Communicating Uncertainty for Global Data Sets

Ashton M. Shortridge & Michael F. Goodchild
National Center for Geographic Information & Analysis, and
Department of Geography
University of California, Santa Barbara, CA, 93106-4060, USA
Email: ashton@geog.ucsb.edu & good@geog.ucsb.edu

Abstract

This paper reviews the modeling approach to characterize spatial uncertainty and examines the prospects and challenges for developing statistically based uncertainty models for comprehensive global datasets like the Shuttle Radar Topography Mission (SRTM). Under this paradigm, data become integral inputs to models which characterize phenomena, rather than the primary representation of phenomena. One challenge is the identification of high quality, well distributed, high accuracy data necessary for developing the uncertainty model. A second challenge is distribution of the extensive processing for the simulation. A final challenge is producer and user acceptance of a new way of looking at the relationship between data and real-world spatial phenomena. How can geographic information science facilitate such a paradigm shift?

1. Introduction

The notion that many social and environmental phenomena are global in scope has gained general acceptance in the last few decades. Today, the general public is aware that individual human impacts on local scales are collectively contributing to global environmental change. To understand global scale phenomena and their importance for humans and the environment, scientists and policy makers require accurate information about the state of the planet. Until very recently however, extant spatial datasets were insufficient for global characterization of the most basic environmental factors.

This decade has seen the wide release of several important datasets with global coverage of satisfactory spatial resolution, and more will be available in the next 3-5 years. Several of these datasets are discussed in this paper's second section, including the Shuttle Radar Topography Mission (SRTM) which will produce a digital elevation model (DEM) at three arc second resolution for much of the land surface of the planet. They are already being used for a wide variety of applications: global process modeling input parameters, navigation in northern Russian waters, continental scale cartography linework, archeology in southern Africa, and as base map data for national GIS projects in developing nations (Smith & Langaas, 1995). On the one hand, the wide array of valuable uses highlights the utility of such datasets. On the other hand, this breadth poses a serious challenge for geographic information science: how can the quality of globally extensive datasets best be characterized for so many potential uses; how can the reliability of application results that employ them be assessed?

A relatively extensive body of research is concerned with characterizing and modeling spatial data uncertainty. The underlying theory, discussed in the third section of this paper, offers robust and general approaches which have been applied experimentally at local and regional scales. This paper covers three fundamental challenges for employing these approaches for global scale spatial data sets. Although these challenges are generally applicable for any phenomenon measurable at any point on the planet, the discussion will focus on elevation. First is the necessity for very high quality "ground truth" information at point locations scattered across the globe. How can well-distributed arrays of such points be developed? In the fourth section we discuss the prospects for using existing global datasets of spot elevations for characterizing uncertainty in global high resolution DEMs.

The second challenge is the distribution of the processing required for the current approaches to characterizing uncertainty. These approaches involve the development of an uncertainty model, the generation (through Monte Carlo simulation) of alternative realizations of the phenomenon, and the propagation of uncertainty through a GIS operation. Simulation can be done by the producer on

servers, either pre-processed and stored in advance or generated "on the fly" as requests are received. Alternatively, processing can become the task of the data user, if appropriate models and algorithms are made available with the data. These three approaches to data distribution are covered in the fifth section.

A third challenge is producer and user acceptance of a new way of looking at the relationship between data and real-world spatial phenomena. Data become integral inputs to models which characterize phenomena, rather than the primary representation of phenomena. The paper concludes with some thoughts on this substantial paradigm shift.

2. Some global terrain dataset accuracy specifications and some problems

The Digital Chart of the World (DCW) consists of vector layers representing a variety of political, human, and natural features on the land surface of the globe. Of particular interest for this paper are the elevation contours and spot heights (ISCGM, 1996). This dataset was compiled from the Operational Navigation Charts (ONC), a 1:1,000,000 paper map series compiled by the US Defense Mapping Agency and others. The ONCs were developed over a period spanning several decades to support aircraft navigation worldwide. Accuracy of contours and spot elevations is divided into horizontal (positional) and vertical components in the specifications. The horizontal accuracy of contours and spot heights derived from the ONCs is "2,040 meters, rounded to the nearest 5 meters at 90% circular error" (ISCGM, 1996). The absolute vertical accuracy statement for contour elevation data is 650 meters at the 90% confidence level, though empirical tests indicate the data is much more accurate, at about 160 meters (USGS 1997). For spot elevations the vertical accuracy is 30 meters (ISCGM, 1996).

The second global dataset containing elevation information is GTOPO30, a 30 arc second (roughly 1 kilometer) resolution raster DEM. A variety of elevation sources were used to compile GTOPO30, including 3 arc second data wherever possible and DCW contours and spot heights. DCW vector hydrography was employed to convert the hypsography to a raster format. Accuracy at any location is related to the data source for that location, and therefore varies widely across the dataset, from an estimated 9 meter RMSE in New Zealand to an RMSE of 304 meters in Peru. GTOPO30 documentation attempts to link the DCW contour accuracy specification to an RMSE; by assuming that error is normally distributed with a mean of zero, an RMSE of 97 meters is estimated (USGS, 1997).

A third source is the forthcoming global DEM derived from the Shuttle Radar Topography Mission (SRTM). This mission is scheduled for launch in September 1999; following the 11 day mission to collect interferometric radar data, processing will take one year. The result will be a publicly available 3 arc second (very roughly 100 meter) resolution raster DEM for the earth's land surface between 60 degrees north and about 60 degrees south. (NASA, 1999a).

The accuracy specification for this product is that of DTED level 2: absolute vertical accuracy (90% Linear Error) is 16 meters (NASA, 1999b). However, actual accuracy of the product will be determined and reported during verification. The elevation data will be released with estimates of random and systematic error for each height posting. This will be the most comprehensive amount of accuracy information ever released with a DEM series.

Accuracy information presented above for elevation data on DCW and GTOPO30 is really quality specification; such accuracy specifications reflect the important requirement of maintaining consistent quality standards in the map production process. However, they are insufficient for identifying a dataset's fitness for uses other than those originally intended. This problem is compounded for global datasets, for which the potential range of applications goes far beyond what data producers can foresee. For example, elevation contours might be adequate for topographic feature recognition at 35,000 feet (an original intent of the ONC map) but be entirely unsuitable for modeling drainage basins.

A second problem with these specifications is that they do not describe elevation error sufficiently. Three assumptions are implicit in both the DCW and the GTOPO30 accuracy statements: that error does not vary across the region in any systematic way (so a single global measure is acceptable), that error is unbiased (and probably normally distributed), and that elevation errors are independent (errors at nearby locations are uncorrelated). Experimental research with many elevation datasets has shown that these assumptions are typically invalid even for fairly small regions (Ehlschlaeger et al., 1997; Fisher, 1991). For global datasets, they can certainly be dismissed.

3. Uncertainty modeling and propagation

Uncertainty modeling approaches the problem of accuracy specification from a much different perspective than the production oriented statements discussed in the previous section. Discrepancies between data and actual terrain are more fully characterized by uncertainty models, which explains their application to many environmental datasets including DEMs (Heuvelink et al., 1989; Fisher, 1991; Lee et al., 1992; Englund, 1993; Ehlschlaeger et al., 1997). Conceptually, this modeling approach treats a surface (here, an elevation surface or an elevation error surface) as a realization of a random function. The random function is comprised of a set of random variables with spatial locations whose dependence on each other is specified probabilistically. The distribution for each random variable is estimated from nearby points for which values are known. Proper estimation of the spatial structure of the surface is critical, since these models use distance and direction from known locations to identify the distribution of elevation or elevation error (see Goovaerts, 1997; Isaaks & Srivastava, 1989, for extensive discussion on the geostatistical approach).

Elevation may be known at a set of points within the study region; for example, a global positioning system (GPS) could have been used to sample several dozen locations within a study area. Conditional methods ensure that the surface model passes through these locations, "honoring" the ground truth data (Goovaerts, 1997). In other cases, true elevations may not be available within the area of interest, but error characteristics are known. Perhaps they are derived from data quality specifications, though these would have to include not only aspatial characteristics like RMSE, but also measures for the spatial structure of error and any correlations between error and slope, or error and absolute elevation. Alternatively, they may be assumed to match those of nearby regions for which this information is available. Unconditional methods are useful in either circumstance, and are used to build error surfaces which are then added to the elevation data. Such methods are termed 'unconditional' because no elevations are specifically honored; instead, elevations at all locations on the surface are perturbed (see for examples Ehlschlaeger et al., 1997; Hunter & Goodchild, 1997).

Regardless of form, an uncertainty model may be employed to characterize uncertainty in a spatial dataset. Various forms of kriging use the uncertainty model to develop a map of the most likely distribution of the phenomenon, given the available information. Kriging also produces a map of the variance, which provides some notion of the uncertainty in the estimate at each location. For many applications, a better approach may be to propagate data uncertainty through the analysis to identify its impact upon the results of the application. This is accomplished by producing, via Monte Carlo simulation, a set of equiprobable realizations of the environmental phenomenon. The application is then run upon all realizations, producing a distribution of results. The general model of propagation is presented in Figure 1.

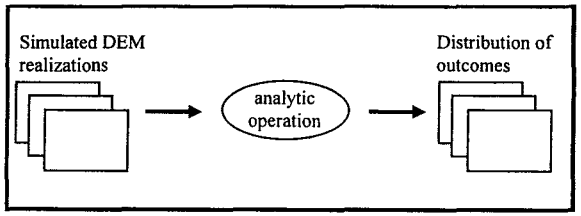

Figure 1. Propagation of uncertainty through a spatial analysis using DEMs.

Geographic information systems require some modification under this paradigm, since they must work on multiple realizations and return meaningful, clear results. Some of these modifications are technological and computer-based; GIS must accommodate multiple realizations of input data and produce different forms of output (Openshaw, 1989, Shortridge, 1999). Others are conceptual and human-based; input becomes a set of realizations, phenomena are stochastic, model output numbers become distributions and maps become animations. Table 1 identifies some relatively straightforward outputs of traditional operators and their uncertainty-enhanced 'counterparts.

Table 1. Output of 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

While the modeling/propagation framework presented here is well developed theoretically and practically for regional datasets, its application to global dataset uncertainty evaluation has not been closely examined. In the next section, the problem of identifying suitable ground truth measures will be explored, focusing particularly on elevation.

4. Global ground truth

High quality, well distributed sample data are required to develop good conditional uncertainty models for spatial datasets like elevation. Ideally, these points should be randomly distributed across the region. The sample should statistically match the population; if elevation values were only sampled from valley floors, lack of information about peaks and ridges will bias the statistics. Many locations should be near other locations so that good models of spatial autocorrelation can be developed. The spatial distribution of the points should also ensure that no large region is unreported. For local or even regional study areas in some parts of the world, such data exist or might conceivably be collected by the researcher.

For datasets of global extent, however, it is clear that the potential for sample data is limited to existing spot height data or future international surveys. We now consider an existing source to highlight some challenges for obtaining global ground truth.

An easily obtainable set of global spot elevations is available in DCW. This dataset does not appear to be accurate enough for use with the 3 arc second SRTM DEM. The vertical RMSE is 30 meters, and the positional accuracy is not good. The accuracy might be adequate for use modeling GTOPO30 accuracy, though in many areas GTOPO was created using these points. A second problem is that DCW spot height locations are not randomly distributed, since many are peaks. For a 30 arc second (1 km) global DEM an ideal set of spot elevations would include many pairs within a few kilometers of one another so that short-range correlation could be adequately modeled. DCW spot height locations are rarely so close. Additionally, the ideal density of the points for even a 30 arc second DEM would be relatively high - something like 100 points per one degree square, or very approximately 1 point for every 120 km². The DCW hypsography spot heights are quite sparse regardless of where in the world one is. For example, the ratios of km²/# spot heights for Belgium, Senegal, and Mongolia are 1900 km²/pt, 1770 km²/pt, and 1150km²/pt, respectively.

An alternative approach to conditional simulation is to develop a model using data in one part of the world and apply it to another part of the world for which elevation and elevation error are believed to have similar properties. For example, adequate high quality spot heights might exist for developing a robust uncertainty model for terrain in the Rocky Mountains of North America. This model could then be applied to similar montane regions in parts of the world where ground truth data are sparse. This approach has the advantage of making the uncertainty modeling approach feasible

with currently available high quality spot height information. Existing data could be employed to rapidly develop models, even in the absence of quality data in some parts of the world. The most significant disadvantage is the necessary assumption that error properties are consistent between regions.

Even if a very high accuracy, global, well distributed set of spot heights were available, problems remain. Chief among these is identifying the comparability of these spot heights to the elevation measures in a global DEM. In global elevation data, the precise meaning of the values is often not well defined. Consider the GTOPO30 definition for each posting. In areas derived from 3 arc second data, the elevation is "representative" of the 100 3 arc second postings within each 30 arc second region. In areas derived from DCW contours, the final elevation is interpolated from vector data. In either case, the elevation is a generalized average (not a well-defined average) that can not be directly evaluated with the real-world terrain it purports to represent. The most obvious approach to evaluation is to compare the GTOPO30 value for a particular location with a high quality measurement for that location and to treat the difference as error, although this is not strictly correct, since the underlying spatial data model is different. DEMs are certainly not the only type of data with this fundamental measurement problem; land cover data sets are particularly subject to it as well. Further research is warranted to identify how such data can adequately be compared with reality, and how their uncertainty might best be modeled.

Once an uncertainty model is developed, the problem arises as to the distribution of simulation processing. The next section covers this issue.

5. Distribution of Processing

Developing a feasible approach for the distribution of computer processing of simulations is critical. It is clear that global data will be stored in and retrieved from digital libraries, and that application processing will take place on users' computers. There are several possibilities for when and where simulation could occur:

- the library could store a large number of realizations of the global data set.
- the library could generate realizations on the fly as they are requested by users.
- the library could transmit the model with the data; users would generate realizations locally.

Under the first two options, the user would request some number of realizations. Realizations would be either retrieved or generated and transmitted via Internet, tape, or CD to the user. The user would then employ these realizations in the propagation schema presented in Figure 1. The relative advantage of one option over the other relates to the tradeoff between storage space and processor speed. Option One entails orders of magnitude more storage than the global dataset itself, while Option Two requires large amounts of processing by the library. Both options require that large quantities of data be transmitted. Consider a user request for 100 realizations of a 5x5 degree tile of SRTM 3 arc second data. In total this request returns 3.6 billion elevations; one only hopes the user is not downloading over a 56K modem!

Therefore the third option seems preferable. In this option the DEM is bundled with an executable simulation routine and any other data and transmitted to the user. Realizations are generated locally and processed through the application model as in Figure 1. While the processing required to develop realizations is substantial, it is no longer practically insurmountable given current technology. Tens of realizations of sizeable datasets (>20,000 points) can be produced using public domain geostatistics software on Pentium II-powered personal computers in under a minute.

A second advantage to this option is that users could be enabled to develop their own uncertainty models from the data and model parameters provided. For example, suppose a user was especially interested in a one degree portion of a 5x5 degree SRTM tile. If the user obtained the "ground truth" data for this block (say, 6000 random spot heights), the user could construct a different variogram model than that developed for the whole tile, and the resulting uncertainty model might be preferable for the application.

6. Conclusion

For the approach advocated in this paper to be accepted, producers and users of global data must agree that it is both practical to bring about and that the implementation pain inflicted is offset by the knowledge gain afforded. We argue that the approach holds significant advantages for both producers and users.

Data producers are in the best position to decide upon the most effective uncertainty models for global spatial data sets, and to allocate resources to implement these models. Their responsibility to do so is clear. United States government data producers, for example, have a mandate 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). Truly achieving this responsibility using traditional data accuracy statements is not possible, while the approach described here is general, meeting the mandate regardless of the multitude of possible uses of global data.

Most spatial data users find spatial statistical theory and techniques challenging topics, since their areas of expertise lie more typically with the phenomena the data represent. The data producer can bridge this knowledge gap for the user community by developing a standard modeling method within the metadata accompanying the spatial data set. The adoption of the modeling/propagation approach requires changes in the way processing is performed and results are analyzed, but these changes will result in increased confidence in the models, the data, and the application results.

Perhaps it is human nature to treat the world as if it is made up of discrete elements, that phenomena can be described crisply with precise numbers. Consider: the Indian city of Bangalore had 4,130,288 people in 1991; the average wind speed in Chicago, USA is 16.7 kph; Asia's mean elevation is 910 meters (Rand McNally, 1997). But, measurement is often not so accurate or precise. More fundamentally, spatial data collection requires that abstracted representations of the earth be developed, immediately creating mismatches between environmental phenomena and the data about them. Under such circumstances, scientists have long accepted that often the best results are stochastic, that estimates without probabilities are less useful. For decision making, the utility of confidence intervals has also been recognized. Here we have argued that the probabilistic view can be usefully extended to modeling spatial phenomena like elevation, and that these more realistic characterizations of the earth lead to more realistic results for projects and models requiring global spatial information.

Acknowledgments

This work was financially supported by the National Center for Geographic Information and Analysis and the National Imagery and Mapping Agency.

References

Ehlschlaeger, C. R., A. M. Shortridge, and M. F. Goodchild (1997). Visualizing spatial data uncertainty using animation. *Computers & Geosciences*, 23(4): 387-395.

Englund, E. J. (1993). Spatial simulation: environmental applications. In Goodchild, M. F., B. O. Parks, and L. T. Steyaert, (Eds.). *Environmental modeling with GIS*. Oxford Press, New York: 432-437.

Fisher, P. F. (1991). Modelling soil map-unit inclusions by Monte Carlo simulation. *Int. J. Geographical Information Systems*, 5(2): 193-208.

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*. Oxford University Press, New York, 483 pages.

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

Hunter, G. J., and M. F. Goodchild (1997). Modeling the uncertainty of slope and aspect estimates derived from spatial databases. *Geographical Analysis*, 29(1): 35-49.

International Steering Committee for Global Mapping (1996). Report of the second meeting of ISCGM, Santa Barbara, USA. 76 pages.

Isaaks, E. H., and R. M. Srivastava (1989). *An Introduction to Applied Geostatistics*. Oxford University Press, New York, 561 pages.

Langaas, S. (1995). Cartographical data and data quality issues. In *Selected papers from the DCW data quality project*. DCW & Data Quality Project Report No.1 /1995, UNEP/GRID-Arendal, Norway, 5-13.

Lee, J., P. K. Snyder, and P. F. Fisher (1992). Modeling the effect of data errors on feature extraction from digital elevation models. *Photogrammetric Engineering & Remote Sensing*, 58(10): 1461-1467.

National Aeronautics and Space Administration (1999a). Shuttle Radar Topography Mission technical fact sheet. http://www-radar.jpl.nasa.gov/srtm/factsheet_tech.html

National Aeronautics and Space Administration, 1999a. Shuttle Radar Topography Mission Data Products. http://www-radar.jpl.nasa.gov/srtm/dataproducts.html

Openshaw, S. (1989). Learning to live with errors in spatial databases. In Goodchild, M. F. and S. Gopal, (Eds). *Accuracy in Spatial Databases*. Taylor & Francis, London: 263-276.

Rand McNally (1996). *Answer Atlas*. Rand McNally & Co, USA, 176 pages.

Shortridge, A. M. (1999). Representing Spatial Data Uncertainty in a Geographic Information System. http://pollux.geog.ucsb.edu/~ashton/demos/aml/amlsim.html

USGS (1996). DEM/SDTS transfers. In *The SDTS Mapping of DEM Elements*. U.S. Dept. Interior, 35 pages.

USGS (1997). Global 30 Arc Second Elevation Data Set (GTOPO30) Documentation. http://edcwww.cr.usgs.gov/landdaac/gtopo30/gtopo30.html