Academic Pursuits

Uncertainty: The Achilles Heel of GIS?

Michael F. Goodchild
University of California, Santa Barbara

The University Consortium for Geographic Information Science

(UCGIS) is a not-for-profit organization of 47 universities and other research institutions in the United States dedicated to the advancement of geographic information science through improved research, education, and access. For information about UCGIS activities and priorities, please visit their World Wide Web site at URL: http://www.ucgis.org.

***Michael Goodchild** is chair of the Executive Committee of the National Center for Geographic Information and Analysis and a professor of geography at the University of California, Santa Barbara. He is also a member of* Geo Info Systems*'s Editorial Advisory Board. His e-mail address is good@ncgia.ucsb.edu.*

A GIS database is a representation of how things are on the surface of the Earth, using binary digits to create an approximation to real phenomena. GIS users work with the database as a surrogate for the real thing, just as an architect works with drawings and models of a building. But the database is only a representation, and because the real world is almost infinitely complex, it is virtually impossible for its digital representation to be completely faithful. *Uncertainty* has emerged as the preferred term for all that the database does not capture about the real world, or the difference between what the database indicates and what actually exists out there. Many forms of uncertainty in geospatial data have been described, including the following:

Positional inaccuracy is caused by our limited ability to measure locations on the surface of the Earth. For example, the National Map Accuracy Standards specify the allowable difference between the positions shown on a map of a given scale and the positions of corresponding features in the real world. As a crude rule of thumb, a database built from a map will have positional inaccuracies of as much as 0.5 millimeters at the scale of the map. For a database built from 1:24,000 mapping, that means inaccuracies of as much as 12 meters.

Errors are caused by blunders, misinterpretations, misclassifications, and a host of other possibilities. Land cover classifications obtained from remote sensing often show accuracies well below 100 percent when subjected to rigorous assessment by comparing them to ground truth (for a review of this area see Fenstermaker, 1994).

Scale. Generalized databases show general trends by removing local detail, creating uncertainty about what is really there.

Fuzziness. The features and classes depicted in geospatial databases are often incompletely defined, leaving uncertainty about exactly what they indicate in the real world.

Sampling creates a representation from limited data, leaving uncertainty as to what actually exists between the sample points.

An example of the consequences of uncertainty in geospatial data is shown in Figure 1. Depicted are two of the six street centerline databases that are available from agencies and vendors for this part of Santa Barbara County, California. Differences caused by positional inaccuracies clearly exist; in addition, streets are shown in one database but are missing from the other because of disagreement about exactly what constitutes a street. Differences also exist in street names and numbering and in the layout of intersections. This much disagreement or uncertainty about the real street network can make it very difficult for users of different databases to communic[ate] effectively — for example, ov[er] which street is referenced by a given Global Positioning Syst[em] (GPS) coordinate. As a result [it] is easy to imagine 911 scenari[os] involving fire trucks dispatche[d] to the wrong address or even t[he] wrong street. Work on these issues at the University of California, Santa Barbara (UCSB) is being led by Val Noronha; more details are avai[l]able at World Wide Web (Web site URL: http://www.ncgia.ucsb.edu/vital.

Uncertainty in geospatial d[ata] is an ancient problem. It existe[d] long before GIS, back in the days when maps routinely showed blank areas or nonexis[]tent sea monsters. But it has come home to roost in GIS in uncertain terms. If we know there is uncertainty in the inpu[t] to GIS analysis, but fail to ide[nti]fy the impact of that uncertai[nty] on the outputs and instead pre[]sent them as correct, then sur[ely] we can and should be held lia[ble] for the consequences.

RESEARCH DIRECTIONS

Because the uncertainty prob[lem] is so serious, it has been the s[ub]ject of a growing volume of research during the past deca[de] and continues to figure promi[]nently in research agendas. T[he] University Consortium for G[eo]graphic Information Science (UCGIS) identified uncertain[ty] as one of its 10 Research Ch[al]lenges (Web URL: http://ww[w.]ucgis.org); it is also one of si areas funded by the 1997 round of the National Imagery and Mapping Agency (NIMA) University Research Initiative.

First, we need to know how to describe uncertainty, so that data producers can make users aware of the level of uncertainty present. Data quality is a major component of the Spatial Data Transfer Standard (SDTS [or Federal Information Processing Standard 173]), in which it is divided into five fundamental components: positional accuracy, attribute accuracy, logical consistency, lineage, and completeness. This list is expanded further in a publication of the International Cartographic Association (Guptill and Morrison, 1995), which is incidentally one of the best overviews of the field. The Federal Geographic Data Committee's (FGDC) Content Standards for Digital Geospatial Metadata (Web URL: http://www.fgdc.gov) also specify quality as a major characteristic of data and lay out how it is to be described so that potential users can make an informed evaluation.

All of these methods suffer from the same basic problem: it takes an expert to make sense of them. Only a very small number of potential users understand statements such as, "Digitizing errors follow a spatially autoregressive model with parameter 3.15," even though statements like this are the only effective way of specifying the exact nature of uncertainty. As research has progressed, the community has come up with more and more models, all of them useful in describing some specific situation but most of them far beyond the comprehension of the average GIS user, let alone the average citizen. Recently we have been experimenting at UCSB with a radically different approach, in which the data are accompanied not by a description but by a method, or piece of code (this work is being carried out with Ashton Shortridge and Chris Funk, both graduate students at UCSB). By running the code, users see a series of simulations of what the true data might look like, all of them equally possible given what is known about uncertainty. Users can then repeat the analysis with a number of simulations and observe the effects on the output, perhaps summarizing them in terms of confidence bands. (A simple demonstration of this concept is available at Web URL: http://www.ncgia.ucsb.edu/~ashton/demos/propagate.html). The big advantage of this approach is that users need no background in statistics and no understanding of the model that lies behind the simulations.

Other researchers are experimenting with ways of getting the uncertainty message across visually. Positional uncertainty can be shown by blurring features or by making them shake in an animation (for example, see Ehlschlaeger et al., 1997, also available at Web URL:

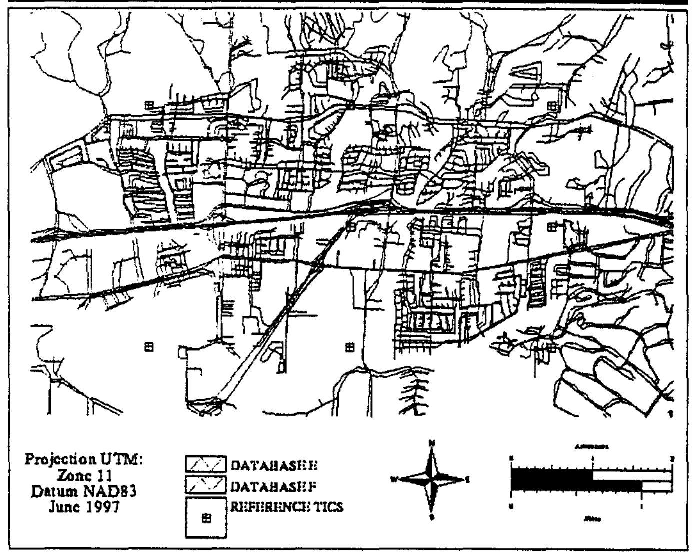

Figure 1. Comparison of two available street centerline databases of part of Santa Barbara County, California, showing differences in feature positions and classifications (see World Wide Web URL: http://www.ncgia.ucsb.edu/vital/research/ncgia.html).

http://www.elsevier.nl:80/homepage/misc/cageo/ehlschl/paper.htm). Uncertainty about the position of a boundary between two classes can be shown by mixing shadings or colors near the boundary. Uncertainty about classification can be shown by fading colors, using the metaphor of the "gray area." There have even been experiments with sound, where quality at a location on a map is communicated by generating a particular tone when users point to it with a cursor. All of these methods seem to work, but only when users are told explicitly that uncertainty is the property being communicated; our natural inclination is to expect maps to say nothing about uncertainty.

Detailed databases contain more data than generalized ones, and normally cost more to produce. So it is common in GIS to be faced with having to use data that are less detailed than one would like. A hydrological model might need a 30-meter digital elevation model (DEM), but might have to be run in an area of the United States where the best data available have a spacing of 3 arc-seconds, or approximately 90 meters. In

From a technical perspective, great progress has been made. On the institutional side, however, the story is much less positive. Uncertainty issues are often swept under the rug by politicians and decision makers who demand "a number, just give me a number."

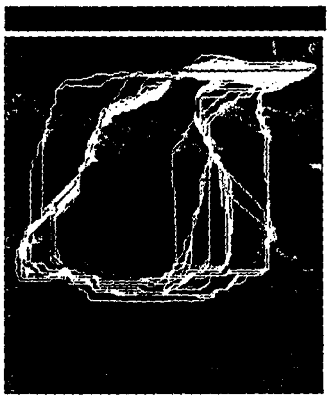

Figure 2. Optimal path for a new road across the Santa Ynez Mountains. The blue line shows the optimal path based on an available 90-meter digital elevation model (DEM); the yellow lines simulate possible optimal paths based on an ideal but nonexistent 30-meter DEM.

these circumstances it is very useful to know how much accuracy is lost in the model. If the result of using 90-meter data is *x*, how different would the result have been if 30-meter data had been available? This may sound like an impossible task, but a number of methods have been developed in the past few years that make it feasible. Essentially, what one does is to run a simulation model that generates a sample of possible 30-meter data sets, consistent with the 90-meter data, and consistent with the general characteristics of 30-meter data obtained from areas where such data are available, preferably as close by as possible.

Figure 2 shows the idea, as implemented by Chuck Ehlschlaeger (now at Hunter College, City University of New York) and Ashton Shortridge. The figure shows a DEM of part of the Santa Ynez Mountains in Santa Barbara County. A route is to be found for a new road from a point in the lower left to one in the upper right, minimizing a function of elevation and slope. The project requires 30-meter data, but only 90-meter data are available. The figure shows the 90-meter solution in blue, and a series of simulations of how different the solution might have been had 30-meter data existed. To do this, essential properties of 30-meter data were obtained from neighboring areas where both types are available. (More information about this experiment can be found at Web URL: http://www.ncgia.ucsb.edu/~ashton/demos/chuck95/stochastic.html.)

With a growing abundance of geospatial data, coming across situations where more than one data set is available to meet a particular need is increasingly common. Figure 1 illustrates this, showing two of the six available street centerline databases for this suburban area. In these circumstances, it is increasingly likely that users will want to see some kind of average, or combination of more than one source. The general term for this is *conflation*, or the combining of elements from more than one source to create a best possible database. Much more research is needed in this area before GIS users will be able to access standard ways of conflating data sets.

THE STATE OF THE ART

Uncertainty research has come a long way in the past decade. Users who want to know something about the quality of the data they receive are now much better served, through standards like SDTS and the FGDC metadata standard, and their international and defense equivalents. Many models have been described in the literature, and made available in software. From a technical perspective, great progress has been made.

On the institutional side, however, the story is much less positive. U

swept under the rug by politicians and decision makers who demand "a number, just give me a number." Commercial software and data vendors often argue that techniques for dealing with uncertainty have no demand in the marketplace or confuse what is otherwise a bullish enthusiasm for the technology. Uncertainty is an essential part of much analysis of risk, as well as many models of decision making, but it has not been addressed to the same degree in GIS. Missing perhaps are the war stories of litigation and disaster resulting from failure to address uncertainty; GIS users have not yet been hit by errors, uncertainties, and lawsuits in places that really hurt. There is no doubt, however, that these will come, and that the potential for disaster is every bit as real in GIS as it is in any other area of human endeavor.

REFERENCES

Ehlschlaeger, C.R., A.M. Shortridge, and M.F. Goodchild. 1997. Visualizing Spatial Data Uncertainty Using Animation. *Computers and Geosciences* 23(4): 387–395.

Fenstermaker, L.K. (ed.). 1994. *Remote Sensing Accuracy Assessment: A Compendium.* Bethesda, Md.: American Society for Photogrammetry and Remote Sensing.

Guptill, S.C., and J.L. Morrison (eds.). 1995. *Elements of Spatial Data Quality.* New York, N.Y.: Elsevier. ■