AURISA 94
The Annual Conference of the Australasian Urban and Regional Information Systems Association, Inc
Sydney, New South Wales
21-25 November 1994

SPATIAL ACCURACY

Michael F. Goodchild

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

Phone 1 805 893 8049
Fax 1 805 893 7095
Email good@ncgia.ucsb.edu

ABSTRACT

Modern spatial information systems create precise digital representations of features that are often inherently fuzzy, inaccurately located, or otherwise subject to a variety of forms of uncertainty. In this environment it is all too easy to impute levels of accuracy to the results of analysis and processing that are wildly inconsistent with the truth. The paper begins with a series of examples of the uncertainty present in map data, and the consequences of its use in GIS. In many cases, the map acts as a channel of communication between the scientist or surveyor in the field, and the eventual user of geographic data; uncertainty can be interpreted from this perspective as the result of inefficiency in the communication channel, providing the user of the data with only a small fraction of the knowledge available to the observer or interpreter in the field. Two approaches are proposed to deal with this problem. The first requires a change of methodology, and makes the database the communication channel rather than the map--cartographic representations are seen in this approach as views of the database with their own objectives and priorities. The second adopts a statistical perspective by modeling uncertainty through the construction of error models. These are illustrated, and an example of the use of an error model to estimate error in the measurement of area in a GIS is discussed. The paper ends with a discussion of alternative methods for visualization of spatial accuracy within the context of error modeling.

INTRODUCTION

The past five years have seen an increasing degree of concern over the issue of accuracy in geographic databases and GIS. "Accuracy of Spatial Databases" was the topic of the first research initiative of the U.S. National Center for Geographic Information and Analysis (NCGIA) from 1988 to 1990 (Goodchild and Gopal, 1989; Goodchild, 1992). Since then there have been conferences devoted to the subject in Australia (Hunter, 1991) and the U.S. (Congalton, 1994). A second NCGIA initiative on "Visualization of Geographic Data Quality" (1991-1994) looked at methods for communicating information on quality and uncertainty to users of maps and GIS. The data quality issue has surfaced in several other NCGIA initiatives, including "Integration of GIS and Remote Sensing" (Lunetta et al., 1991).

Although the terms used to describe the problem vary widely—"accuracy", "uncertainty", "error", "quality" are all used, and with substantial overlap of meaning—the core issue is really quite simple. Traditionally, maps have been viewed as comfortable, scholarly, pleasant, civilized things, to be displayed on walls just as we display books on shelves. But over the past 30 years we have brought the contents of maps into the precise, brutish world of digital computers, where they are treated as measurements of the world, or collections of facts with fixed, well-defined meaning. When the subjective and somewhat vague content of a map is digitized, analyzed, and the results of analysis returned to the user, the authority of the digital machine replaces that of the map author or the collector of the data. Suddenly, contours that were drawn to convey an impression of the form of the Earth's surface become the basis for precise measurements of elevation, slope, or aspect, to as many digits of precision as the machine can produce.

To provide context for this discussion, it might be useful to review recent figures on the magnitude of this activity. Daratech (1994) reported total sales by the GIS software industry of $450 million in 1993. A survey by the U.S. Office of Management and Budget reported annual expenditures on digital geographic data activities within the federal government alone of $4.5 billion. Estimates of the total annual value of U.S. activity range as high as $10-14 billion.

The digital world of geographic data processing is very precise. The term "single precision" implies about 1 part in 10^7, but many GIS vendors now offer double precision storage and manipulation of coordinates and data, or 1 part in 10^{14}. In the largest possible coordinate system for geographic data, one covering the entire Earth, 1 part in 10^{14} implies a need for positional accuracy of about 10^{-7}m, which is about the size of a molecule. None of our measuring instruments are capable of capturing data that even approach that level of accuracy, so why do GIS vendors perceive a market for it? Does the GIS user's sense of accuracy depend more on the machine's precision than on the accuracy of the data source?

Almost too late, the geographic information community has come to realize how unrealistic this collective view can be. Our knowledge of the shape of the Earth limits our ability to establish absolute position to better than about 10m, although much higher levels of relative accuracy can be achieved over short distances. The maps that form the source of much of our digital geographic data resource are no more than perspectives, and often embed the agenda of the creator, a point that is made very clearly in Dennis Wood's book *The Power of Maps* (Wood, 1992). Ormeling (1992, p. 65) writes:

> "If cartographers are aware of the fact that maps can never be scientifically objective, this does not include users of geographical information systems. Nowhere else is the infallible character of maps adhered to to such a degree as in GIS user circles, where various map-based files are combined without even an approximate knowledge as to how this combination will affect the accuracy of the information."

Ormeling provides a succinct illustration in a pair of topographic maps, both of the same area of the South Tirol, but one made under the Austrian administration prior to 1919 and the other under the subsequent Italian administration.

Two years ago, I was a participant in a geographic data planning exercise in the State of Victoria. One of the GIS applications identified in that study was a simple cadastral map of suburban land parcels, to be created by digitizing a paper map, construction of a database, and then plotting. It was proposed that the plotted map show the identification of each parcel, along with its area. Area could be calculated from the database using the routine GIS function, so there would be no need to input it as a parcel attribute.

The discussion turned to the precision of parcel area. Given the precise nature of a GIS, it seemed reasonable to ask that each area be given in square meters to two decimal places–certainly within the precision limitations of a modern GIS. The average parcel area was about 1,000 sq m, so this represented precision of 1 part in 10^5, far less than the limits of double precision. But how did this apparently reasonable request compare to the accuracy actually achievable by this process? Chrisman and Yandell (1988) have provided a statistical analysis of the problem, but in this case a simple back-of-the-envelope calculation was adequate. Assuming that paper-based mapping and digitizing are capable of sustaining a positional accuracy of 0.5mm at the scale of the map, a figure that is broadly compatible with map accuracy standards and common digitizing practice, Table 1 shows the original map scales needed to sustain given levels of accuracy of an area estimate, based on different parcel sizes. For a 1,000 sq m parcel, a map of 1:3,000 will sustain accuracy of 10%, that is, will ensure reliability of the 100s digit in the area estimate. To achieve 1% accuracy, or reliability of the 10s digit, it is necessary to map at 1:300 or better. For accuracy of an area estimate to the second decimal place, we would need an original cadastral map at a scale of 3:1, or three times larger than reality.

Parcel Area	1% Error	10% Error
100 sq m	1:100	1:1,000
1000 sq m	1:300	1:3,000
1 ha	1:1,000	1:10,000
10 ha	1:3,000	1:30,000
100 ha	1:10,000	1:100,000
1000 ha	1:30,000	1:300,000

Table 1: Minimum scales for a digitized map to achieve prescribed levels of accuracy in the estimation of a given polygon area.

This is a simple example of error propagation, or the effects of underlying data uncertainty on the products of GIS analysis. Given the simple model used and the straightforward nature of the calculation, it is perhaps surprising that current GIS software does not attempt to estimate confidence limits on area estimates, or to truncate the digits of reported area calculations to make them consistent with known accuracy. It is necessary to supply an estimate of positional accuracy, but that is often readily available in metadata, or in the positional tolerance values established for cleaning operations or for overlay.

What is perhaps more significant is that the achievable accuracy comes as a surprise, not only to naive users of GIS but to those who are supposedly wise to the accuracy issue. A simple sense of achievable accuracy in GIS operations should surely be part of every GIS education. We currently devote time and attention to accuracy standards, and students are expected to know the positional accuracies of standard topographic maps. But we do not currently insist on the ability to translate such knowledge into its implications for basic GIS functions.

The experience of the past decade has shown that accuracy in geographic data is a substantial problem. To address it will require the combined efforts of cartographers, GIS users, geographers, and specialists with an understanding of the levels of accuracy present in each major type of geographic data. The remainder of the paper discusses two approaches to the problem, one methodological and one statistical. Because the approaches are rooted in different paradigms, it is possible to pursue both of them simultaneously.

METHODOLOGICAL APPROACH: RETHINKING THE ROLE OF MAPS

Traditionally, geographic databases have been viewed as containers of maps, a perspective that is reinforced by the familiar layer-cake view of GIS. The map captures knowledge of the world, and provides input to a database through processes of digitizing and scanning. In some cases, the process of creation of the map is measurement-like, as in photogrammetry and other aspects of topographic mapping. But in other cases the process uses the knowledge of a field expert, such as a soil scientist, or forester, and is not likely to be replicable from one observer to another.

In the latter case, the map becomes a kind of communication channel between the expert and the user, a means of transferring the expert's knowledge and interpretation of a complex real world in a form useful to the end user. Several trends are making the operation of this communication process increasingly critical. The widespread adoption of GIS has meant that integrative studies are now being done that combine input from a variety of disciplines, but without specialized knowledge of each individual layer of data. For example, global climate models are now incorporating information on soil properties, but soil science is not likely to become a major element in the training of atmospheric scientists. Increasing emphasis on sharing of scientific data, and the widespread popularity of the Internet for scientific communication, have meant a steady increase in the physical and intellectual separation between data producer and data user. In this context, the notion that spatial metadata might allow a user to determine the fitness of a data set for use in a particular project takes on even greater significance–are we really asking a soil scientist to educate an atmospheric scientist through the medium of a metadata record?

There is an apocryphal story about a study that once tried to measure the efficiency of the map communication channel for soil science. If one took the total information available to the soil scientist following a standard field survey, and compared it to the total information available to the end GIS user of the soil map, one could evaluate a simple ratio as an estimate of the channel's communication efficiency. Of course numerous assumptions would have to be made about the measurement of information, such as those made in communication theory. But it is clear that the number would be low–very little of the field knowledge actually gets through to the end user if the medium of communication is the map. For example, it is impossible to include the chemical properties or stratigraphic characteristics obtained from field samples, or interpreted air photos, yet all of these are potentially useful items of information.

A revised vision that might help to address many of these issues has emerged repeatedly in discussions of GIS. In this view, the database captures what is known about reality, within a data model or structure that is designed to fit the observer's method of data collection. In the soil case, it might include digital representations of air photos, stratigraphic records, and records of chemical analysis. Far more information can be put into a database than can be shown on a map, because the database is not constrained by limitations of planarity, or information density. Boundary widths can be adjusted to reflect transition zones, rather than forced to be infinitely thin, and soil properties can be described in terms of continuous spatial variation, rather than forced to be homogeneous within areas. Maps become views of the database, created using rules that emulate the cartographic process, and with user control of parameters such as scale, and minimum mapping unit. In this model, the analyst has access to the contents of the database rather than to a digital representation of a map, allowing access to all of the soil scientist's knowledge.

Clearly we are a long way from being able to implement this vision at this time. We do not have the rule base that captures the cartographic process, and we have massive cultural and institutional investments in the traditional process. But several recent developments may bring the revised vision closer to reality.

First, we now have many geographic technologies that allow direct measurement in the field, transferring measurements to database without the intervention of maps. They include the global positioning system (GPS), remote sensing, and various forms of automated photogrammetry and direct capture of topographic features. The GPS-equipped vans that are now in use capturing data on highways also represent a powerful new technique for direct measurement.

Second, we are now seeing the introduction of software environments that support creation of databases in the field. This is sharply distinct from the GIS tradition of supporting analysis of maps, as it integrates the process of observation and field interpretation with analysis and modeling in one homogeneous environment. Condor Earth Technologies Inc recently announced VECTORMAP®, an integration of binoculars and a drawing pad that allows direct GIS data collection with automated measurement of direction, range, and elevation. Several vendors now offer field mapping systems using various forms of personal digital technology that make use of new framework data sets such as the Digital Orthophoto Quads (DOQs) as field digital backcloths.

STATISTICAL APPROACH: ERROR MODELING AND VISUALIZATION

While new visions of the interaction between field expert, user, and database may address many of the underlying issues of geographic data accuracy, it is clear that uncertainty is an inevitable property of geographic data. Very few geographical facts can be said to be free of any kind of uncertainty, and in general any geographic database is an approximation to geographic truth. Thus methods must be devised for capturing knowledge of uncertainty, storing it in databases, communicating it to users, and determining its impacts on decisions.

Over the past five years significant advances have been made in modeling uncertainty, through the development of error models. In this context, the term "error" is used in its statistical sense, and does not carry its normal connotation of "mistake". Tools based on these models are becoming available, and we may even see error modeling tools embedded in the major GIS packages before long.

The concept of an error model is familiar and intuitive for simple measurements, but much more complex for geographic databases. Consider for example the case of an opinion poll, and a result expressed as a percentage, say 43%. We know that other samples of respondents might have produced a slightly different result, say 45%, or 42%, and we may even be familiar with the concept of a bell curve that describes the relative likelihood of different results, centered around the known value of 43%. Its width or spread describes the variability to be expected between different samples, and is usually expressed as the standard deviation, standard error, or root mean square error (RMSE)–these terms are essentially synonymous. For public consumption, the measure of width is normally taken to be the one that encloses 95% of likely results and excludes 5%–the 95% confidence limits, or 1.96 times the standard deviation, or the limits of the true value "19 times out of 20".

For geographic data, the equivalent of a single percentage measurement is an entire map. Different soil scientists or photogrammetrists or digitizer operators will produce different digital maps of the same area, and the variation between them is captured by a statistical error model. Parameters are needed to describe the amount of variation. Each possible map that might be generated by the error model is termed a "realization".

It is not immediately clear why geographic databases cannot be regarded as collections of single measurements, and their uncertainty modeled as separate error models for each measurement. For example, a digital elevation model (DEM) consists of a regular array of point elevation measurements, and the RMSE of each measurement is often known and published. From this, it is possible to make simple calculations using the error model. For example, suppose the RMSE for a DEM is 7m, and consider a point with a published elevation of 100m. From the statistics of the bell curve (normal or Gaussian distribution) we know that the probability that the true elevation is above 100m is 50%, and the probability that it is above 107m is 34%, and similar probabilities can be determined for any range of elevations. However, a common form of GIS analysis is the calculation of slope, based on elevations of neighboring DEM points. To estimate the uncertainty in a slope calculation, it is necessary to have an error model for the joint behavior of errors at neighboring points. If errors are the same for all points in a DEM, e.g., all high by 1m, or all low by 5m, then there is no error in slope however large the elevation errors. But if errors are independent at neighboring points, such that one point may be high by 1m and its neighbor low by 5m, then uncertainty in slope is so large that estimates are effectively useless. Technically, uncertainty of slope is estimated as follows:

$$\sigma^2_{slope} = 2\sigma^2_e \frac{(1 - r)}{h^2}$$

where σ_e is the RMSE of elevation estimates (7m), h is the spacing, and r is the correlation between neighboring elevation errors, for slopes calculated from two adjacent point elevations (see Heuvelink, 1993, for a more comprehensive analysis of this problem).

Error models for DEMs have been explored by Fisher (1992), and Haining and Arbia (1993) discuss a general error model for raster data layers where the attribute is measured on a continuous scale. A powerful package for modeling and propagating errors in this type of data known as ADAM has been developed at the University of Utrecht (Heuvelink, 1993), and tools based on standard packages such as ARC/INFO, GRASS, and IDRISI are becoming available.

An equivalent error model for categorical data has been described by Goodchild, Sun, and Yang (1992). In this case, the objective is to model the variability between digital representations of an area class map, such as a land cover map, that can be ascribed to uncertainty or error. The approach includes the ability to model both the transition of classes between adjacent zones, where maps force instantaneous change, and also inclusions of other classes within zones. In this case also it is necessary to model the spatial structure of errors, since one cannot assume that every cell in a raster representation has an error that is independent of its neighbors.

Given an error model, it is possible to investigate the propagation of uncertainty into GIS products, as discussed earlier in the case of area measurement from a cadastral map. Just as in the case of slope estimation discussed above, knowledge of the spatial structure of errors is essential for an analysis of most GIS operations, with the exception of those that address only simple locations. For example, although the confusion matrix is a common method for describing uncertainty in area class maps or classified remotely sensed scenes, it contains no information on the spatial structure of errors. Thus it is not possible to use it as the basis for an error model, and to make estimates of the uncertainty present in simple derivative GIS products such as the area of a given class. Similarly, the spatial interpolation procedure known as Kriging (Isaaks and Srivastava, 1989) provides estimates of both interpolated value and uncertainty at points. But Englund (1993) shows in a well-conceived example that such point uncertainty estimates are of misleading value in determining the uncertainty in a simple site suitability analysis using GIS.

Following Heuvelink, Burrough, and Stein (1989), the general approach to error propagation in GIS operations is as follows:

1. Define an error model for each type or layer of information required in the analysis.

2. Depending on the operation to be performed and the nature of the error model, adopt one of the following three strategies:

    a. find an algebraic expression for the uncertainty in the result of the operation (limited to a few simple operations, and must be defined and coded in advance);

    b. find an approximation to uncertainty using a Taylor series expansion (less limited, but again must be defined and coded in advance);

    c. generate a suitable number of realizations of the error model for each input layer, perform the analysis repeatedly, each time using a different realization of each input layer, and compute uncertainty by measuring variation across the products from each analysis (the most robust method, computationally intensive but applicable for any conditions).

Although the example of error propagation in area measurement was deliberately chosen as a case of serious underestimation of the effects of uncertainty, our experience with this approach indicates that the opposite is also sometimes true (Hunter and Goodchild, 1994; Hunter, Goodchild, and Robey, 1994). In fact, the principal objective of making explicit uncertainty estimates in GIS analysis is not to show that uncertainty is always a problem, but to identify those cases where it is a problem, and to distinguish them from cases where it is not. At the same time, the approach draws attention to the need for error models, and appropriate information on which they can be based. At this point, we know far too little about the error structures of our most popular databases, including DEMs.

Visualization

Thus far, the emphasis has been on the effects of uncertainty on the products of GIS analysis. But many GIS applications are far less sophisticated, being concerned with little more than the production of a map. Until the end of the last century, map-makers frequently made use of techniques for showing lack of knowledge or uncertainty, leaving areas blank, or showing mythical features. But today we expect our maps to be correct, believing the surface of the Earth to be well understood and well mapped. It is only since the advent of GIS that we have begun to realize how inappropriate this assumption can be, particularly when information of limited accuracy is placed in an almost infinitely precise machine.

There are three possible approaches to communicating uncertainty in map displays (Leung, Goodchild, and Lin, 1992):

1. Ignore uncertainty, showing the observed data as if they were the only possible data.

2. As (1), but include descriptions of uncertainty through the use of appropriate techniques.

3. Using an error model, show multiple realizations, as possible and equally likely versions of the truth.

The first approach is the one that currently dominates practice. We show contours on topographic maps with constant widths that bear no relation to actual positional uncertainty. We show maximum likelihood classifications of remotely sensed scenes, ignoring the less likely classes. Boundaries are shown on soil maps with constant widths that bear no relation to field knowledge about the widths of transition zones. While this approach may be comforting, in its ability to present the world as simpler and clearer than it really is, it is clearly misleading to the GIS user.

The second approach is analogous to the case of a simple measurement, such as the poll example used earlier. Descriptions of uncertainty are often included in the legends of soil maps. Boundaries can be blurred to indicate positional uncertainty, or colors of adjacent zones blended together. DEMs can be displayed as maps of probability. But all of these approaches suffer from a fundamental problem that occurs with spatial data but not with simple measurements. If a boundary is blurred to indicate uncertainty of position, its blurred form is always less contorted than the original, and so gives the user a false impression of geographic detail. In the example of spatial interpolation by Kriging cited above, the technique gives a correct estimate of mean and variance at every point, but the spatial variability of the map of mean estimates is misleadingly smooth. In general, maps such as these that are useful as descriptions of uncertainty are misleading because they are not themselves possible maps.

This makes the third approach the only fully satisfactory one. In this approach, multiple realizations of the error model are shown, either by animation (Ehlschlaeger and Goodchild, 1994) or in multiple windows. In the poll example, the analogous approach would be to present a sample of possible poll results—42, 46, 44, 42, 43—rather than the mean and standard deviation of the error model. This is redundant in this case because intuition gives us the same information, but in the map case there is a clear practical difference between the two approaches.

The third approach is undoubtedly provocative, and provides grounds for interesting debate. Given a series of possible maps, will users feel free to select the map that best suits their personal agenda? The poll example suggests not–if 43% is the published result, with a margin of error of 3%, we do not see political analysts proclaiming that the result is actually 46%, although they may wish it were. Will unequivocal presentation of the implications of uncertainty destroy faith in geographic data, and the GIS industry with it? The experience of fields that regularly make decisions under uncertainty suggests that it will not, and that decision making under uncertainty is both possible and acceptable.

CONCLUDING REMARKS

What impact will growing concern for uncertainty have on decision making, and the application of GIS? First, the impact will be significant. The popularity of GIS depends at least in part on the ability of the computer to confer authority and precision on results. Many decisions might have had different outcomes if the same analysis had been performed by hand instead of by GIS.

Second, concern for accuracy and effective modeling of uncertainty are inevitable. As long as GIS was in the exclusive hands of one side in arguments over spatial decisions, it was possible to ignore the accuracy issue to some extent. But increasingly GIS is being used by both sides, as it becomes more widely available to government agencies, citizen groups, and companies. It has become too easy to point out errors and inaccuracies in the other side's argument, and a few errors can easily undermine a well constructed case.

Finally, concern for accuracy is long overdue. Uncertainty has long been a recognized component of decision making in other fields, and it is no longer possible to ignore its presence in geographic data.

ACKNOWLEDGMENT

The National Center for Geographic Information and Analysis is supported by the U.S. National Science Foundation, grant SBR 88-10917.

REFERENCES

Chrisman, N.R., and B.S. Yandell, 1988. Effects of point error on area calculations: a statistical model. *Surveying and Mapping* 48: 241-246.

Congalton, R.G., editor, 1994. *Proceedings, International Symposium on the Spatial Accuracy of Natural Resource Data Bases*. Bethesda, MD: ASPRS.

Daratech, Inc., 1994. New GIS market study. *Geodetical Info Magazine* (May) pp. 36-37.

Ehlschlaeger, C.R., and M.F. Goodchild, 1994. Uncertainty in spatial data: defining, visualizing, and managing data errors. *Proceedings, GIS/LIS '94, Phoenix, October 25-27 1994*. Bethesda, MD: ASPRS, pp. 245-252.

Englund, E., 1993. Spatial simulation: environmental applications. In M.F. Goodchild, B.O. Parks, and L.T. Steyaert, editors, *Environmental Modeling with GIS*. New York: Oxford University Press, pp. 432-437.

Fisher, P.F., 1992. First experiments in viewshed uncertainty: simulating fuzzy viewsheds. *Photogrammetric Engineering and Remote Sensing* 58: 345-352.

Goodchild, M.F., 1992. *Research Initiative 1: Accuracy of Spatial Databases, Closing Report*. Santa Barbara, CA: National Center for Geographic Information and Analysis.

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

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

Haining, R.P., and G. Arbia, 1993. Error propagation through map operations. *Technometrics* 35: 293-305.

Heuvelink, G.B.M., 1993. *Error Propagation in Quantitative Spatial Modelling*. University of Utrecht.

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.

Hunter, G.J., editor, 1991. *Proceedings, Symposium on Spatial Database Accuracy*. Melbourne: University of Melbourne, Department of Surveying and Land Information.

Hunter, G.J., and M.F. Goodchild, 1994. Quality assessment of GIS products for resource management. *Proceedings, Resource Technology '94, Melbourne*. University of Melbourne, pp. 56-67.

Hunter, G.J., M.F. Goodchild, and M. Robey, 1994. A toolbox for assessing uncertainty in spatial databases. *Proceedings, AURISA '94, Sydney*.

Isaaks, E.H., and R.M. Srivastava, 1989. *Applied Geostatistics*. New York: Oxford University Press.

Leung, Y., M.F. Goodchild, and C.C. Lin, 1994. Visualization of fuzzy scenes and probability fields. *Proceedings, Fifth International Symposium on Spatial Data Handling, Charleston, SC*, 2: 480-490.

Lunetta, R.S., R.G. Congalton, L.K. Fenstermaker, J.R. Jensen, and others, 1991. Remote sensing and geographic information system data integration: error sources and research issues. *Photogrammetric Engineering and Remote Sensing* 57(6): 677-687.

Ormeling, F., 1992. Brian Harley's influence on modern cartography. *Cartographica* 29(2): 62-65.

Wood, D., 1992. *The Power of Maps*. New York: Guilford.