regular lattice. The results provide an indication of the sample sizes for which the asymptotic properties of the tests can be considered to hold. They also illustrate the power of Lagrange Multiplier tests to distinguish between substantive spatial dependence, and spatial dependence as a nuisance (error autocorrelation).

Arbia, G. and **Haining, R.P.** (in press) Error propagation through map operations. *Technometrics*.

In image processing and GIS, a new map is constructed by carrying out a sequence of operations on a set of stored source maps. These operations typically include addition, ratioing and overlaying of two or more maps. But each source map may contain error. The paper investigates the effects of different types of source map error on the error structure of the resulting map and shows how these effects also depend on the spatial structure of the true source map.

Berman, O., Chiu, S.S., Larson, R.C., Odoni, A.R. and **Batta, R.** (1990) Location of mobile units in a stochastic environment. In P.B. Mirchandani and R.L. Francis, editors, *Discrete Location Theory*. Wiley, New York.

Much previous work on finding locations for mobile units has assumed that the characteristics of the street network are known and fixed, and optimized for such parameters as travel time. Unfortunately congestion levels on street networks vary dramatically, leading to severe sub-optimality. Uncertainty is represented by allowing the network to adopt several discrete states, and finding probabilistic optima.

Buyong, T. (1989) Utility mapping systems based on measurements. *Proceedings, URISA '89, Boston MA 2*: 222-230.

Many utility companies have taken significant steps towards implementing computerized utility mapping systems in a multipurpose cadastral system setup. The cadastral base map needs frequent upgrades due to ongoing improvements and revisions. Current implementations of coordinate-based utility mapping systems do not permit these upgrades to be reflected in the utility mapping system easily. A computerized mapping system based on measurements is proposed in which the changes in the base map are automatically propagated in the utility mapping system. This approach guarantees correct registration of utility maps and base maps.

Buyong, T. and Frank, A.U. (1989) Measurement based multipurpose cadastre. *Technical Papers, ASPRS/ACSM Annual Convention, Baltimore MD 5*: 58-66.

Many local level governments have taken significant steps towards implementing a multipurpose cadastre. Most of them follow the National Research Council study which recommends starting a multipurpose cadastral system with a good network of geodetic control points. This approach, however, has several major problems. The establishment of geodetic network is very costly and time consuming and often cannot be provided by local level governments. We propose a multipurpose cadastral system based on measurements where the implementation does not require the immediate completion of a geodetic control network.

Buyong, T. and Kuhn, W. (1990a) Local network adjustment for a measurement-based multipurpose cadastre. *FIG, XIX International Congress Proceedings, Helsinki, Finland 3*: 525-537.

In a measurement-based system, it is impractical to adjust all measurements in the database every time coordinate values of some points are required. Instead, the adjustments include only those neighboring measurements that significantly influence the results of the adjustment of the desired area. The local adjustment bears the same results as the global adjustment for most cadastral purposes.

Buyong, T. and Kuhn, W. (1990b) Local adjustment for cadastral measurement databases. *Proceedings, 1990 ACSM/ASPRS Annual Convention, Denver CO 3*: 19-27.

Early parcel-based land information systems were created mainly for land taxation and land management. Today's computerized systems however, serve multiple purposes with a large variation of functionalities. Some of the functions demand more up-to-date metric information. This scenario has resulted in the emergence of the concept of a cadastral measurement database. The demand for fast response and the high frequency of query make adjustment of the complete measurements in the database impractical. This paper proposes that only measurements in the locality of the query area need to be adjusted.

Buyong, T.B., Kuhn, W. and Frank, A.U. (in press) A conceptual model of measurement-based multipurpose cadastral systems. *Journal of the Urban and Regional Information Systems Association*.

A measurement-based multipurpose cadastral system uses measurements as the basic carrier of metric information. This concept is realized by allowing the processing of the measurements to be suspended until an item of information is needed. Least squares adjustment is the tool used to process the measurements. A direct manipulation user interface provides friendly interaction with the system. A measurement database furnishes convenient management of measurements and related data. The advantages of a measurement-based system are: incremental implementation, ease of updating, improvement of accuracy over time, correct integration of different data layers, preservation of background information as well as several economic benefits.

Carson, Y.M. and **Batta, R.** (1990) Locating an ambulance on the Amherst campus of the State University of New York at Buffalo. *Interfaces 20*(5) 43-49.

The paper applies a stochastic network approach to model uncertainty in travel times on the road network of the SUNY campus, and uses this to optimize locations for mobile units. Primary listing under Initiative 6.

Deichmann, U., Goodchild, M.F. and Anselin, L. (1990) A general framework for the spatial interpolation of socioeconomic data. *Proceedings, Advanced Computing in the Social Sciences Conference, Williamsburg VA.*

Spatial data are collected and represented as attributes of spatial objects embedded in the plane. Basis change is defined as the transfer of attributes from one set of objects to another. Methods of basis change for socioeconomic data are reviewed, and seen to differ in the assumptions they make about underlying density surfaces. These methods are extended to more general cases, and illustrated using California data. The implementation of this framework within a GIS is discussed.

Goodchild, M.F. (1989) Accuracy of spatial databases: Initiative 1 specialist meeting report. *Technical Paper 89-1.* National Center for Geographic Information and Analysis, Santa Barbara, CA.

This report on the I1 Specialist Meeting contains a summary of the discussion and abstracts of the paper presentations.

Goodchild, M.F. (1990) Modeling error in spatial databases. *Proceedings, GIS/LIS 90, Anaheim CA 1:* 154-162.

Geographic information systems process spatial data to generate information products through such operations as overlay, area measurement and buffer zone generation. The paper reviews the techniques available for modeling error in the various data models commonly used in GIS, and the relationships between them. Methods are discussed for incorporating error model parameters in the database, for tracking the propagation of error through computational processes and for reporting uncertainty in system products. The paper provides a general introduction to the research completed under the NCGIA's Initiative 1: Accuracy of Spatial Databases.

Goodchild, M.F. (1991a) Keynote address: symposium on spatial database accuracy. *Proceedings, Symposium on Spatial Database Accuracy, Melbourne, Australia.* Department of Surveying and Land Information, University of Melbourne: 1-16.

The accuracy of spatial databases has been the focus of NCGIA's Research Initiative 1. Uncertainty is a particularly significant problem in GIS because spatial data tend to be used for purposes for which they were never intended, and because the accuracy problem in GIS requires consideration of both object-oriented and field-oriented views of geographic variation. The paper defines the relevant terms, and provides a review of the current state of knowledge in the area of error models for spatial data. Efforts to build data quality standards are also reviewed.

Goodchild, M.F. (1991b) Issues of quality and uncertainty. In J.C. Muller (ed.) *Advances in Cartography.* Elsevier: 113-139. Also appears (1991) in *Proceedings, State of Indiana Geographic Information System Conference, Indianapolis, November.* Indiana University GIS Alliance: 17-53.

A general overview of the data quality issue in GIS.

Goodchild, M.F. (in press) Geographical data modeling. *Computers and Geosciences.*

Data modeling is defined as the process of discretizing spatial variation, but is often confused with issues of data structure, and driven by available software rather than by a concern for accurate representation. The paper reviews the alternative data models commonly available in spatial databases, and assesses them from the perspective of accurate representation of geographical reality. Extensions are discussed, particularly for three dimensions and time dependence. (Presented at Inter Agency Committee on Geomatics seminar, Ottawa, March 1990; GIS Design Models Conference, Leicester, March 1990).

Goodchild, M.F., F.W. Davis and D.M. Stoms (1991) The use of vegetation maps and geographic information systems for assessing conifer land in California. *Technical Paper 91-23.* National Center for Geographic Information and Analysis, Santa Barbara, CA.

This report to the California Department of Forestry and Fire Protection summarizes research into the nature and sources of error in medium to small-scale vegetation maps used for state-wide forestry conservation planning.

Goodchild, M.F. and Gopal, S. (1989) *Accuracy of Spatial Databases.* Taylor and Francis, New York, 290pp.

This book is a collection of 23 of the papers presented at the Specialist Meeting for I1 and subsequently revised and edited. It contains several contributions from NCGIA personnel (Santa Barbara: Goodchild, Theobald, Tobler, Slater, Kennedy; Buffalo: Batta, Fotheringham, Amrhein), plus introductions to each section by the editors. These NCGIA contributions are not listed separately in this report.

Goodchild, M.F. and Klinkenberg, B. (in press) Statistics of channel networks on fractional Brownian surfaces. In N. Lam and L. de Cola (eds.) *Fractals in Geography*. Wiley, New York.

The random topology model of channel networks is inappropriate as a null hypothesis in testing for the presence of geological controls, because packing and surface continuity impose additional constraints of a geometric nature. A range of surface conditions are simulated using fractional Brownian processes, and drainage networks extracted under a number of assignment rules. The paper raises a number of questions concerning interpretation of significance tests using the random model.

Goodchild, M.F., Sun Guoqing and **Yang Shiren** (in press) Development and test of an error model for categorical data. *International Journal of Geographical Information Systems*.

An error model for spatial databases is defined here as a stochastic process capable of generating a population of distorted versions of the same pattern of geographic variation. The differences between members of the population represent the uncertainties present in raw or interpreted data, or introduced during processing. A new error model is defined in this paper for categorical data. Its application to soil and land cover maps is discussed in two examples: the measurement of area, and overlay. Specific details of implementation and use are reviewed. The model provides a powerful basis for visualizing error in area class maps, and for measuring the effects of its propagation through GIS processes.

Goodchild, M.F. and Wang Min-hua (1989) Modeling errors for remotely sensed data input to GIS. *Proceedings, AutoCarto 9, Baltimore MD 530-537*.

Different views of spatial resolution and accuracy present a major obstacle to the integration of remote sensing and GIS. Accuracy in remote sensing is modeled using probabilities of class membership in each pixel; in vector-based GIS it is modeled using concepts such as the epsilon band. The problem of linking the two views of accuracy reduces to one of realizing a stochastic process which must satisfy conditions of prior and posterior probabilities, and spatial dependence. The authors propose two suitable methods, one storage intensive and the other computationally intensive.

Goodchild, M.F. and **Yang Shiren**. (1989) A hierarchical spatial data structure for global geographic information systems. *Technical Paper 89-5*. National Center for Geographic Information and Analysis, Santa Barbara, CA. Also appears (1990) in *Proceedings, Fourth International Symposium on Spatial Data Handling, Zurich 2*: 911-917. Also (1992) *Computer Vision, Graphics and Image Processing: Graphical Models and Image Processing 54(1)*: 31-44.

Hierarchical spatial data structures such as the quadtree offer distinct advantages of data compression and fast access, but are difficult to adapt to the globe. Following Dutton, we propose to project the globe onto an octahedron, and then recursively subdivide each of its triangular faces into four triangles. We provide procedures for addressing the hierarchy, and for computing addresses in the hierarchical structure from latitude and longitude, and vice versa. At any level in the hierarchy the finite elements are all triangles, but are only approximately equal in area and shape; we provide methods for computing area, and for finding the addresses of neighboring triangles. (Presented at Advanced Computing in the Social Sciences Conference, Williamsburg, VA, April 1990; University of Victoria, Simon Fraser University, February 1990).

Goodchild, M.F., Yang Shiren and **Dutton, G.** (1991) Spatial data representation and basic operations for a triangular hierarchical data structure. *Technical Paper 91-8*. National Center for Geographic Information and Analysis, Santa Barbara, CA.

A triangular hierarchical data structure has been proposed as the basis for a global geographical information system. In this paper we briefly review one such scheme based on recursive subdivision of an octahedron, and conversion algorithms to and from latitude/longitude. Schemes for representing point, line and area objects on the earth's surface are described. We present algorithms for identifying triangle neighbors, region filling and object dilation.

Gopalan, R., Kolluri, K.S., Batta, R. and **Karwan, M.H.** (1990) Modeling equity of risk in the transportation of hazardous materials. *Operations Research 38(6)*: 961-973.

The authors develop and analyze a model to generate an equitable set of routes for hazardous material shipments. The objective is to determine a set of routes which while minimizing total risk of travel, will also spread the risk equitably among the zones of the geographical region in which the transportation network is embedded. The findings indicated that one can achieve a high degree of equity by modestly increasing the total risk, and by embarking on different routes so as to evenly spread the risk among the zones.

Klinkenberg, B. and **Goodchild, M.F.** (in press) The fractal properties of topography: a comparison of methods. *Earth Surface Processes and Landforms*.

The fractal characteristics of fifty-five digital elevation models from seven different US physiographic provinces are determined using seven methods. The self-similar fractal model is found to provide a very good fit for some landscapes, but an imperfect fit for others. Thus, outright rejection of this model does not appear to be warranted, but neither does a blind application.

Kumler, M.P. (1990) A quantitative comparison of regular and irregular digital terrain models. *Proceedings, GIS/LIS 90, Anaheim CA 1*: 255-263.

This paper presents the results of an investigation into the relative accuracies of two competing digital terrain models: the gridded digital elevation model (DEM) and the variable-resolution triangulated irregular network (TIN). Study sites were selected from around the US on the basis of their different surface characteristics and the availability of digital elevation data.

Kumler, M.P. and **Goodchild, M.F.** (1991) A new technique for selecting the vertices of a TIN, and a comparison of TINs and DEMs over a variety of surfaces. *Technical Papers, 1991 ACSM-ASPRS Annual Convention, Baltimore MD 2*: 179.

A technique is presented for selecting the vertices for a TIN from digitized contour lines and selected elevation points. The procedure relies on a Douglas-Peucker generalization of the contour lines to capture a large set of points along ridges and channels. This set is supplemented by elevation values at peaks and pits, along edges and at corners of the model, and at other locations necessary to improve the overall fit of the TIN. The procedures for identifying the supplementary points are described in detail.

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

This paper presents a research paradigm for exploring the formulation of source accuracy indices and functions to propagate such indices and to document the quality of derived GIS data products. Error propagation functions are defined by their association with a particular GIS transformation and a particular accuracy index. These resulting associations and indices are metadata manipulatable by spatial, thematic and temporal error through GIS applications. The result is a way in which users can experiment with measures that document the accuracy of GIS derived spatial data products.

Lanter, D.P. and **Veregin, H.** (1991) A lineage information program for exploring error propagation in GIS applications. *Proceedings, 15th ICA Conference, Bournemouth UK 1*: 468-472.

A lineage-based system is described for propagating error through data processing steps in layer-based GIS. The system can be used as an exploratory tool for assessing the impacts of competing assumptions about error propagation mechanisms. The system also facilitates the development of optimal strategies for improving the quality of derived data products.

Lanter, D.P. and **Veregin, H.** (in press) A research paradigm for propagating error in layer-based GIS. *Photogrammetric Engineering and Remote Sensing*.

This paper focuses on the nature of error in spatial databases and the implications of this error for spatial data transformations in GIS applications. It describes an error propagation research paradigm as an information flow linking successively more formal components of error propagation in a GIS context. These components include development of conceptual models of error, creation of formal indices to measure error in spatial databases, implementation of mathematical functions to transform error indices and model the propagation of error as it is processed, and evaluation of the indices to gain insight into the utility of conceptual models used in error measurement and propagation. The paradigm enables researchers to formulate, manipulate and experiment with components of error propagation to determine their implications for decision making. The applicability of the paradigm is illustrated with a simple GIS application in which error is propagated from sources to final product through a sequence of data transformation functions.

Mark, D.M. and **Csillag, F.** (1989) The nature of boundaries on area-class maps. *Cartographica 21*: 65-78.

Most research on cartographic line generalization has concentrated on linear features such as coastlines, rivers and roads, but these methods may not be appropriate for generalization of other types of lines. The paper presents a model of boundaries between classes on area-class maps, such as soil maps. Appropriate generalization may involve the construction of surfaces representing probabilities of class membership.

Noronha, V.T. and **Goodchild, M.F.** (in press) Modeling inter-regional interaction: implications for defining functional regions. *Annals, Association of American Geographers*.

The functional region is a spatial phenomenon arising from selectivity in human interaction. Conceptualizing and delimiting the functional region must therefore be based on a concept of spatial interaction in anisotropic environments. This paper extends the well known spatial interaction model, and offers a theoretically derived model of inter-regional spatial interaction. The Inter-Regional Interaction Model explains the behavioral phenomena that we recognize as functional distance and functional regions. Calibration of the model amounts to objective delimitation of functional regions. Regionally biased interactions are simulated, and the embedded regional structure successfully recovered in a series of tests, even when substantial error components are introduced into the simulations. A heuristic is developed for large problems. The calibration technique is demonstrated on US student migration matrices. A cluster of southeastern states consistently emerges as a functional region distinct from the north.

Painho, M.O., Duncan, B.W. and Davis, F.W. (1990) Utilizing GIS technology to assess boundary consistency and thematic map accuracy. *Proceedings, GIS/LIS 90, Anaheim CA* 1: 68

Geographic Information System technology is now used extensively both in the private and public sectors for resource analysis and planning. Vegetation cover is an especially important and dynamic layer in most GIS applications. Use of the vector-based (polygonal) data structure to represent vegetation can involve considerable cartographic generalization and simplification, with important consequences in land surface inventory and analysis. In this paper we use GIS capabilities of point, line and polygon overlay to examine the accuracy and reliability of land surface analysis.

Pullar, D. (1990) Comparative study of algorithms for reporting geometrical intersections. *Proceedings, Fourth International Symposium on Spatial Data Handling, Zurich* 1: 66-76.

Two broad approaches to solving the geometrical intersection problem can be distinguished based on how adaptive the solutions are to the input data. "Theoretical" algorithms are only dependent upon the number of line segments, whereas the behavior of "pragmatic" algorithms depends on the statistical character of the underlying data. This paper compares the performance of theoretical and pragmatic solutions to the segment intersection problem using both algorithmic analysis and empirical tests.

Pullar, D. (1991) Spatial overlay with inexact numerical data. *Proceedings, AutoCarto 10, Baltimore MD* 6: 313-329.

A methodology for the operation of spatial overlay is presented in this paper. A general framework for spatial overlay based on concepts in epsilon geometry is developed to cope with the problems of computational errors and handling inaccurate numerical data. These problems normally cause topological inconsistencies and generate spurious effects in the result. A mapping is defined to accommodate the edges and vertices in all spatial layers so they are unambiguously aligned within a prescribed tolerance. Geometrical arguments are given to show the correctness of this approach.

Pullar, D. and Beard, M.K. (1990) Specifying and tracking errors from map overlay. *Proceedings, GIS/LIS 90, Anaheim CA* 1: 79-87.

This paper presents tools to treat spatial uncertainty for representing and combining the geometry for map layers. As a speculative approach, we adopt an error model for geometrical objects based upon a normal probability distribution. The error model is used to track error in two applications, line simplification and map overlay.

Ray, C-W. (1990) Remote sensing and GIS: establishing consistency between raster and vector data models. Unpublished Masters Thesis, Department of Geography, University of California, Santa Barbara.

This thesis is a case study in which an attempt was made to establish consistency between raster based satellite imagery and vector based cartographic data models. The GIS or cartographic data model was based on 1987 land-cover and land-use maps in a polygon format and was compared to a raster-based Landsat TM image from June 11, 1986. There were six bands of 30m resolution in the image and 251 polygons in 14 classes in the GIS. Techniques of spatial statistics were used to evaluate the relationship between the two datasets.

Rogerson, P.A. (1990a) Buffon's Needle and the estimation of migration distances. *Mathematical Population Studies* 2: 229-238.

The paper suggests a procedure for estimating migration distances from data on the proportion of migrants crossing regional boundaries. The method makes use of Buffon's Needle, a problem in geometrical probability from the 18th Century. An application to migration distances in the US is given.

Rogerson, P.A. (1990b) Migration analysis using data with time intervals of differing widths. *Papers of the Regional Science Association* 68: 97-106.

Migration data collected for periods of differing lengths will yield inconsistent population forecasts and alternative interpretations of mobility levels and migration patterns. Examples are given to illustrate the effects of migration interval choice. In addition to the level of mobility, the geographic pattern of migration flows is also affected by the choice of interval width.

Slater, P.B. (1991a) A quasi-dynamical model of the space-time evolution of a national population through internal migration. *Geographical Analysis* 23(2): 174-178.

Models are proposed for taking either a recorded internal migration table for the geographic units of a nation and two times (start and end) or simply the population distributions over the units at the two times, and estimating the probability distributions over the possible migration histories of individuals. The probability distribution is sought which reproduces the recorded population distributions and the naturally weighted convex combination of these distributions at the intermediate points in time.

Slater, P.B. (1991b) Physical-based models of internal migration. *Applied Mathematics and Computation* 43(1): 95-103.

The use of physical principles as paradigms to model a socioeconomic/demographic phenomenon - the migration of people within a nation - is discussed. In particular, an entropy-maximization model, analogizing migrants to particles interacting through a Lennard-Jones-type potential, is proposed. Certain constraints - dictated by conservation principles and the specific nature of internal-migration data (irregularly coarse-grained configurations of residences at two points of time) - are imposed on the solution. The results of this model can serve as estimates of the probability transition rates employed by Weidlich and Haag in their master-equation approach to the analysis of internal migration.

Smith, T.R., Zhan, C.X. and Gao, P. (1990) A knowledge-based, 2-step procedure for extracting channel networks from noisy DEM data. *Computers and Geosciences* 16(6): 777-786.

Noise in DEM data creates significant problems because of its influence on the estimation of aspect, which in turn is used to determine flow directions in hydrological networks. A knowledge-based procedure is presented for using higher-level information about the nature of stream networks and terrain to improve on the extraction process in useful ways.

Stoms, D.W., Davis, F.W., Cogan, C.B., Painho, M.O., Duncan, B.W. and Scepan, J. (1990) Sensitivity of habitat models to uncertainties in GIS data: a California Condor case study. *Proceedings, GIS/LIS 90, Anaheim CA* 1: 69-78.

Spatial decision makers need to know the reliability of the output products from GIS analysis. For many GIS applications, however, it is not possible to compare these products to an independent measure of truth. Sensitivity analysis offers an alternative means of estimating reliability. In this paper, we present a GIS-based statistical procedure for measuring the sensitivity of wildlife habitat models to data quality and model assumptions. The approach is demonstrated in an analysis of habitat associations derived from a GIS database of the distribution and habitat of the endangered California Condor. Sensitivity analysis indicated that the condor habitat associations are relatively robust and thus has increased our confidence in our initial findings.

Theobald, D.M. (1990) Automated delineation of hydromorphological features on a triangular irregular network-based digital elevation model. Master's thesis, Department of Geography, University of California, Santa Barbara.

Information about terrain is especially useful in hydrological and geomorphological modeling. Automated techniques to represent terrain surfaces and delineate drainage networks have recently been developed, making it feasible to describe the terrain quantitatively. Not only do different data structures have different characteristics, but the manner in which flow is modeled with these data structures has a large impact on the quality of the simulation. The thesis examines hydrological modeling on TIN surfaces, and artifacts which arise under different TIN generation strategies.

Theobald, D.M. and Goodchild, M.F. (1990) Artifacts of TIN-based surface flow modeling. *Proceedings, GIS/LIS 90, Anaheim CA* 2: 955-967.

Information about terrain is basic to nearly any type of environmental research and is especially useful in hydrological and geomorphological modeling. A useful criterion for determining terrain accuracy for the purpose of hydrological modeling is aspect, since it alone determines flow direction on a surface. This paper reviews the three conventional models of terrain: DEM, contour and TIN. Arguments for the efficiency, accuracy and consistency of the TIN model are summarized. Formal

concepts important in modeling hydrology using TINs are presented, and algorithms are developed for the major functions. These are tested on a model of a test topography in Kansas, and several classes of artifacts are identified. These originate in the methods used to construct the TIN, but have significant effects on the success of the hydrological modeling. Measures to reduce their frequency and consequences are discussed.

Tobler, W.R. (1990) GIS transformations. *Proceedings, GIS/LIS 90, Anaheim CA 1*: 163-166.

If one considers GIS information to be interpretable as point, line, area or field phenomena, then there are sixteen common classes of transformation. Within these classes one can further distinguish between categorical and numerical data, to obtain some eighty distinct possible classes of transformation. An attempt is made to enumerate, via example, many of these transformations, and to comment on implementation algorithms.

Tobler, W.R. (in press) The resel based GIS. *International Journal of Geographical Information Systems*.

There now exist several micro-computer processing systems incorporating algorithms and display techniques which are appropriate for the class of objects known as images. Zooming, histogram equalization, contrast stretching, edge enhancement, filtering and ratioing are common options used in these systems. The author proposes a system for the equivalent processing of general resolution elements (resels). Such a system requires a high resolution display and a new repertoire of processing algorithms.

Veregin, H. (1989a) Accuracy of spatial databases: annotated bibliography. *Technical Paper 89-9*. National Center for Geographic Information and Analysis, Santa Barbara, CA.

Veregin's bibliography includes over 250 references on accuracy, error and uncertainty in spatial data. Each includes a short paragraph of annotation. The bibliography is intended as a tool for supporting research both inside and outside the Center.

Veregin, H. (1989b) Taxonomy of error in spatial databases. *Technical Paper 89-12*. National Center for Geographic Information and Analysis, Santa Barbara, CA.

The taxonomy is intended as a user's guide to current knowledge of error problems in GIS, and is keyed to Veregin's bibliography. Its 90 pages include discussion of error source identification; detection and measurement; propagation modeling; strategies for error management; and strategies for error reduction.

ADDITIONAL REFERENCES

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

Fisher, P.F. (1991a) Simulation of the uncertainty of a viewshed. *Proceedings, AutoCarto 10, Baltimore* 205-218. ASPRS/ACSM, Bethesda, MD.

Fisher, P.F. (1991b) Modelling soil-map inclusions by Monte Carlo simulation. *International Journal of Geographical Information Systems* 5(2): 193-208.

Goodchild, M.F. and Wang Min-hua (1988) Modeling error in raster-based spatial data. *Proceedings, Third International Symposium on Spatial Data Handling, Sydney*, 97-106. International Geographical Union, Commission on Geographical Data Sensing and Processing, Columbus, Ohio.

Heuvelink, G.B.M. (in press) An iterative method for multi-dimensional simulation with nearest neighbour models. In P.A. Dowd, editor, *Proceedings of the Second CODATA Conference on Geomathematics and Geostatistics, 10-14 September 1990, Leeds*. Sciences de la Terre.

Heuvelink, G.B.M., P.A. Burrough and A. Stein (1989) Propagation of errors in spatial modelling with GIS. *International Journal of Geographical Information Systems* 3(4): 303-322.

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

Lodwick, W.A., W. Monson and L. Svoboda (1990) Attribute error and sensitivity analysis of map operations in geographical information systems: suitability analysis. *International Journal of Geographical Information Systems* 4(4): 413-428.

NCGIA (1989) The research plan of the National Center for Geographic Information and Analysis. *International Journal of Geographical Information Systems* 3(2): 117-136.

Openshaw, S. and P.J. Taylor (1979) A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In N. Wrigley, editor, *Statistical Applications in the Spatial Sciences*. Pion, London.