SHARING IMPERFECT DATA[1]

Michael F. Goodchild

National Center for Geographic Information and Analysis

University of California

Santa Barbara, CA 93106-4060

ABSTRACT

Quality is almost always an issue in working with spatial data, since almost all spatial data are of limited accuracy. Access to appropriate information on quality is essential if agencies or scientists are to share spatial data, since no user is willing to trust data of unknown accuracy. The concept of metadata provides an effective means of ensuring such access, by integrating information on quality with the data themselves. Standards for the description of quality and for metadata are emerging, and procedures for quality assurance and quality control of GIS data are under development.

INTRODUCTION

The theme of this paper is that data quality is a significant hindrance to the sharing of spatial data between institutions. At the simplest level, we argue that data that meet the quality standards and needs of one agency will frequently fail to meet the different, or more exacting needs of others. But at a more subtle level, the difficulties inherent in documenting data quality, in passing that information to the consumer of data, and in making it general enough to apply in a wide range of applications, in themselves form a much more problematic and challenging barrier. Not only must agencies sharing data trust eachother's quality control procedures, but they must also be able to communicate knowledge of data quality - sources of error, types of imperfection, measures of accuracy - in an objective and useful way. We argue that the lack of suitable methods for formulating and communicating information on quality currently forms a significant impediment to sharing, particularly given the awkward quality problems inherent in spatial data.

---

[1] Paper prepared for the specialist meeting of NCGIA Initiative 9: Institutions Sharing Spatial Information, San Diego, February 26-29, 1992.

The first part of the paper introduces the data quality problem, and explains its dimensions as it applies to spatial data. The next section looks at the problem of communication, through concepts of metadata and lineage, and at why these are frequently poorly developed in the spatial data handling field. This is followed by a section on current developments in quality assurance and quality control (QA/QC). The last section examines possible courses of action that may help to remove the data quality impediment to greater sharing.

THE DATA QUALITY PROBLEM

Observational data are almost always subject to error, but spatial data seem to suffer more from imperfect quality than do other kinds of data. In part this is due to the fact that spatial data can at best only approximate the complexity of real geographic distributions, through processes of generalization, abstraction or aggregation. It may also be due to the fact that much geographical data is the result not of precise measurement but of partly subjective interpretation - two soil scientists are unlikely to make identical soil maps of the same area. Finally, spatial data may be unusually imperfect because of the measurement errors inherent in many instruments, and the simplifying assumptions about such things as the figure of the earth that are implicit in their use. The next five sections describe some of the common sources of error in spatial data. Veregin has published a bibliography and associated taxonomy (Veregin, 1989a,b), and reviews of many aspects of the spatial data quality problem can be found in Goodchild and Gopal (1989).

Sources of error

Measurement

The instruments that produce raw geographic data often have well known accuracies. GPS, for example, provides positional coordinates with levels of accuracy that are a function of the experimental design (one receiver or two), and can be degraded through the use of Selective Availability (Leick, 1990). Land cover classes determined through aerial photography or remote sensing often have known confusion matrices, which give the probability that any one cover class may be confused with any other. Such errors enter the raw data, and stay with them through the entire process of database creation and use.

Definition

In photogrammetry, remote sensing, geodesy, surveying and other measurement-oriented spatial

sciences, the variables being measured are mostly well-defined. However in the more interpretational spatial sciences, such as soil science, geology, forestry or geography this is far from true, and one person's moraine can be another's kame. Measurements are not necessarily replicable between observers, and uncertainty of definition is a major source of inaccuracy in data. Although one can argue as to whether variation between observers really deserves to be called error, a system that fails to identify different parts of the map with different observers has no way of distinguishing this type of error from other types.

Lack of documentation

To extend the previous point, differences between observers become indistinguishable from other types of errors if not identified as such. For example, suppose two maps of the same area use different geodetic datums or different map projections. Without adequate documentation, the user may ascribe the differences in the two maps to error. In effect, then, lack of documentation is in itself a contributor to error: we could go so far as to define error as the unexplained differences between observations of the same phenomenon. Once the difference in datums is known, the discrepancies between the maps can be removed by a precise series of mathematical transformations.

Distortions in physical media

Much spatial data originates from maps, either printed or hand drafted. Although some media are very stable, paper is known to stretch with changes in humidity, and folding can also introduce unwanted distortion. Digitizing and scanning of paper documents produce precise measurements based on distances from a defined origin point, so distortions of the base produce errors that accumulate with distance from the origin. Thus although a typical digitizer can achieve precisions of fractions of a millimeter, errors on the order of several millimeters in position commonly occur due to distortion of the base. Some types of copying process are known to introduce distortions that vary markedly with direction.

Processing

GIS is a precise technology, and is usually carried out with precisions far in excess of the inherent accuracy of the data. Nevertheless, some GIS processes are capable of introducing substantial errors. The use of snap tolerances in digitizing, automatic sliver removal in overlay and raster/vector conversion are all sources of uncertainty at levels comparable to the accuracy of the data. For example, if a tolerance of 2mm is used to snap lines in digitizing, it is impossible to distinguish correctly between a lake with a narrow neck,

and two lakes, if the neck is 2mm or less across.

Interpretation

Very few systems for spatial data handling offer any extensive capabilities for formal documentation of data. The meaning attached to an attribute, such as class 1 on a land cover map, is more often entirely the responsibility of the user. The label "soil" attached to a file will have different meanings to different users, each of whom may have different needs. One user may be looking for geotechnical information on soil mechanics in connection with highway construction, while another may be looking for soil fertility information in connection with forestry. Far more documentation is needed if each is to be able to make effective use of the data, but it is very unlikely that that documentation will be in the database - more often it is on a sheet of paper somewhere which may or may not have been filed with the data, may or may not be legible, and may or may not be correct. Lack of documentation in the database, and the responsibility for interpretation that this places on the user, is in itself a significant source of error in spatial data, because of the different meanings that users are consequently free to attach to the same information.

The structure of errors

Errors of position are perhaps the more obvious type of spatial data error, although certainly not the only type. Most of the examples in the previous sections were of positional error. The error in the position of a point can be described by confidence limits in the x and y directions. However the impact of errors in spatial data is not just a function of their magnitudes, but also responds to the inherent relationships between them. For example, if the point is part of a larger object, such as a line, it may behave independently, leading to uncertainty in the shape of the object; or the object may behave as a rigid whole, every point in the object being subject to precisely the same error. Conceptually, we distinguish between cases like these by differentiating between absolute and relative positional error. Statistically, the distinction lies in the structure of correlations between errors: correlated errors (the case of the rigid object) produce very different effects from independent errors (each point behaving independently, distorting the shape of the object).

Unfortunately few geographical objects can be described as collections of discrete points - more common are continuous curves, which may be approximated as discrete points in the database. While we can describe uncertainty in the positions of continuous curves through such devices as buffer zones and epsilon bands, it has proven much more difficult to develop appropriate theories and robust measures.

Another common source of errors in spatial data is in the attributes of objects, rather than their positions. For example, the accuracy of digital elevation models (DEMs) is commonly described as an expected error in the elevation specified at each point, rather than in the positional accuracy of those points. Structure is important here also. For example, in the viewshed operation commonly used in GIS the accuracy with which a visible area can be estimated is strongly affected by the correlations between errors in the digital elevation model. Independent errors at each point will create a rugged distortion in the horizon, whereas correlated errors will raise or lower the entire horizon (Fisher, 1991).

METADATA AND LINEAGE

Thus far, we have established that data quality is a significant problem in spatial data handling. There are many sources of error in the pathway from observation to database, analysis and decision, and each has its own characteristics and processes. Error is a function of information. Once differences can be described by simple mathematical functions, they can be eliminated, but differences of unknown structure - random differences - can only be treated as error. Moreover these mathematical functions can be lost along the way, creating additional sources of error. If the map-maker forgets what projection was used to make a particular map, then in effect another source of error has been introduced, unless someone is clever enough to deduce the projection and its parameters.

We define metadata as information about data - knowledge of scale, date of creation, projection, legends and the meanings of symbols, the name of the mapmaker, and any other information that may be useful to the potential user. Of course, the more widely distributed the data, the more difficult it is to anticipate uses, and thus to determine what to include as metadata. The use of GIS merely makes this situation worse, by making it easier to copy, distribute and analyze data for a more and more varied set of purposes.

Two different approaches are useful in describing the accuracy of data to a potential user. One is by direct comparison between the data and the truth which it supposedly represents. Thus we could describe the digital elevation data above as having an accuracy of 7m, because the root mean square difference between the elevation in the database at a given point, and the true elevation at that point, is 7m. Because of frequent difficulties in defining "truth", it is often acceptable to compare the data to an independent source

known to have higher accuracy.

The alternative approach is to describe accuracy through the processes known to contribute to error. In the digital elevation model case, we could describe the various processes used to generate the data, and include estimates of the magnitudes of errors introduced at each step. Assuming that the errors at each step are statistically independent, total error is found by summing the squares of the individual error contributions, and taking the square root of the total. We will refer to these two as the "measurement" and "process" approaches to error estimation respectively. Both are useful, depending on the context and the information available to the user of the data.

Spatial Data Transfer Standard

The proposed federal Spatial Data Transfer Standard (SDTS; DCDSTF, 1988) contains a lengthy section on data quality. It establishes the principle of "truth in labeling" - that it is more useful to report all available information on data quality, and to lay out the minimal set of parameters that must be measured, than to establish thresholds and associated classifications. SDTS identifies five dimensions of data quality - positional accuracy, attribute accuracy, consistency, completeness and lineage - which can be defined for the entities in any digital cartographic database. Under each heading there are prescribed indicators of quality that must be measured or estimated, and reported in the data quality statement accompanying the dataset. Under the lineage heading, the distributor of the dataset is expected to provide information on the stages in the creation of the data, and the uncertainties introduced at each stage.

SDTS provides a comprehensive template for data quality reports. It falls short of defining precise format standards for metadata, or of requiring that data quality reports become an integral part of the formatted database and inseparable from it, and if adopted as a Federal Information Processing Standard it can be enforced only for datasets handled by the federal government. Finally, as the name of the original task force suggests, its concern is primarily with cartographic data, and the uses to which such data is commonly put. Traditionally, cartographers have assumed that the user will "read" the map visually, but increasingly the same maps are being used to create digital databases and support detailed and precise analysis. A standard oriented towards the quality of maps clearly cannot anticipate all of these new uses, or the error descriptors relevant to each of them. For example, spatial dependence of errors, which we have argued to be relevant in viewshed estimation from DEMs and many other applications of spatial data, is not

listed as a parameter to be estimated in SDTS data quality reports.

Lineage handling

GIS software has been very slow to adopt the idea of integrated metadata. Even such obvious parameters as the scale of the map from which the data was digitized, the date of creation, or positional accuracy must be communicated to the user through written documentation, as the system has no means of storing, handling or reporting them. Soil maps often carry elaborate text explanations of classes, but the digitized soil map in the database simplifies these to simple alphabetic codes, and leaves it up to the user to search out more complete descriptions of each code's meaning. Some systems will go so far as to identify class 1 with "Grimsby Loam", but will not help the user to access the 1,000 words of detailed description of Grimsby Loam provided by the mapmaker, often on the back of the sheet.

There are some exceptions to this general lack of interest in metadata on the part of GIS designers. The recent version of IDRISI includes several standard metadata fields for each layer, and in relational systems such as ARC/INFO it is possible for the database designer to link metadata files to objects or coverages in particular implementations. However neither of these systems provide capabilities for handling metadata, such as the automatic propagation of metadata from input to output in GIS processes. For example, a metadata handler should attempt to estimate the accuracy of the output of an operation like overlay given the accuracies of the input layers.

Recently the American Society for Testing and Materials (ASTM) has proposed a draft standard for GIS metadata. The draft lays out the fields appropriate to particular types of data, and in many cases defines methods of measurement and formats for description. The draft has been discussed at several recent GIS meetings, including URISA in San Francisco in August, 1991, and GIS/LIS '91 in Atlanta. A continuing problem, however, is the need to anticipate data uses in deciding what metadata fields to include in the standard.

Lanter and Veregin (Lanter, in press; Lanter and Veregin, in press) have described a research effort at UC Santa Barbara to develop a metadata handler, called Geolineus, which provides a user interface to a GIS command language with full access to appropriately defined metadata. The user interacts with information on each layer in the database through icons and a "point and click" interface, and the system automatically propagates metadata using appropriate rules. Systems like this replace the traditional approach,

in which metadata are entirely the responsibility of the user, and are often ignored, with a set of tools to inform the user of the nature of the data being processed.

3. QUALITY ASSURANCE AND QUALITY CONTROL

Quality assurance and quality control (QA/QC) are major concerns of agencies and activities that must deal with the potential threat of litigation. Standards such as SDTS lay out the parameters that must be measured and described in a quality report, but they fall short of providing the kinds of assurances demanded in a legal setting, that all reasonable efforts have been made to control and maintain the quality of the data. Such assurances are more likely to be provided by procedures and protocols, and evidence that these have been enforced and followed in practice, than by objective measures of quality. In other words, the legal context invariably emphasizes process rather than measurement.

In the extreme, QA/QC can place entirely unreasonable demands on process. In GIS applications, the highest level of QA/QC would require that every item of information in the database have attached to it a documented line of responsibility - that someone "signed off" on every item. While this is possible in principle, its imposition in practice would make it impossible to build the kinds of multithematic, detailed GIS databases now commonly in use. Because the data are processed by software, the same standard would apply to the code - that every line and every change have some documented line of responsibility attached to it.

The measurement approach to quality is fundamentally statistical. For example, the Circular Map Accuracy Standard, a commonly used measure of positional accuracy of points, is defined as the radius of a circle within which the true location of the point is expected to lie 90% of the time. The remaining 10% of cases are expected to lie outside the circle, and it is possible for individual points to lie at distances of many times the CMAS. But while averages and statistical expectations are the basis of such measures, they may have little value to QA/QC procedures concerned with maintaining rigorous levels of quality throughout the database. A single mislocated point may be entirely acceptable within a statistical framework, but can be very damaging to credibility in a courtroom, where maximum error is likely more relevant than average error.

Various efforts are under way to establish QA/QC procedures for GIS applications. The Environmental Protection Agency, for example, is developing a set of QA/QC procedures to cover the wide variety of spatial databases and GIS-supported decision-making within the agency. These, coupled with

effective reporting of data quality and handling of metadata, should help to remove some of the barriers to greater data sharing.

CONCLUSION

The central theme of this paper has been that lack of information on data quality, and lack of quality itself, is a major impediment to data sharing. No one is willing to make use of data they do not trust, or whose accuracy they do not understand. But since the effective description of quality is a function of use, the task of describing quality against all possible uses is formidable, especially given the ease with which data can be disseminated and applied in novel ways in the world of GIS. In this sense GIS is its own worst enemy: by inviting people to find new uses for data, it also invites them to be irresponsible in their use.

In this situation some degree of compromise is essential, and traditionally mapmakers have provided data quality reports that satisfy the largest possible number of users while requiring only a reasonable level of effort. Thus users of topographic maps must often infer quality from simple diagrams showing the level of accuracy of various parts of the map using terms such as "relatively low" and "relatively high". These can be made with little effort, and are adequate for many purposes.

For better or worse, the widespread use of GIS to process spatial data has changed the parameters of this compromise. Simple quality diagrams are no longer sufficient to help the GIS user assess the accuracy of his or her output. Instead, GIS has led to an explosive increase in the applications of spatial data, and this requires a parallel increase in the effort invested in measuring and documenting data quality. Furthermore, because such documents are of no value unless they are available to the user, it is essential that quality be handled and transmitted as metadata, fully integrated with the data themselves.

Data quality is a messy issue, and the average GIS user would probably far rather ignore it than confront it. But as GIS matures, we can expect more and more to find GIS results challenged in court, and it will be less and less acceptable to rely on the overused notion that computer output speaks louder - that the output of a GIS beats a handdrawn map every time. Increasingly, both sides in a suit will have access to the same technology, and the winner will be the one that can demonstrate higher quality, in data as well as in analysis.

ACKNOWLEDGMENT

The National Center for Geographic Information and Analysis is supported by the National Science Foundation, Grant SES 88-10917.

REFERENCES

Digital Cartographic Data Standards Task Force (1988) The proposed standard for digital cartographic data. *The American Cartographer* 15(1): 1-140.

Fisher, P.F. (1991) First experiments in viewshed uncertainty - the accuracy of the viewshed area. *Photogrammetric Engineering and Remote Sensing* 57(10): 1321-1327.

Goodchild, M.F. and S. Gopal (1989) *Accuracy of Spatial Databases*, Taylor and Francis, London.

Lanter, D. (in press) A lineage-based meta-database for GIS. *Cartography and GIS*.

Lanter, D. and H. Veregin (in press) A research paradigm for propagating error in layer-based GIS. *Photogrammetric Engineering and Remote Sensing.*

Leick, A. (1990) *GPS Satellite Surveying*, Wiley, New York.

Veregin, H. (1989a) *Technical Paper 89-9*, National Center for Geographic Information and Analysis, Santa Barbara, CA.

Veregin, H. (1989b) *Technical Paper 89-12*, National Center for Geographic Information and Analysis, Santa Barbara, CA.