CHAPTER 14

Encapsulating Simulation Models with Geospatial Data Sets

M.F. Goodchild, A.M. Shortridge, and P. Fohl

INTRODUCTION

Differences exist between real world phenomena and their digital portrayal in geospatial data sets. There is general agreement that these differences must be described and reported along with the data, so that users can make informed decisions on the fitness of the data for specific applications. There is "a strong need...to obtain detailed understanding of how errors propagate through the large number of possible combinations of model types, data types, data sources, and kinds of error, and to make this available to users in an easily accessible form" (Burrough et al., 1996). Some have argued that indeed, in the absence of metadata accuracy reports, spatial data are virtually useless (Smith et al., 1996).

Accuracy reporting typically consists of summary statistics derived from ground measurements upon a subsample of the data. For land cover maps derived from remotely sensed images, this might be the percent correctly classified for each category (Lunetta et al., 1991). For a digital elevation model, the statistic might be a root mean square error (RMSE) for a set of locations at which the true elevation is known (Shearer, 1990). These sorts of global measures of uncertainty are inadequate by themselves for analysis of uncertainty, since they provide no information about spatial structure. Indeed, map and data accuracy standards in general are not sufficient to characterize the spatial structure of uncertainty (Goodchild, 1995; Unwin, 1995).

The current paradigm holds that data producers are responsible for providing such (often inadequate) summary statistics with their data, and that data users are responsible for translating these statistics into meaningful estimates for the suitability of these data for their applications. Just what users are expected to make of these summary reports is unclear; how, for example, does a forester use RMSE to decide whether a particular elevation data set is suitable for fire tower site selection? Openshaw (1989) described general simulation approaches to modeling uncertainty in spatial data for geography, and the past decade has seen considerable progress. This research supports the notion that the general simulation and error propagation method is a complete characterization of uncertainty in spatial data and its effects on analysis. However, these methods remain both theoretically and technically challenging to implement for most spatial data users.

This chapter describes a new paradigm for both data producers and data users. Under this paradigm, data producers replace current accuracy information with an "uncertainty button" in metadata. The button ties an appropriate simulation method to the data quality report. In essence, the button becomes the accuracy metadata; the method replaces the measure. Data users adopt a new view of spatial data; instead of employing the original dataset for an application, they will use one or more realizations to produce a distribution of potential outcomes. By studying this distribution, users gain an understanding of how uncertainty in the data affects their application. Enhanced spatial operations in GIS will facilitate this approach to data handling, and provide more sensitive methods for understanding what is known about the real-world phenomenon modeled by the data.

The following section of the chapter reviews the current metadata accuracy reporting method, and describes the simulation/propagation approach for characterizing uncertainty. We suggest that this be substituted for traditional metadata reporting in the form of an uncertainty "button." The third section introduces various data examples to illustrate the proposed approach. The chapter concludes with a discussion of prospects and challenges for this framework.

METADATA CHARACTERIZATION OF SPATIAL DATA UNCERTAINTY

Much spatial data production, particularly that of federal government agencies like the USGS, is now impacted by a range of metadata specifications. The objective of the development of these specifications is to enhance the sharing of spatial information, to encourage consistency in data generation and use, and to reduce redundancy in data compilation (SDTS, 1996). Government agencies engaged in spatial data production subject their data to accuracy assessments, typically disqualifying any that fail to meet quality specifications and reporting summary information from the assessments in metadata reports for data users. These reports summarize the quality of the data as it relates to some predefined specification. As an example, consider a USGS level 2 digital elevation model (DEM). An approved DEM file must have an RMSE of less than one-half of the source contour interval, with no error exceeding one contour interval (USGS, 1995). With regard to these reports, then, producers are primarily concerned that their data meets a somewhat abstract measure of accuracy.

In contrast, data users are normally not interested in the accuracy of the data set itself, but rather in the spatial phenomenon that the data set imperfectly represents. They need to know how imperfect this representation is, as it relates to their applications. Consider a forester who wishes to use a USGS DEM to help identify promising sites for a new fire tower. The forester has calculated the size of the viewshed for a set of locations, and is interested in determining how closely the calculated viewshed matches the actual viewshed at these sites. That the RMSE for the quadrangle does not exceed 7 meters is not a detail which the forester can easily use to determine the quality of the viewshed calculations.

Indeed, analytically deriving the uncertainty of spatial attributes is frequently difficult or impossible.

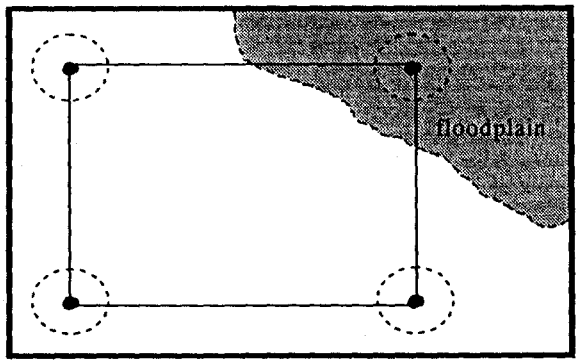

Figure 14.1. Land parcel; positional uncertainty of corner points indicated by circles.

Consider the relatively simple case presented in Figure 14.1. The area of the rectangular land parcel is defined by four corner points. According to the survey, these points are arrayed in a square one hundred meters on a side. However, the surveyed points are subject to positional uncertainty; this uncertainty is characterized by a Gaussian distribution with a mean of zero and a standard deviation of 10 meters, as depicted by the dashed circles. The application question is, what is the standard error associated with the area of the land parcel, given the positional uncertainty information?

In fact, this can only be calculated directly from the available information with some difficulty (Griffith, 1989). However, the standard error may be estimated more simply through a Monte Carlo simulation procedure, which would proceed as follows (and as illustrated in Figure 14.2). Positional error is simulated for each corner using a distribution meeting the criteria specified above. The resulting quadrilateral is a potential realization of the actual parcel. The area of this quadrilateral is calculated and stored. Then, positional error is simulated again, and the area is again noted. This process is repeated a large number of times. For each realization, uncertainty in position of the corners is propagated to variation in parcel area. By analyzing the resulting distribution of area measurements, one can estimate the standard error and characterize the variation in area due to the positional uncertainty of the corner points. Figure 14.3 portrays a histogram of areal estimates derived from 100,000 simulations. The simulation method is general, in the sense that uncertainty can be propagated to answer other questions as well. For the parcel example, the following questions might be of interest and could be answered: what is

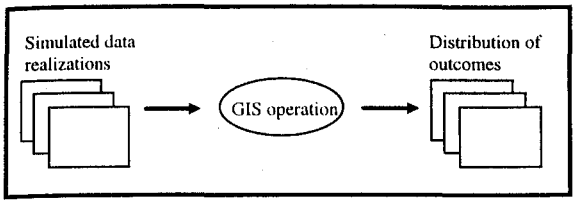

Figure 14.2. Error propagation approach.

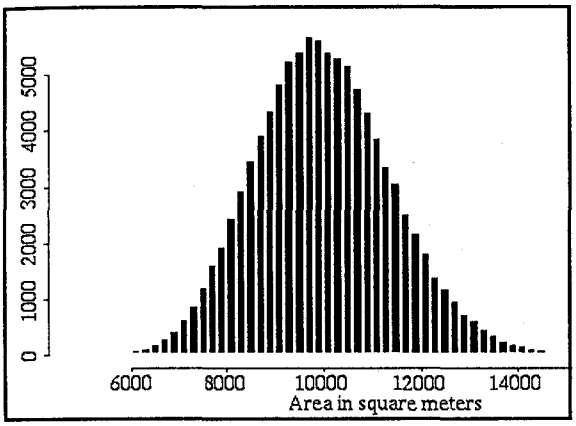

Figure 14.3. Distribution of parcel areas from 100,000 simulations. Mean is 10,000.4, standard deviation is 1427.

the chance that the parcel area is smaller than 9,000 square meters? How likely is it that more than 10% of the parcel is actually in the floodplain? A working prototype for this example is available on-line at:

http://www.ncgia.ucsb.edu/~ashton/demos/propagate.html

These examples—the dilemma of the forester, as well as the parcel area puzzle—illustrate that traditional metadata summaries generally fall short of providing adequate measures of spatial data uncertainty to the user. The general Monte Carlo propagation approach demonstrated above, coupled with data-specific uncertainty simulation algorithms, appears to be the most adequate way of expressing what is known about some spatial phenomenon by combining the data collected about the phenomenon with relevant data quality information (Heuvelink et al., 1989; Fisher, 1991; Lee et al., 1992; Englund, 1993; Ehlschlaeger et al., 1997). For characterizing the uncertainty due to imperfect spatial data in many applications, the user requires a set of equally probable simulations of the spatial phenomenon rather than an incomplete set of summary statistics and a data set known to be in error. In a sense, the simulations themselves become the uncertainty metadata, since the user can see the variation between them, as well as the distribution of application results across the realizations.

We propose that the responsibility for providing these simulations—as with metadata in general—rests with the data provider, not the user, since the provider has much more information concerning the quality of the data and is more equipped to perform an accuracy assessment sensitive to measuring spatial patterns of error. Additionally, simulation theory and techniques are challenging topics for most spatial data users, whose areas of expertise lie more typically with the phenomenon the data represent. The data producer can bridge this knowledge gap for the user community by encapsulating an appropriate simulation method within the metadata accompanying the spatial data set. At the U.S. federal level, at least, the mandate for this exists; data quality specification documents emphasize the responsibility of USGS data producers to "report what data quality information is known," so that users can make informed decisions about the applicability of the data for their applications (USGS, 1996).

What would such a metadata record look like? Figure 14.4 shows what might appear on the computer monitor when a user is electronically browsing a spatial data library. An "uncertainty button," following the GIS error handlers of Openshaw (1989), replaces the usual statistic or table. A short simulation algorithm replaces a line or two of text, or a number, in the record. When the user presses this uncertainty button, the specified number of simulations are generated using the producer-specified uncertainty model and simulation algorithm. These simulations are then processed by the user's GIS and a distribution of results is returned.

GIS operation functionality must be enhanced to effectively incorporate this information about uncertainty from the many data realizations. It is obvious that the main difference in computation is that the same operation must be performed n times, where n is the number of realizations. A somewhat more difficult step is deciding what the operation should return. Table 14.1 presents three very typical results from a GIS operation in the left-hand column. The central column suggests what the results might be from a compound operation, performed upon a set of realizations. The final column provides an example of each type of op-

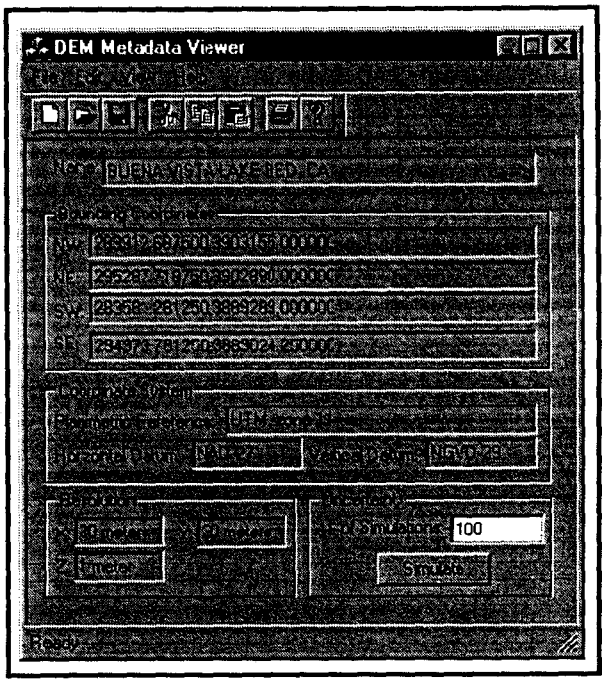

Figure 14.4. Example metadata browser window. Clicking the "Simulate" button generates DEM realizations for uncertainty propagation.

Table 14.1. Output from Different GIS Operations upon Uncertain Spatial Data.

Traditional Result	Result Incorporating Uncertainty Information	Example
Number	Mean, Standard Deviation	Query: Polygon Area
Object	Probability Field	Calculate: Buffer
Surface	Animation Frames	Generate: Cost Surface

eration. The extension of GIS functionality necessary for the implementation of uncertainty propagation is beyond the scope of the present work, but it is certainly realizable.

ILLUSTRATIVE EXAMPLES OF THE APPROACH

The previous section introduced a concept for incorporating uncertainty simulation directly into the metadata associated with an individual spatial data file. Three examples are presented here to indicate how this general method could be implemented. Point, line, and surface area data models are represented in these examples, to demonstrate the breadth of spatial data types that are amenable to this approach.

Scattered Point Data

The first example is a map of tree locations in a forest. Studies of seed dispersal for this species of tree indicate that the maximum range of dispersal from any individual is 100 meters. The data consist simply of the coordinate locations for each tree. Spatial uncertainty in this data is limited to the positional uncertainty of these coordinates. For simplicity, we assume that the organization producing these data has determined the positional uncertainty to be isotropic, with

a Gaussian distribution centered on each observed location and a specified standard deviation. Uncertainty is independent for each point. An application question for these data is, given the uncertainty in the tree locations, how likely is it that all tree locations are within 100 meters of one another?

Coastline Data

The second data set is a vector coastline for a section of the central California coast near Point Conception. The data set itself is a "mean line," an average of several coinciding coastline data sets. Of potential importance is the notion that the mean line is not itself a potential coastline, due to the smoothing effect of the averaging (Goodchild et al., 1995). Many applications may require coastlines of statistically realistic texture rather than the smoother mean. The generation of such coastlines from the mean line requires simulation of removed variation, and is complicated by the prospect that this variation is spatially autocorrelated. The simulation model proposed in Goodchild et al. (1995) uses a distance decay exponent and a range parameter to characterize spatial autocorrelation of the variation about the coastline. Their model is considerably more complex to implement than the preceding point data model. However, the byte size of the algorithm code itself is not large, and could easily be transmitted with the data. An application question of interest is, how long is this stretch of coastline? A mean length (which is not equivalent to the length of the mean coastline), and standard deviation is returned.

DEM Data

The third example uses digital elevation data. The data set is a subset of the USGS one-degree DEM, Los Angeles-west. Studies comparing these data to higher resolution and accuracy collocated 7.5' USGS DEMs have developed measures of mean and variance of the difference, and of the spatial structure of this difference (Ehlschlaeger et al., 1997). An uncertainty model has been developed that uses this information to produce realizations of the difference surface. Each difference surface realization is then added to the one-degree DEM, creating a statistically probable simulation of the "actual" 7.5' DEM-quality surface. If 7.5' DEM data are adequate for a particular application, then this model creates a set of potential realizations of adequate surface representations. This is an unconditional simulation, meaning that no locations on the DEM are necessarily spared from perturbation. An application question for these data is, what is the expected cost of a least-cost path traversing the terrain, where cost is a function of path length, path steepness, and elevation range of the path?

In each of the three cases, the uncertainty model/simulation algorithm is encapsulated in the metadata in the form of a simulation button. Users pressing this button generate a series of simulated data sets. Through the error propagation approach, the questions proposed in these illustrations, and many others, can be answered. These answers come with confidence intervals or other measures of reliability, providing a more realistic depiction of the effects of data uncertainty on the application question.

DISCUSSION

Taken together, the three data examples are representative of much of spatial data. We chose two different object data models and a field data model. The simulation models chosen are also representative. While the first, operating on the point data, was spatially independent for each location, the remaining two simulation models directly accounted for spatial dependence in the error field. The approach advocated in this chapter is very general and extendable to any spatial data set that can be stored in a computer and can be assessed for its fidelity to the phenomenon it represents.

Several critical issues present themselves. The first is the choice of spatial uncertainty model. A growing body of research on spatial uncertainty modeling indicates the diversity of approaches, methods, and results. In the face of such diversity, how is a data producer to choose the most acceptable model? On the other hand, how is the resource manager, the ecologist, or the environmental engineer to choose? These users undoubtedly lack expert knowledge about both the data collection methods employed by the data producer and the spatial simulation model theory and implementation in vogue with spatial information scientists. By working with uncertainty modelers, data producers are in the best position to decide upon the most effective simulation approaches for specific spatial data sets. Data users can have increased confidence both in the uncertainty simulation models and in the data itself. Research on simulation model efficacy must be done to enable data producers to make informed decisions about which models to use, and to indicate

needed changes to accuracy assessments to accommodate model requirements.

A second research topic concerns the distribution of computer processing for simulations. Data will be stored and queried in digital libraries. However, when the user wishes to check the uncertainty of the data, and "clicks the button," what should actually happen? One possibility is that the library maintains a large number of stored realizations for each data file. This seems unwieldy, particularly in light of the continuing rapid increase of processor speed. Instead, realizations could be generated on the fly. Where should the generation occur—at the library site or on the user's machine? From a computational perspective, it might make sense for the processing to occur on the user's machine. In this case, users would download the data file, bundled with an executable simulation routine, and generate simulations locally.

Geographic information systems algorithms require some modification under this paradigm, since they must work on multiple realizations and return meaningful, clear results. Table 14.1 identifies some relatively straightforward outputs of traditional operators and their "uncertainty-enhanced" counterparts. Research topics remain; for example, how will this method fare in compound spatial analysis, in which a large number of input data layers are combined using numerous spatial operations? How can the output of one GIS function easily be used as the input to another? How can the contribution of uncertainty from different spatial sources be easily quantified and expressed to the user? Significant representation issues arise as well. How can information about uncertainty best be communicated and understood? Which, if any, spatial models are especially resistant to effective characterization and communication of uncertainty?

Traditional metadata accuracy reports must change. Those who use spatial data increasingly demand to know how reliable their GIS results are, and standard accuracy statistics are inadequate to supply answers. Simulation-based uncertainty models have been developed for spatial data, but they remain difficult to understand and utilize for most end users. We have argued that the producer, not the user, should be responsible for providing adequate measures of spatial data uncertainty; by adequate, we mean encapsulating the simulation algorithm with the data set. This approach was demonstrated on three representative illustrations. While many challenges remain, we believe that this chapter has introduced and demonstrated a viable, general solution for adequately reporting spatial data uncertainty.

REFERENCES

Burrough, P.A., R. van Rijn, and M. Rikken. Spatial Data Quality and Error Analysis Issues: GIS Functions and Environmental Modeling, in *GIS and Environmental Modeling: Progress and Research Issues,* M. Goodchild et al., Eds., GIS World Books, Fort Collins, CO, 1996, pp. 29–34.

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

Englund, E.J. Spatial Simulation: Environmental Applications, in *Environmental Modeling with GIS,* M.F. Goodchild, B.O. Parks, and L.T. Steyaert, Eds., Oxford Press, New York, 1993, pp. 432–437.

Fisher, P.F. First Experiments in Viewshed Uncertainty: The Accuracy of the Viewshed Area. *Photogrammetric Eng. Remote Sensing,* 57(10), pp. 1321–1327, 1991.

Goodchild, M.F. Attribute Accuracy, in *Elements of Spatial Data Quality,* S.C. Guptill, and J.L. Morrison, Eds., Elsevier, London, 1995, pp. 59–79.

Goodchild, M.F., T.J. Cova, and C.R. Ehlschlaeger. Mean Objects: Extending the Concept of Central Tendency to Complex Spatial Objects in GIS, in *Proceedings, GIS/LIS '95,* ASPRS/ACSM, Nashville, TN, 1995, pp. 354–364.

Griffith, D.A. Distance Calculations and Errors in Geographic Databases, in *Accuracy in Spatial Databases,* M.F. Goodchild and S. Gopal, Eds., Taylor & Francis, London, 1989, pp. 81–90.

Heuvelink, G.B., P.A. Burrough, and A. Stein. Propagation of Errors in Spatial Modelling with GIS. *Int. J. Geogr. Inf. Syst.,* 3(4), pp. 303–322, 1989.

Lee, J., P.K. Snyder, and P.F. Fisher. Modeling the Effect of Data Errors on Feature Extraction from Digital Elevation Models. *Photogrammetric Eng. Remote Sensing,* 58(10), pp. 1461–1467, 1992.

Lunetta, R.S., R.G. Congalton, L.K. Fenstermaker, J.R. Jensen, K.C. McGwire, and L.R. Tinney. Remote Sensing and Geographic Information System Data Integration: Error Sources and Research Issues. *Photogrammetric Eng. Remote Sensing,* 57(6), pp. 677–687, 1991.

Openshaw, S. Learning to Live with Errors in Spatial Databases, in *Accuracy in Spatial Databases,* M.F. Goodchild and S. Gopal, Eds., Taylor & Francis, London, 1989, pp. 263–276.

SDTS Task Force. The Spatial Data Transfer Standard: Guide for Technical Managers. U.S. Dept. Interior, 1996, <ftp://sdts.er.usgs.gov/pub/sdts/articles/pdf/mgrs.pdf>

Shearer, J.W. Accuracy of Digital Terrain Models, in *Terrain Modelling in Surveying and Civil Engineering*, G. Petrie and T.J.M. Kennie, Eds., Thomas Telford, London, 1990, pp. 315–336.

Smith, T.R., D. Andresen, L. Carver, R. Dolin et al. A Digital Library for Geographically Referenced Materials. *Computer*, 29(7), pp. 54, 1996.

Unwin, D.J. Geographical Information Systems and the Problem of 'Error and Uncertainty.' *Prog. Human Geogr.*, 19(4), pp. 549–558, 1995.

USGS. DEM/SDTS Transfers, in *The SDTS Mapping of DEM Elements*, U.S. Dept. Interior, 1996, <ftp://sdts.er.usgs.gov/pub/sdts/datasets/raster/dem/demmap3.ps>

USGS, National Mapping Program Technical Instructions, Standards for Digital Elevation Models, U.S. Dept. Interior, 1995.