SPATIAL DATA QUALITY

Gary J. Hunter and Simon Jones
University of Melbourne, Australia

Arnold Bregt
Center for Geo-Information, Wageningen University and Research Center, The Netherlands

Ewan Masters
University of New South Wales, Australia

Keywords: Spatial data, accuracy, error, uncertainty, quality elements, current problems

Contents

Summary

Spatial data quality is an important topic, for both the users and the creators of spatial information. Today, policymakers, resource mangers, scientists, and the general population all make increasing use of spatial information. This increase in the volume of spatial data being processed, the routine but complex nature of the spatial analytical operations being conducted, and the provision of "non-expert" driven computer information systems, have resulted in an "empowered" information-rich environment where decisions can be taken by "informed" managers and scrutinized by an "informed" populace. However, these advances have also resulted in a vastly increased potential for the introduction of error into derived spatial data products. Furthermore, these spatial data quality issues introduce uncertainties into any subsequent decisions that are made using such data. This article discusses the topic of spatial data quality by considering the various issues and problems associated with understanding, communicating, and ultimately living with the uncertainties inherent in spatial information. To begin with, the importance of spatial data quality is outlined in the context of data standards, protecting professional reputations, litigation, and scientific enquiry. Next, the various elements of spatial data quality are presented, including sections on the sources of error and a classification of error types. This leads to a discussion of error modeling, the communication of uncertainty, and data quality management issues. The practical application of these theories is illustrated throughout by reference to various case

studies with data quality/decision uncertainty themes. Finally, current issues and future trends in spatial data quality are presented for readers aiming to pursue the subject in greater depth.

1. Introduction

Let us begin this article by considering the following real-life situations. First, a truck driver running low on fuel, and using a vehicle navigation system, needs to know whether all re-fueling stations that exist for the next 100 km along the truck route have been entered in the navigation system. Second, a water supply authority intends to acquire land for a proposed reservoir and needs to know exactly which land parcels will be inundated by the reservoir when it is complete, and who the current legal owners are so that it can start legal negotiations with them to purchase their properties. Third, an advertising company executive wants to be reassured that the mail addresses are at least 80 percent accurate in an expensive commercial address database that has just been purchased for a nationwide advertising campaign. In each of these cases, the issue that matters most to the people who will be using the information is one of spatial data quality, and in each situation quality is of key importance to them and the decisions they will ultimately make.

In fact, the topic of quality in spatial databases, and the resultant accuracy of the information products derived from this data, has now assumed far greater importance than when all we had were traditional paper-based maps. This is owing in part to the large variety and number of spatial operations that can now be performed with relative ease to create a host of dazzling maps, displays, and images (see ***Spatial Data Management : Topic Overview***). Spatial data has also become far more widely available to users, and it is quite easy for data to be applied for purposes for which it was not originally intended. Add to this the mistaken impression, at least in the mind of novice users, that "the computer is always right," and the potential for introducing uncertainty into spatial data and the decisions taken with it can be considerable. Many spatial data users are now realizing that cartographers in fact possessed a wide range of technical and analytical skills, the complexity of which are only being understood now as researchers try to model those same processes through the application of knowledge-based techniques. In fact, the ease with which digital spatial data can be manipulated has the potential to become a double-edged sword. Limitations and difficulties that were once obvious during manual processing are now often completely hidden from view, and are prone to "infect" or "taint" even the most handsome of map products.

Of course, the data quality issue is not new to professionals such as land surveyors, cartographers, and soil scientists, who always had a sound knowledge of the errors that could arise during field surveys and map production. To overcome these effects, they tried to convey estimates of map product accuracy to users through information in the map margins such as reliability diagrams, special symbolization for uncertain areas, and horizontal and vertical positional accuracy statements based on testing to established mapping standards. However, there was a period in the 1970s and 1980s when much of this core information disappeared from new digital map products, mainly because the software available at that time only stored the graphics and attribute information and not the associated metadata (that is, data about the data) that had traditionally been

associated with conventional map products. To compound this situation, the creation and management of spatial databases has now passed from the domain of the "expert few" into the hands of a much wider section of the community. This includes many new users who, while being attracted to the new-found possibilities open to them through spatial processing, do not have the necessary background knowledge concerning the inherent errors that might be present in their datasets, or the errors that might have inadvertently been propagated through their own actions or the operations of their software (see ***Advanced Geographic Information Systems).***

Leading researchers realized that errors in digital spatial data had the potential to cause problems that had not been experienced before with paper maps. With these warnings an international trend started in the early 1980s to develop data transfer standards that would include data quality information (see *Spatial Data Standards*). The standard that led the way in documenting data quality was the US Government's Spatial Data Transfer Standard (SDTS), and it divided the task of data quality reporting into five essential elements: dataset lineage (that is, the history of where the dataset came from and how it was developed); positional accuracy; attribute accuracy; logical consistency; and completeness. (For more on standards, see ***Spatial Data Standards.***)

However before we examine these and related subtopics in more detail, a few words must be said about the terminology used in this subject. As can be seen in the glossary, "accuracy" relates to the closeness of data to the truth, and in situations where it can be measured, the term "error" is used to express the difference that exists. On the other hand, "uncertainty" implies a lack of knowledge or sureness about just what the truth might be and/or how accuracy may be measured. Finally, "quality" in the sense of spatial data implies the "fitness for use" of the data to be applied to a particular task. In essence, there is no such thing as "bad" data, but there are instead data that lack the necessary quality for a given application (since data that are unsuitable for one user may be quite acceptable to another). Throughout the remainder of this article, examples will be given that help illustrate these terms.

The rest of this article is structured such that the reasons data quality has become an important issue are presented next, followed by a practical explanation of the elements of spatial data quality and sources of error, and finally closing with a review of the main problems that still need to be addressed in this topic, and some of the future trends that may lead to their solution.

2. The Importance of Spatial Data Quality

During the past 20 years, both the producers and users of spatial data have become increasingly concerned at their inability to measure and communicate the accuracy and quality of their information products. Their anxiety is well placed for the following reasons:

- There are growing statutory and business pressures to provide data quality statements when transferring spatial data;

- There is a natural need to protect individual and agency reputations, particularly when spatial data are used to support administrative decisions that may be subject to legal challenge (see ***Geographic Information Legal Issues);***
- The issue of consumer rights has grown in importance, and data producers feel the need to minimize the risk of any litigation that may arise as a result of alleged harm being suffered through the use of spatial data; and
- There is an obligation to satisfy the basic scientific requirements of being able to describe how close spatial information is to the truth it supposedly represents.

Of these motives, those causing greatest concern at present are the threats to individual and agency reputations, and the question of legal liability. Therefore, although the question of accuracy and data quality may not be a "money-making" subject, it may well be a "money-losing" one if it is not dealt with appropriately in the future. This section discusses why the spatial information community should be concerned with the accuracy and quality of spatial data.

2.1. Increased Pressure to Report Data Quality

First, numerous governments around the world have established a mandatory requirement for public sector agencies to provide data quality statements when transferring data. This enables users to decide more efficiently whether the data are suitable for their needs, without having performed their own detailed analysis of it or conducted lengthy interviews with the data custodian. Similarly, private sector data producing agencies are being pressured to provide data quality reports for their products through competitive business forces. While it is possible for private data producers to provide minimal statements in which the various quality elements are listed simply as "unknown," in reality this is not a sound business strategy when other producers are taking the effort to inform users, to the best of their knowledge and ability, of the accuracy and quality of the data being supplied.

2.2. Protecting Reputations

Next, because spatial information is invariably used for decision making at differing levels, the lack of accuracy estimates has the potential to cause harm to both personal and agency reputations and the public's confidence in them—particularly in cases where administrative decisions are subject to judicial review and the use of geographic information systems (GIS) may be open to challenge before courts of law. For example, during the 1990s spatial data accuracy became a key factor in questioning government orders to restore contaminated private land in the United States, and the Environmental Protection Agency (EPA) in that country was estimated to have allocated US$1 billion to fund expected court cases with affected land owners seeking to overturn decisions against them which were made in part using GIS. Clearly, when the stakes are high enough (and some land owners received clean-up bills for many millions of dollars even though the toxic waste problems occurred on their land many years before they became the owners), issues such as accuracy quickly come to the forefront. In addition, in natural resource and environmental disputes it is now quite common for GIS to be used by each party in the dispute, and cases of "GIS versus GIS" are increasingly occurring,

in which the legal tactics invariably include arguments designed to discredit an opponent's use of GIS, and the spatial operations and models employed to arrive at a particular finding.

-
-
-

Bibliography

Australian Surveying and Land Information Group (AUSLIG). (undated). *TOPO-250K GEODATA Specifications.* AUSLIG website, URL: <http://www.auslig.gov.au/mapping/specs.htm> [This site contains an excellent example of a spatial data quality report.]

Beard M.K. and Mackaness W. (1993). Visual access to data quality in geographic information systems. *Cartographica*, **30**, 37–45. [A definitive paper on the various methods by which spatial data quality may be visually communicated to users.]

Burrough P.A. and McDonnell R.A. (1998). *Principles of Geographical Information Systems*, 333 pp. New York: Oxford University Press. [This book introduces readers to the essential elements of GIS and has excellent chapters dealing with errors and quality control in geographic data, and error propagation in numerical modeling.]

Chrisman N.R. (1991). The error component in spatial data. *Geographical Information Systems: Principles and Applications*, 1st edition, vol. 1 (eds. D.J. Maguire, M.F. Goodchild and D.W. Rhind), pp. 165–174. London: Longman. [A clear and concise introduction to spatial data error and the elements of data quality. This paper and others from this publication are now freely available via the Internet at <http://www.wiley.com/legacy/wileychi/gis/resources.html>.]

Crosetto M. and Tarantola S. (2001). Uncertainty and sensitivity analysis: tools for GIS-based model implementation. *International Journal of Geographical Information Science*, **15**(5), 415–47. [A very practical paper describing how different types of spatial data error can be taken into account to test the reliability of different types of flood forecasting models.]

Epstein E.F. and Roitman H. (1987). Liability for information. *Proceedings of the 1987 Urban and Regional Information Systems Association Annual Conference*, Chicago, URISA, vol. 4, 115–125. [This conference paper contains many real-life examples of the effects of spatial data error and its consequences in a legal setting.]

Goodchild M.F. and Gopal, S. (1989). *The Accuracy of Spatial Databases*, 290 pp. London: Taylor and Francis. [This book is for the advanced reader and summarizes the research issues associated with spatial data accuracy]

Heuvelink G.B.M. (1999). Propagation of error in spatial modeling with GIS. *Geographical Information Systems: Principles, Techniques, Management and Applications*, 2nd edition, vol. 2 (eds. D.J. Maguire,

M.F. Goodchild, D.W. Rhind and P.A. Longley), Chapter 14, pp. 207–217. London: John Wiley. [This paper describes how the error that propagates through numerical models using spatial data can be quantified.]

Hunter G.J. (1999) Managing uncertainty in geographic information systems. *Geographical Information Systems: Principles, Techniques, Management and Applications*, 2nd edition, vol. 2 (eds. D.J. Maguire, M.F. Goodchild, D.W. Rhind and P.A. Longley), pp. 633–641. London: John Wiley. [This paper describes how the effect of spatial data quality may be managed through the use of risk management.]

Ordnance Survey. (undated). *SUPERPLAN Data Users Guide.* British Ordnance Survey website, URL: <http://www.ordsvy.gov.uk/ > [Contains an excellent example of digital data quality reporting]

Biographical Sketches

Gary J. Hunter came to academia in 1988 after 17 years in industry, and his experience includes engineering construction, cadastral and topographic mapping projects in Australia, Indonesia, and Papua New Guinea. In 1990 he was the first lecturer appointed in Geographic Information Systems at the University of Melbourne, and he is now an Associate Professor and Reader in the Department of Geomatics. In 1994 he gained his Ph.D. from the University of Melbourne on the subject of managing uncertainty in spatial databases. He is a regional/section editor of *Transactions in GIS* and the *URISA Journal*, and serves on the editorial boards of the *International Journal of Geographical Information Science* and *GeoInformatica*. In 1996 he served as President of the Australasian Urban and Regional Information Systems Association (AURISA), and was appointed a Life Member of AURISA in 1999.

Simon Jones is a Research Fellow in the Department of Geomatics at the University of Melbourne. He gained his Ph.D. from the University of Leicester (UK) on the subject of understanding remotely sensed data from tropical forested areas. His research interests are in spatial data uncertainty and its impact on environmental decision making, land cover mapping, and monitoring. Before joining the University of Melbourne he worked as a researcher for European Commission TREES project.

Arnold K. Bregt is Professor of Geo-Information Science at Wageningen University in the Netherlands. Following more than 15 years of experience in the field of GIS research and application, his current areas of interest are spatial data quality, dynamic modeling of land use change, and spatial data infrastructures. From 1996 he was one of the project leaders and initiators for the development of the Dutch spatial data infrastructure.

Ewan Masters has a degree in Surveying and a Ph.D. in Geodesy from the University of New South Wales, Australia. He has been working in various research and teaching capacities at the University of New South Wales since 1981. He joined the academic staff in 1986 with teaching responsibilities in the areas of spatial information systems and data analysis, and is currently a senior lecturer in the School of Surveying and Spatial Information Systems. He is the postgraduate coordinator for the school as well as being a member of the Postgraduate Committee and Computing Committee for the Faculty of Engineering. His main research interests include managing and improving the quality of spatial databases.