METHODS: UNCERTAINTY

Michael F. Goodchild, University of California, Santa Barbara

Michael F. Goodchild
Department of Geography
University of California
Santa Barbara, CA 93106-4060
USA

METHODS: UNCERTAINTY

Keywords: representation, error, vagueness, fuzzy set, spatial dependence, Tobler's First Law, geostatistics, probability, confidence limits, geographic information system, Global Positioning System

Glossary

adjustment theory: the theory of errors in survey measurements

frequentist: holding to the view that probability must be defined through experiment, as a proportion of events

fuzzy set: a set defined by membership in a specific class, with membership allowed to be partial

Gaussian (or Normal) distribution: a description of the relative sizes of errors in repeated measurements, a mathematical function resembling a bell

geostatistics: the theory of regionalized variables, or spatial variables subject to spatial dependence

~~metadata: the description or documentation of a data set~~

Monte Carlo simulation: simulation of multiple realizations of a statistical model, for example by tossing a coin

rough set: a set defined by a mix of full and partial membership

subjectivist: holding to the view that probability can be defined subjectively, as a value that determines individual behavior

Taylor series expansion: a form of approximation of a nonlinear function

Tobler's First Law: nearby things are more similar than distant things, a widely observed tendency of geographic information

Synopsis

The representations that constitute the bulk of geographic knowledge must leave their users uncertain to some degree about the nature of the real world. Knowledge of geographic uncertainty is essential in geographic research, in decision making, and in everyday life. A variety of methods have been devised for measuring the degree of uncertainty inherent in geographic information. Models of uncertainty make use of the theoretical constructs of fuzzy and rough sets, spatial statistics, and geostatistics. Much progress has been made in recent years in developing methods for visualizing uncertainty, and for analyzing its propagation into the results of geographic research. Progress has also been made on the description of quality in data documentation, and in the use of data set quality as a dimension of the processes of search and discovery,

Introduction

As humans we experience only a very small fraction of the geographic world. While seasoned travelers may claim to have visited all of the 50 US states, or a large proportion of the world's countries, or all of its continents, even they will have sensed directly only a fraction of the Earth's 500,000,000 sq km of surface. In the temporal domain our experience is even more limited, as none of us can expect to live much longer than 100 of

the Earth's 5,000,000,000 year history. All other knowledge about the geographic world must come to us through representations of various kinds, in the form of text, speech, images, digital databases, or physical models. Moreover such representations must inevitably be generalized, approximated, abstracted, or subject to measurement error. It follows then that virtually all knowledge of the geographic world is subject to uncertainty. No representation can provide more than an approximation to the truth, forcing geographers always to look at the world through a fuzzy, clouded, and distorting lens. The advent of digital technology and powerful geographic information systems has changed this not one bit, though the precision of computing devices often encourages a false sense of accuracy on the part of their users.

It is clearly in the interests of users of representations to understand something of their inherent uncertainty. Researchers are interested in knowing something about the uncertainty of the results of their research; decision-makers should be aware of the uncertainties surrounding their decisions, particularly if they expect to be held legally or administratively accountable; and the general public should be aware of the mistakes that can result from overly precise interpretation of geographic information. Thus research on uncertainty has tended to address four specific issues:

- the measurement of uncertainty, expressed as probabilities, confidence limits, or qualitative judgments;
- the modeling of uncertainty within the theoretical frameworks of fuzzy and rough sets, spatial statistics, and geostatistics;
- the visualization of uncertainty, in order to convey knowledge about uncertainty to the user of information in intuitively straightforward ways; and
- the propagation of uncertainty, tracking the effects of uncertainty as information is processed and manipulated, and expressed as measures of confidence in results.

Humans are inveterate explorers, whether as hunter-gatherers searching for new food sources and reporting their discoveries to the band, or as earlier Europeans setting out to explore and conquer new worlds, or as modern cavers searching for previously unexplored passages. The notion that exploration can be complete, that every needle in the geographic haystack can eventually be found, has surfaced at various times in human history. In the late 19th century, for example, it was widely believed that the few remaining areas of the planet -- the interior of Africa, the Antarctic continent – would eventually be explored and mapped; by the 1960s space was being described as the Final Frontier; and some years later the US Geological Survey completed its national topographic program of mapping the 48 contiguous states at 1:24,000. Of course all of these efforts leave representations that are uncertain; but it is far more compelling to think about exploring the surface of Mars, or finding unexplored cave passages, than to update or improve the detail of existing mapping. Thus the US is currently anticipating a massive program of human exploration of Mars, while allowing its national topographic maps to become increasingly out of date; their inherent uncertainty about the world they purport to represent is increasing rather than decreasing.

The following sections follow the outline above: first, methods for describing and measuring uncertainty; second, theoretical constructs that frame the modeling of uncertainty; third, recent developments in visualizing uncertainty in geographic information; and fourth, methods for the propagation of uncertainty.

Describing and measuring uncertainty

Two methods are often used to express uncertainty in common parlance: probabilities, as in such statements as “there is a 40% chance of rain today”; and confidence limits, as in “there is a 2% margin of error on this opinion-poll result”, or “this thermometer has an accuracy of 0.5 degrees”. Both are grounded in extensive conceptual and theoretical frameworks. Statements of probability may be interpreted in one of two ways: in a frequentist context as the results of real or imagined experiments; and in a subjectivist context as relative expressions of likelihood. For example, a frequentist might interpret a 40% chance of precipitation as “on 2 out of 5 days with atmospheric conditions similar to this it rained”, or as “40% of the area covered by the forecast will experience rain”. A subjectivist, on the other hand, would understand that rain is more likely than if the probability was given as 20%, and less likely than if it was given as 60%.

Confidence limits are associated with measurements, and express the uncertainty resulting from the use of a specific instrument. The theory of errors asserts that under very general conditions, repeated measurements of the same phenomenon using the same instrument will follow a Gaussian distribution defined by a mean and standard deviation. If the mean is different from the true value the measurement is said to be biased, and the root of the mean squared error is known as the standard error. These concepts are routinely applied to the measurement of position using the Global Positioning System (GPS), to the measurement of elevation, and to many other common forms of measurement. The theory of errors also addresses the impact of errors on the results of calculations from measurements.

Several issues make this simple picture inappropriate for geographic information. First, and perhaps most importantly, the equating of uncertainty and error assumes the existence of a truth, since error is defined as the difference between a measurement and its true value. While this may be a reasonable assumption in some domains, it is rarely appropriate in a geographic context, particularly in human geography. Even a task as apparently straightforward as the measurement of latitude and longitude with a GPS unit is fraught, because of widespread variation in the definitions of those quantities, variations in the Earth’s axis, and the effects of tectonic movements and Earth tides. The classification schemes often used in mapping land use or land cover are based on inherently vague definitions, such that two observers, however experienced, cannot be assumed to share precisely the same definitions of urban, suburban, or rural, and are almost certain to produce different mappings of the same area.

Although the early literature on uncertainty in geographic information was dominated by concepts of error and accuracy, the literature of the past two decades has increasingly avoided these terms, preferring uncertainty as the umbrella term, and making frequent reference to imprecision, vagueness, and fuzziness. Traditional approaches to mapping

have emphasized so-called Boolean classification, in which every point on the landscape must be assigned to exactly one soil class, or exactly one land cover class. However more recent work has opened the way partial or fuzzy membership, by assigning degrees of membership to several classes. For example, a point near the boundary between the urban core and the suburbs might be assigned a 40% membership in the urban class and a 60% membership in the suburban class; the boundary between the two classes would no longer be an infinitely thin line, but a zone of transition.

Second, the theory of errors assumes, quite reasonably, that repeated measurements are made independently. While this is a reasonable assumption in measuring temperature with a thermometer, it flies directly in the face of the pervasive property of geographic data described as Tobler's First Law: nearby things are more similar than distant things. Consider, for example, the measurement of a street's location using a GPS unit, and suppose that a vehicle equipped with the GPS is driven along the street, recording location every second. The horizontal accuracy of the GPS is known to be 5m. If each second's measurement is independently subject to an average error of 5m, the result would be a very jagged track, much longer than the true length of the street (Figure 1). In reality errors in GPS measurements persist over extended periods of time, and GPS tracks are automatically smoothed by the unit's own software. As a result, the shape of the track bears an acceptably close resemblance to the street's true shape, and its length is only slightly longer than the true length This property of positively correlated errors is very common in geographic data; if it were not, many of the tasks routinely performed by geographic information systems (GIS) and other technologies would be much less successful than they are. Together with mean and standard error, spatial correlation ranks as one of the most important measures of uncertainty in geographic information.

[Figure 1 about here]

Much effort has been expended in recent years in devising standard ways for describing the uncertainties that are known to be present in geographic data sets. By the early 1990s five dimensions of quality had been identified, and written into standards by various national and international bodies:

- positional accuracy, summarizing the differences between positions recorded in the data and the corresponding true positions on the Earth;
- attribute accuracy, summarizing the differences between recorded properties of geographic features and corresponding true properties;
- logical consistency, a measure of the degree to which the contents of the data set match the rules used to define it (for example, are the boundaries of area features truly closed?);
- currency, a measure of the degree to which the data set is complete and up to date, or the degree to which its contents match the real world at the stated time of validity; and
- lineage, a history of the actions and processes by which the data set was compiled.

The emphasis in these standards is typically on truth in labeling: the statements should simply record what is known, rather than referencing some threshold standard of quality. What is known might be quantitative, but it might also be qualitative and to some degree subjective. Typically such statements will be assembled by the producers of data, who are likely to be staff members of the government agencies that have traditionally dominated the production of geographic data. But the development of powerful tools over the past two decades has meant that increasing quantities of geographic information are now available from local agencies and even individuals, who may be much less likely to spend the time needed to determine and document quality.

These standards for the description of data quality have become an accepted part of metadata, or the data used to describe data sets and to assess their fitness for a particular application. Today many large Web sites provide catalogs of geographic data sets, allowing users to specify requirements, search the catalog, and in many cases to retrieve and use selected data sets in their own work. Data quality plays a very important role in this process, by allowing the searcher to assess whether levels of uncertainty are above or below the appropriate thresholds for a given application. Contemporary tools allow metadata to remain attached to a data set as it is retrieved, providing an effective means of documentation.

Modeling uncertainty

A large number of models of uncertainty in geographic information have been proposed over the past four decades. In essence, these models allow the results of limited analyses to be generalized, as for example when the manufacturer of a thermometer asserts that the accuracy of the instrument is 0.5 degrees, based on a small number of sample measurements. Models also allow for the calculation of related properties, such as the impact of an uncertain slope measurement on predictions of soil loss. Most usefully, models allow for the simulation of uncertainties, through the generation of a number of alternative realizations of the model, each of which might be the truth, but showing a degree of variation that matches the known level of uncertainty. Such simulations have become a powerful basis for visualizing uncertainties (see next section), and for investigating the influence of uncertain data on the results of analysis (see final section).

A simple example is shown in Figure 2. Each of the four corner points of this parcel of land have been surveyed, with an uncertainty that is a small fraction of the length of each side. Uncertainty can be modeled by assuming that both coordinates of each surveyed point are disturbed by an error drawn from a Gaussian distribution. The figure shows two simulated realizations of this model.

[Figure 2 about here]

In surveying, the well-developed body of theory known as adjustment allows practitioners to make use of known levels of measurement error in their instruments, and to assess the accuracy of positions determined from those measurements. Adjustment theory is based on the theory of errors and the Gaussian distribution. More broadly, however, this rigorous approach is problematic for geographic information for the

reasons discussed earlier. Geostatistics, or the theory of regionalized variables, addresses directly the spatial dependence that is endemic to geographic information, and has provided a very powerful basis for modeling uncertainty.

Finally, the theory of fuzzy and rough sets has found an appropriate home in the study of uncertainty in geographic information, particularly in addressing issues related to vaguely defined classifications. The notion of replacing Boolean or hard classification with measurements of membership is very appealing, and has been adopted enthusiastically in several areas. Rough sets provide something of a common ground between Boolean and fuzzy classification, and are also finding applications in this domain. The lack of a rigorous way of dealing with spatial dependence in this framework is a major barrier to greater progress, however.

Visualizing uncertainty

Cartographers have long been concerned with the portrayal of certain specific types of uncertainty on maps. The depiction of mythological beasts on early maps and globes and white areas in the center of Africa are instances of practices that over the years have become less and less common as the geographic world has become better known. Nevertheless dashed lines that still denote uncertainty in some national boundaries (e.g., in the Arabian Peninsula) and in intermittent watercourses.

Today's paper maps and atlases show little regard for the uncertainty that is almost universally present in geographic information. The width of contours on topographic maps reflects the width of the pen that drew them, not uncertainty over their positions. Maps showing classifications almost always employ Boolean methods, and show boundaries between classes as a single pen width. While this might be acceptable as long as printed maps had to be drafted using pens, the digital world is far more flexible in the options it provides for cartographic design. Colors and shading can be continuously graded, line widths can be varied, and assorted methods can be used to convey the uncertainty message.

Over the past two decades researchers have investigated several methods. Maps have been produced with varying line widths and continuous variations of shading and color. Features have been blurred to indicate uncertainty about their positions, and colors have been greyed to indicate uncertainty about classifications or measured attributes. All of these methods are based on assessments of uncertainty for individual features, however, and fail to address the spatial dependence that is almost always present. For example, blurring of points may indicate uncertainty in their positions, but may mislead the user who is interested in knowing the distance between the points, because if both points are misplaced by the same error in the same direction, the distance computed between them will be perfectly accurate.

Addressing this problem has proven particularly difficult. While many models exist that incorporate spatial dependence, including geostatistical models, it makes little sense to attempt to describe uncertainty by specifying the parameters of such models, when few users will a sufficient statistical background to interpret them appropriately. Instead,

researchers have experimented with various forms of animation. In the case of positional error, for example, animation allows the features to be shown dancing either independently or in a correlated fashion, clearly demonstrating the difference between absolute and relative positional error, and its implications for analysis.

Propagating uncertainty

As tools for processing geographic information have become more sophisticated, it has become possible to evaluate the extent to which the results of analysis are affected by known uncertainties in the underlying data. What, for example, is the impact of known errors in census population counts on forecasts of future school populations, or analyses of social deprivation in neighborhoods? Such questions are particularly fraught when they involve changes of scale, or changes in the zones used to report statistics and to conduct analysis.

Significant research progress has been made in our understanding of the propagation of uncertainty, using tools ranging from differential calculus to Taylor-series expansion to Monte Carlo simulation, and powerful packages are now available to analyze the effects of uncertainty on many standard manipulations of geographic information. Some progress has also been made on automating the modification of metadata as data sets are manipulated, in order to create data quality statements for the output of operations.

This work on uncertainty propagation has led to some useful conclusions. First, many of the operations performed on geographic information are essentially nonlinear in their responses, leading to surprising large impacts from comparatively modest uncertainties. Missing a key feature in a geographic data set, for example, can sometimes have dramatic consequences. Second, in other instances the impacts of uncertainty are remarkably low, a point that has already been made in the context of Tobler's First Law and the estimation of such simple properties as distance or shape – the distance between two features may be estimated very accurately despite large uncertainties about the features' positions.

Conclusion

Uncertainty is in some ways the Achilles heel of modern geographic information processing. Despite the apparent precision of computing systems, and the use for example of double precision (15 significant digits) in the representation of many aspects of the geographic world, the actual accuracy of geographic information and the products of processing may be alarmingly low. Several instances have surfaced in recent years of the potentially disastrous and expensive consequences of uncertainty, ranging from the severing of a cable in the Italian Alps by an aircraft (the cable was not marked on the maps available to the pilot) to disputes over land ownership that result from poor-quality surveying. Some of these have resulted in expensive court settlements, when it became apparent that the users of geographic information technologies had not taken adequate note of the known uncertainties in their data.

Despite extensive research over the past two decades, this message still has not filtered through into the design of maps and GIS, and comparatively simple steps to reveal and evaluate issues of uncertainty have not been taken. It would be simple, for example, to

provide estimates of the uncertainties associated with measurements of area, when these are made from representations with known levels of positional error. In reality, uncertainty is a messy concept that decision makers would often rather sweep under the carpet, and its successful analysis often requires an understanding of sophisticated concepts of probability and correlation. Nevertheless, any knowledge of uncertainty, however crude, is better than none, and the consequences of ignoring it can be severe.

Further reading

Guptill, S. C. and Morrison, J. L. (1995). *Elements of spatial data quality.* Oxford: Elsevier.

Heuvelink, G. B. M. (1998). *Error propagation in environmental modelling with GIS.* Bristol, PA: Taylor and Francis.

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

MacEachren, A. M. (1995). *How maps work: representation, visualization, and design.* New York: Guilford.

Monmonier, M. S. (1996). *How to lie with maps*. Chicago: University of Chicago Press.

Shi, W., Fisher, P. F., and Goodchild, M. F. (eds.) (2002). *Spatial data quality*. New York: Taylor and Francis.

Zhang, J. X. and Goodchild, M. F. (2002). *Uncertainty in geographical information*. New York: Taylor and Francis.

Figure captions

1. An illustration of the effects of independent errors in the representation of a complex geographic feature such as a road.

2. Two realizations of a model of uncertainty in the four surveyed corner points of a simple land parcel.