GENERALIZATION, UNCERTAINTY, AND ERROR MODELING{PRIVATE }

Michael F. Goodchild

National Center for Geographic Information and Analysis, and
Department of Geography
University of California
Santa Barbara, CA 93106-4060, USA
+1 805 893 8049; FAX +1 805 893 7095; good@ncgia.ucsb.edu

Abstract

I argue in this paper that users of GIS expect the contents of a database to follow principles of scientific measurement, and that these conflict in several ways with the objectives of cartographic generalization. The recent literature on uncertainty and error modeling in spatial databases has proposed measures and models of error, defined in this context as the difference between an observation and its true value. Since true value must be interpreted in GIS in the context of the user's understanding of the meaning of the observation, incomplete specification must be included as a potential source of error. Generalization has two distinct roles in GIS, although they are traditionally merged in cartography: as a set of rules for generating displays, and as a component of data specification. The discussion focuses on the latter, and reaches a series of conclusions about the value of concepts of cartographic generalization in digital databases, notably with regard to metric scale.

INTRODUCTION

In GIS, the juxtaposition of issues of generalization with those of data quality seems to bring out a fundamental tension. In this paper I hope to position the discussion that follows towards one end of that tension, reflecting in part my own personal views, while at the same time exploring the degree to which accommodation is possible. While one might argue that the tension has existed for a very long time, the advent of GIS, and in particular the expectations of its users about the contents of GIS databases and the products of GIS analyses, seem to force us to confront it explicitly.

The approach that forms the basis of this paper might be termed "scientific measurement", although many other terms might be appropriate, some of them possibly more so. The primary tenet of this approach might be expressed as the concept of generality—that certain constructs have the same meaning to all of us, and can therefore form the basis of reliable communication and sharing of information. For example, while the term "hot" has some value and is used in everyday speech, its meaning is not sufficiently stable, independent of context, and generally agreed to allow it to be used for reliable communication. Instead, we have devised the construct of Celsius temperature, based on certain clear and agreed context-independent principles, including the freezing and boiling points of water at sea level. It may be of momentary interest to two people that one considers today "hot" and the other "cool", but it is of much greater interest for many different purposes that the temperature is 22 degrees Celsius. Obviously I am ignoring here the possiblity that one can address the question "What do people mean by 'hot'?" on a scientific basis.

Scientific measurements can have precision, which is generally defined (though not always) as the degree of detail in the reporting of a measurement, and accuracy, which is defined as the expected difference between an observation and its true value. Inaccuracy can result from the use of a biased instrument, or from carelessness, or misreporting—all cases where a different value might have been obtained using "best practice".

The arguments in this paper are based on the assertion that users of GIS expect the contents of its database to be a collection of scientific measurements, subject to agreed definitions, best practice observation methods, and replicable between observers. These are the assumptions of scientific measurement, but they are also deeply ingrained in our legal system, and in the entire set of societal arrangements and practices within which GIS use is embedded. Of course such practices are constantly under attack from many directions (Pickles, 1995), but they remain largely in place, and I believe that the assertion holds true of the vast majority of GIS practice at this time.

DATA QUALITY AND DATA SEMANTICS

Even though the contents of a geographic database may be understood within the context of scientific measurement, their accuracy is not perfect for a number of reasons. First, the instruments used for measurement of position on the surface of the Earth have limited accuracy—even the best instruments for measuring angle, for example, are limited to one part in 10^4 or 10^5. Second, the models used to represent the Earth are imperfect. For example, the figure of the Earth used in modern geodetic datums is only an approximation to the true geoid; and a differentiable mathematical function can only approximate the true shape of local terrain. Third, the real world is infinitely complex, and digital representations must be simplified and approximated to fit within the limited capacity of digital stores. Finally, a form of inaccuracy results when a measurement is not fully defined. For example, I might measure the slope at a point as 25%, using a definition and method known only to me. You might visit the same point and measure the slope as 30% using a different definition of slope. While we might agree to the extent that we both use a similar method of slope measurement based on a regular grid and the fitting of a plane to a 3x3 neighborhood centered on the point, we introduce a form of inaccuracy by failing to include the sampling interval in the definition or semantics of the measurement.

The fractal literature contains numerous instances of measurements of geographic phenomena that depend in similar ways on the linear dimensions of the measuring instrument (Mandelbrot, 1982, and see also Maling, 1989). Other geographic variables depend on linear dimensions that are often unspecified, or uncontrolled. For example, the population density at a point can be defined only if population is counted in some area around the point, and divided by the size of the area. The definition of the variable depends on the area used—a population density surface based on scanning a circle of radius 200m is clearly not the same as one based on a radius of 1km. Yet databases of population density are rarely annotated with this aspect of their semantics, and indeed are often obtained by making calculations of density over variably-sized subdivisions of space, such as counties, giving the analyst only a very indirect control over this property of the data.

The standard scientific response to this fourth problem is to enrich the semantics—to make as many significant aspects of the definition of a measurement explicit. For example, one might claim in the context of terrain that there is no such thing as slope, but only slope at a particular sampling interval using a particular method of calculation. In this sense one should always specify the linear dimensions of the measuring instrument, the sampling interval, the method used to calculate slope from a DEM, and any other aspects of the specification that might otherwise be uncertain to the user. In the best of all possible GIS worlds these properties would be always available to the user, stored in the form of metadata perhaps. But in reality, with the increasing separation that is now common between creator and user of data as a result of increasing sophistication of digital communication technologies, lack of explicit semantics is often a major source of inaccuracy. Moreover, semantics may be lost along the communication channel, during format conversion, or due to a failure to document on the part of the vendor of proprietary software. How many GIS vendors, for example, document the full technical details of their method for calculation of slope from a DEM? Too often the details are assumed to be part of a shared set of conventions, or to be part of a technical world that is of no interest or beyond the comprehension of the user.

STATISTICAL MODELING OF ACCURACY

From a statistical perspective, the term "error" includes any difference between a measurement and its true value, however caused. Following the previous discussion, failure to specify the complete definition of a measurement can appear to the user as error. Thus we should generalize the definition of accuracy somewhat to make it specific to the user's semantics—accuracy depends on the meaning given to the measurement by the user, and is a measure of the difference between the measurement and the truth defined according to that meaning. Thus perfect geographic data is data that is perfectly consistent with a specification (Salgé, 1995).

Accuracy is an expression of average or expected difference, rather than the magnitude of any given instance. For example, an inaccurate measuring instrument might happen to give the correct value by chance—its inaccuracy would become apparent only over a series of observations. A metric of inaccuracy, such as the mean absolute error, or root mean square error, is termed an error descriptor, and might be a characteristic of a measuring instrument, or a measuring method, or an individual observer.

The effects of error in measurements on the results of analysis are the subject of the field of error analysis (Taylor, 1982). To understand such error propagation effects, it is necessary to have some kind of error model to emulate the distribution of errors that might occur in practice. The best-known error model is the Gaussian or normal distribution. But conventional error analysis is based on the assumption that each observation is independently subject to errors. In GIS, the individual

elements of a database are rarely independent observations—in constructing a DEM, for example, any connection between the eventual grid and the original measurements of control points and positions on stereo pairs is lost in the complex processing that occurs during photogrammetry. So error modeling in GIS is forced to treat an entire data set as a single observation, from a population of possible data sets that might have been observed. Thus Goodchild, Sun, and Yang (1992) and others define a GIS error model as a stochastic process capable of generating a population of GIS data sets, the differences between them being indistinguishable from differences due to errors of observation, definition, processing, etc. Each simulated data set is termed a realization of the error model. For example, Hunter, Caetano, and Goodchild (1995) discuss an error model for DEMs that is consistent with published USGS data quality statements. Given a suitable error model, it is possible to simulate a sample of equally possible realizations, and by repeating the same analysis on each of them to estimate the effects of error on the outcome of any GIS analysis.

CARTOGRAPHIC GENERALIZATION

McMaster and Shea (1988, p. 242) define cartographic generalization as "the application of both spatial and attribute transformations in order to maintain *clarity*, with *appropriate* content, at a given *scale*, for a chosen *map* purpose and intended *audience*" (emphasis mine). The connection between generalization, defined in this way, and the earlier discussion of this paper is problematic, for several reasons. First, there has been no previous reference in this paper to maps. Maps can be defined as a form of display of spatial data in which position is represented in analog form on a two-dimensional medium, such as paper or a computer screen. As such, they have two very distinct roles in GIS: first, as a medium for storage of spatial information as it makes its way from observation into digital form; and second, as a method for display of the contents of a spatial database. While they are clearly effective and virtually indispensible in the second case, there are many examples of spatial data sets that pass from observation to database without being displayed as maps.

From the perspective of scientific measurement and the first half of this paper, this definition of generalization has elements of subjectivity, notably in the suggestion that content might depend on the audience, and in the word "appropriate". The term "scale" is also problematic. Metric scale is defined as a ratio of distance on a map to distance on the ground, the former being well-defined in the definition of map given earlier. But it clearly has no meaning for digital data, where there are no distances to be measured (distances between bits on the hard drive?). Because maps must have fixed scale (ignoring the effects of projection), they provide strictly limited space for the clear portrayal of information. Thus in areas where information is inherently rich it is necessary to thin, approximate, or generalize in the interests of clarity. Lines on maps must be shown with certain minimum widths, because they must either be drawn with pens, or be wide enough to be visible. Clearly, then, the scale of the map and a range of other factors determine the types and amounts of information that can be shown, and it is not surprising that metric scale is the primary factor in determining how maps are drawn and what they show.

Although these rules of mapmaking are still relevant to the display of digital spatial data, such displays are inherently transitory, easily changed by zooming and redrawing, and readily edited. While the generalization decisions of the mapmaker are likely to impact the value of a printed map for many years, similar decisions in digital display will change as a function of the display device, user preferences, the selection of objects for display, and a host of other transitory factors. In this context it makes no sense for generalization to constrain the contents of a database; rather, generalization issues should only come into play on the fly in the generation of map views of the database.

On the other hand, the level of geographic detail is clearly an important and relatively immutable property of any digital spatial database. It determines the likely volume of the data, and thus affects the costs of storage and processing. It determines the quality of decisions that can be made with the data, and its range of uses. But there are many contenders as suitable metrics of the level of geographic detail in a database: the sample interval of a DEM, the cell size of a raster image, the size of the smallest recorded feature. Attempts to use metric scale as a universal indicator of level of geographic detail for both maps and digital databases have led to numerous problems and conflicts, since the term is only well-defined for maps. For example, a digital orthophoto quad is said to have a "scale" of 1:12,000 because its positional accuracy (about 6m) is consistent with the National Map Accuracy Standards for maps at that metric scale; a scanned digital air photo is said to have a "scale" of 1:35,000 because that was the metric scale on the focal plane of the camera; a digital soil map is said to have a "scale" of 1:15,840 because it was digitized from a paper map of that metric scale. The term has come to mask a complex series of conventions that have little meaning except within a small community of experts.

TOWARDS AN ACCOMMODATION

The literature is rich with reports of efforts to automate processes of generalization so they can be implemented on the fly in the generation of displays of spatial databases (Buttenfield and McMaster, 1991; Müller, Lagrange, and Weibel, 1995). More problematic is the more limited literature that has attempted to formalize the specification of generalized geographic data in ways that are consistent with the concepts of scientific measurement discussed earlier, and has compared such methods for vector and raster representations (McMaster and Monmonier, 1989; Monmonier and McMaster, 1990; Schylberg, 1993). In this section I propose that this latter issue can be better approached within the framework of the field/discrete object dichotomy (Goodchild, 1992), which is arguably closer to our conceptualization of geographic variation, and hopefully more productive.

Fields

A field is defined as a function z over the spatial variables x and y. Geography can be conceived as a collection of fields, each describing the variation of one variable, such as elevation or rainfall, over the surface of the Earth. For a variable z measured on an interval or ratio scale, a convenient error descriptor is the mean absolute error, $E|z-z^*|$, or the expected difference between the observed value z and the true value z^* at any point. For nominal or ordinal variables, the equivalent is the probability $p(z,z^*)$ that a point observed to be of class z is actually of class z^*.

To observe a field in practice it is invariably necessary to observe over some extended area, parametrized by a linear measure, as discussed earlier. The role of such a linear measure as a measure of geographic detail is well understood in Fourier analysis, where it is related to wavelength, and in geostatistics. Convolution with a filter can change the effective value of the linear measure by smoothing the field. Commonly used terms for such characteristic linear measures of fields are spatial resolution, and scale.

Raster representations of fields provide an explicit linear measure of geographic detail in the cell size; and regular grids do the same in the sampling interval. More problematic are the vector representations of fields—TINs, polygons, digitized contours, and irregular point samples—where there is no single linear measure that is characteristic of the representation's level of geographic detail. In the case if irregular point samples the level of geographic detail may be controlled by the algorithm used to interpolate a continuous field; or it may be inferred from the distances between sample points. In the case of polygons, it might be argued that level of geographic detail is implicit in the size of the minimum mapping unit, or smallest recorded polygon. In general, however, there is no clear, objective connection between the dimensions of the objects used to represent the field and the level of geographic detail available to the user, and this remains a powerful argument against such representations of fields.

Discrete Objects

In this alternative view, geographic phenomena are conceived as collections of discrete point, line, or area objects, potentially overlapping, and embedded in an otherwise empty space. Discrete object representations have much richer semantics, since it is much more difficult to establish widely shared rules about what constitutes a mountain, or a city, than their field equivalents—a surface of elevation, or population density.

Level of geographic detail is similarly important, but is now embedded in sets of specification rules. We must decide, for example, what constitutes a lake at different levels of geographic detail. Statement such as "Manitoba has 100,000 lakes" clearly require a specification of level of geographic detail to qualify as scientific observations.

In the cartographic tradition, metric scale has understandably become the primary means of specifying level of geographic detail in feature specifications. But the numerical value of metric scale is of no particular significance—specifications could just as easily refer to "high", "medium", or "low" detail as to "1:24,000", "1:100,000", and "1:250,000", provided these terms were used consistently throughout the set of specification rules. As noted earlier, significant uncertainty about the meaning of data results if the level of geographic detail is not made explicit.

It is easy to understand why metric scale became the measure of choice in map specification rules. But if all that is required to drive specification rules for discrete objects is some ordinal scale of geographic detail, then there are no grounds for maintaining metric scale as the preferred measure for digital databases, and good grounds not to if the measure has no well-defined meaning in this context.

CONCLUSIONS

At the outset, I indentified what seemed to be a tension between user expectations of scientific measurement on the one hand, and the traditions of fields like cartographic generalization on the other. As in many other areas of human activity, the technological transition to digital spatial databases has led to numerous problems and conflicts, as traditional ideas either succeed or fail in making themselves relevant to the new context. The discussion seems to lead to five distinct conclusions:

1. The expectation of GIS users that spatial databases will contain collections of scientific measurements raises problems. It suggests the need to measure accuracy, in order to characterize the expected difference between database contents and the specifications understood by the user. It also conflicts with aspects of the cartographic tradition of generalization.

2. Echoes of the cartographic tradition of generalization appear in two very distinct contexts in GIS: in the generation of displays of the database, and as contributions to its specification. In the former case, the momentary and context-specific nature of GIS display argue for generalization almost entirely on the fly. In the latter case, and given the user's expectations, it is essential that specifications be as explicit as possible.

3. Metric scale is not appropriate as a specification parameter for digital databases. Efforts to make metric scale a universal property of spatial data, invariant under analog/digital conversion, should be abandoned in favor of more appropriate linear measures of level of geographic detail.

4. For representations of fields, linear measures are available and explicit for raster and grid data, but not for other (vector) representations. There is a need to agree on appropriate equivalents that are related to the linear dimensions of the primitive objects of the representation.

5. For geographic phenomena represented as discrete objects, it is possible to write the complex rules of feature specification using any ordinal scale of geographic detail. But lack of explicit specification of level of geographic detail in such rules contributes significantly to inaccuracy.

Acknowledgments

The National Center for Geographic Information and Analysis and the Alexandria Digital Library are supported by the National Science Foundation.

References

Buttenfield, B.P., and R.B. McMaster, editors, 1991. *Map Generalization: Making Rules for Knowledge Representation.* Harlow: Longman Scientific and Technical.

Goodchild, M.F., 1992. Geographical data modeling. *Computers and Geosciences* 18(4): 401-408.

Goodchild, M.F., Sun Guoqing, and Yang Shiren, 1992. Development and test of an error model for categorical data. *International Journal of Geographical Information Systems* 6(2): 87-104.

Hunter, G.J., M. Caetano, and M.F. Goodchild, 1995. A methodology for reporting uncertainty in spatial database products. *Journal of the Urban and Regional Information Systems Association* 7(2): 11-21.

Maling, D.H., 1989. *Measurements from Maps: Principles and Methods of Cartometry*. New York: Pergamon.

Mandelbrot, B.B., 1982. *The Fractal Geometry of Nature*. San Francisco: Freeman.

McMaster, R.B., and M. Monmonier, 1989. A conceptual framework for quantitative and qualitative raster-mode generalization. *Proceedings, GIS/LIS '89, Orlando, FL*, pp. 390-403.

McMaster, R.B., and K.S. Shea, 1988. Cartographic generalization in a digital environment: a framework for implementation in a geographic information system. *Proceedings, GIS/LIS '88, San Antonio, TX*, Vol. 1, pp. 240-249.

Monmonier, M., and R.B. McMaster, 1990. The sequential effects of geometric operators in cartographic line generalization.

International Yearbook of Cartography 30: 93-108.

Müller, J.C., J.-P. Lagrange, and R. Weibel, 1995. *GIS and Generalization: Methodology and Practice*. London: Taylor and Francis.

Pickles, J., 1995. *Ground Truth: The Social Implications of Geographic Information Systems*. New York: Guilford.

Salgé, F., 1995. Semantic accuracy. In S.C. Guptill and J.L. Morrison, editors, *Elements of Spatial Data Quality*. Oxford: Elsevier.

Taylor, J.R., 1982. *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. Mill Valley, CA: University Science Books.

Proceedings, GIS/LIS, 96, Denver. November 19-21 pp 765-774. 1996