Foreword

The quality of spatial data, as indeed of any data, is crucial to its effective use. Spatial data purport to represent aspects of the spatial world, and in the context of this book that means primarily the geographic world, the world defined by human experience and by the surface and near-surface of the Earth. Quality, as the definitions in this book attest, is a measure of the difference between the data and the reality that they represent, and becomes poorer as the data and the corresponding reality diverge. Thus, if data are of poor quality, and tell us little about the geographic world, then they have little value.

This argument seems watertight, and the examples in the Introduction provide ample illustration. But as Nicholas Chrisman notes in Chapter 1, the initial reaction of many leaders of the field was negative. My own experience was similar; in 1977, I gave a presentation on data quality to an international conference of experts in spatial data, and was met with something between indifference and outright opposition. In the almost 30 years since then, there has been no massive outcry among users of geographic information systems (GIS) and spatial databases, demanding better methods of handling the uncertainty present in data; despite many warnings, there have been few court cases over bad decisions that resulted from poor data, and the major GIS vendors still provide little in the way of support for handing information about data quality. Yet the geographic information science (GIScience) community continues to identify data quality as a topic of major significance, and much progress continues to be made, as the chapters of this book will confirm.

I think that there are several explanations for this apparent contradiction, and they lie at the very heart of GIScience. First, these issues cannot be ignored by anyone with a scientific conscience. It makes no sense whatsoever for a GIS vendor to claim that his or her software stores coordinates to double precision, in other words to 14 significant digits, when one part in 10^{14} of the linear dimension of the Earth is approximately the size of a molecule. None of our devices for measuring position have accuracies that are any better than one part in 10^7, or single precision, but in this and in many other instances it is the precision of the digital computer that masquerades as accuracy. Any self-respecting scientist knows that it is misleading to report any result to a precision that exceeds its accuracy, yet our GIS software does so constantly. It is clearly the responsibility of the GIScience community to draw attention to such issues, to reflect on the role that software plays, and to demand that it adhere to the best of scientific principles.

Second, there is a long tradition in map-making of compromising the objective of portraying the world accurately with the potentially conflicting objective of visual clarity. A contour might be kinked, for example, to emphasize the presence of a stream, whether or not the stream's course is actually indented in the landscape. A railway running close to a road might be separated from it in the interests of avoiding visual confusion. Thus a map can be far from a scientifically accurate representation of the Earth's surface, yet it is natural to assume that the contents of a map, once digitized, stored in a database, and analyzed using GIS software, are indeed scientifically accurate. The cartographic perspective even leads to a somewhat different interpretation of data quality – a digitized map can be said to be perfect if it exactly represents the contents of the paper map from which it was obtained, whether or not the paper map exactly represents phenomena in the real world.

Third, while many of the methods discussed in this and other books on the topic of spatial data quality are intuitive and simple, the theoretical frameworks in which they are grounded – spatial statistics, geostatistics, and set theory – are complex and difficult. Quality is difficult to attach to individual features in a database, but instead must be described in terms of the joint quality of pairs of features, through measures of relative positional accuracy, covariance or correlation. Surveyors have dealt with these problems for decades, and have developed appropriate training regimes for their students, but many users of GIS lack the necessary mathematical skills to handle complex models of spatial data quality. Instead, researchers have had to look for clever visual ways of capturing and communicating what is known about quality, about how quality varies from one type of feature to another, and about how it varies from one geographic area to another. And as with any technology that makes difficult mathematical concepts accessible to a broad community of users, there is always the potential for misinterpretation and misuse.

These three issues are the common threads that have run through the work that I have done on the topic of spatial data quality over the past three decades. Thinking back, my own interest in the topic seems to have stemmed from several intersecting themes and ideas. First, I was fascinated with the field of geometric probability, and the elegant results that had been obtained by such mathematicians as Buffon and Coxeter – and thought that these ideas could be applied to maps. As Ashton

Shortridge and I showed in a paper many years later, the statistics of a vector crossing a raster cell can be related to Buffon's famous problem of the needle randomly dropped on a set of parallel lines ([SHO 02]). Second, I was bothered by the lack of any simple models of the errors introduced by digitizing, or of the uncertainties inherent in such simple GIS operations as the measurement of area. One of the paradoxes of GIS is that it is possible to estimate properties such as slope accurately even from a very inaccurate digital elevation model, because slope responds to the covariance of errors in addition to the variance, as many GIScientists have shown (see, for example, [HUN 97]). Third, I was struck by the wealth of knowledge in disciplines such as geostatistics and surveying that was virtually unknown to the GIScience community. A paper on measurement-based GIS ([GOO 02]), for example, was prompted by what I perceived as a need to bring research on adjustment theory into the GIScience literature.

Much of the early work in GIS was dominated by the desire to create accurate digital representations of the contents of maps. The Canada Geographic Information System of the 1960s, for example, saw as its primary mission the capture of mapped information on land, followed by calculation and tabulation of area; many other early GIS projects had similar goals. Much later, GIScientists began to look systematically at the results of these projects, and the degree to which their results replicated not the contents of maps, but the contents of the real world. By that time, of course, many of the fundamental design decisions of GIS had been made. Those decisions were predicated on the assumption that it was possible to create a perfect representation of the contents of a map, and even today that assumption seems reasonable. But, as I am sure all of the authors of the chapters of this book would argue, it is not possible to create a perfect representation of the infinite complexity of the real world. If the field of GIS had begun in the mid-1960s with that assumption, one might reasonably ask whether the design decisions would have been the same. Does a technology designed for the goal of perfect representation of the contents of maps adapt well to the imperfect representation of the contents of the real world? This seems to me to be one of the most profound questions that GIScience can ask – in effect, it asks whether the ontological legacy of GIS is consistent with its fundamental objectives.

The chapters of this book present an excellent overview of the dimensions of spatial data quality research, from the most theoretical and abstract to the most practical and applied. The book has no epilog or concluding chapter, so perhaps I might be permitted to offer a few comments on where the field might be headed. Previous comments notwithstanding, there does appear to be steady progress in the adoption of a greater sensitivity to spatial data quality issues among the user community and GIS software vendors. Better standards for description of spatial data quality are being adopted, and are being supported by software. Suppliers of data are more likely to provide statements of data quality, and to test products against ground truth, than they were in the past. A greater range of examples is available in the literature, and spatial data quality is now an obligatory part of the GIS curriculum. This book, and its availability in English, will add substantially to that literature.

That said, however, the central problem seems as unsolved as ever – how to communicate what is known about spatial data quality to an ever-expanding population of users, many of whom have very little understanding of the basic principles of GIScience. In the past year, we have seen a massive expansion of access to spatial data, through the introduction and widespread popularity of Google Earth and similar tools. Very few of the people recruited to the use of spatial data by these technologies will have any understanding of spatial data quality issues, but many of them will likely have the motivation to learn, if the research community can develop and implement appropriate techniques. This book and its coverage of the important issues should keep us moving in the right direction.

Michael F. Goodchild

Bibliography

[GOO 02] M.F. GOODCHILD (2002) "Measurement-based GIS" in W. SHI, P.F. FISHER and M.F. GOODCHILD, editors, *Spatial Data Quality*. New York: Taylor & Francis, pp. 5–17.

[HUN 97] G.J. HUNTER and M.F. GOODCHILD (1997) "Modeling the uncertainty in slope and aspect estimates derived from spatial databases", *Geographical Analysis* 29(1): 35–49.

[SHO 02] A.M. SHORTRIDGE and M.F. GOODCHILD (2002) "Geometric probability and GIS: some applications for the statistics of intersections", *International Journal of Geographical Information Science* 16(3): 227–243.

Introduction

The quality of geographic information (or geospatial data quality) has always presented a significant problem in geomatics. It has experienced significant growth in recent years with the development of the World Wide Web, increased accessibility of geomatic data and systems, as well as the use of geographic data in digital format for various applications by many other fields.

Problems regarding data quality affect all fields that use geographic data. For example, an environmental engineer may need to use a digital elevation model to create a model of a watershed, a land surveyor may need to combine various data to obtain an accurate measurement of a given location, or someone may simply need to surf the Web to find an address from an online map site. Although many people associate the term "quality" only with the spatial accuracy of collected data (e.g. data collected by Global Positioning System (GPS) in relative mode will be of "better quality" than those digitized from paper maps at a scale of 1:100,000), the concept of quality encompasses a much larger spectrum and affects the entire process of the acquisition, management, communication, and use of geographic data.

Research into certain aspects of geographic data quality has been ongoing for a number of years, but interest in the subject has been increasing more recently. In the 1990s, the scientific community began to hold two conferences on this subject:

– *Accuracy* (*International Symposium on Spatial Accuracy Assessment in Natural Resources & Environmental Sciences*): addresses uncertainty regarding, primarily, the field of natural resources and the environment. This conference has been held every two years for the past ten years.

– *ISSDQ* (*International Symposium on Spatial Data Quality*): addresses geographic data quality in general. This conference has been held every two years since 1999.

Written by Rodolphe DEVILLERS and Robert JEANSOULIN.

In addition to events dealing entirely with spatial data quality, many large conferences on geomatics and on other related fields hold sessions on spatial data quality. Several international organizations also have working groups that address issues regarding quality, such as the International Society for Photogrammetry and Remote Sensing (ISPRS) (WG II/7: *Quality of Spatio-Temporal Data and Models*), the Association of Geographic Information Laboratories in Europe (AGILE) (*WG on Spatial Data Usability*), and the International Cartographic Association (ICA) (*WG on Spatial Data Uncertainty and Map Quality*). Quality is also a concern for standardization bodies that address it in a general way for the production and distribution of goods and services (for example, ISO 9000), but also for the field of geographic information specifically, often with regard to the standardization of metadata (for example, FGDC, OGC, CEN, ISO TC/211).

Several books have been published over the last decade on the subject of quality [GUP 95; GOO 98; SHI 02]. The first [GUP 95] presents reflections on the elements of quality by members of the *Spatial Data Quality* commission of the International Cartographic Association. The two other books [GOO 98 and SHI 02] present research breakthroughs on issues regarding quality and the uncertainty of geographic information. Other books have also been published about the problems of uncertainty in geographic data (for example [ZHA 02; FOO 03]). The present book differs from the above-mentioned publications by attempting to offer a more global, complete, and accessible vision of quality, intended for a wider range of users and not only for experts in the field.

Decision and uncertainty: notion of usefulness of quality

The notion of usefulness of quality depends first on the use that is made of the data it quantifies, particularly during and after a decision. Decision and quality are part of a dialectic that is similar to that of risk and hazard. Decision may be regarded as the conclusion of an informed and logical process, in which the treatment of uncertainty[1] must be present. The purpose of the usefulness of quality is then measured by its ability to reduce the uncertainty of a decision. Research into probability and statistics, artificial intelligence, and databases have examined the question of uncertainty for many years. Spatial and temporal data are largely present today: for example, in conferences such as UAI (*Uncertainty and Artificial Intelligence*), in international conferences on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU), and in European Conferences on Symbolic and Quantitative Approaches to Reasoning and Uncertainty (ECSQARU).

[1] The concept of uncertainty is presented in more detail in Chapter 3.

The purpose of this book is to present a general view of geographic data quality, through chapters that combine basic concepts and more advanced subjects. This book contains 15 chapters grouped into four parts and includes an appendix.

Organization of the work

The first part, entitled “Quality and Uncertainty: Introduction to the Problem”, introduces the work and provides certain basic concepts. Chapter 1 presents a brief history of the study of geographic data quality, showing how the field developed and grew from strict problems of spatial accuracy to include much more comprehensive considerations, such as the assessment of fitness for use. This historical perspective allows a better understanding of how certain aspects of the study of quality developed to become what it is today. Chapter 2 presents general concepts on quality, definitions, certain sources of the problem with quality, as well as the fundamental distinction in geographic information between what is called “internal quality” and “external quality”. Chapter 3 positions and defines different terms found in the field of data quality and uncertainty (for example, uncertainty, error, accuracy, fuzziness, vagueness) and describes these different types of uncertainties using examples.

The second part is titled “Academic Case Studies: Raster, Choropleth and Land Use”. Data quality is often determined at the moment of data acquisition (for example, depending on the technique used). Chapter 4 presents problems associated with quality in the field of imagery, taking into account geometric and radiometric factors. Chapter 5 addresses the inaccuracy of maps as a result of misinterpretation, highlighting the difficulty of classifying natural environments in which precise boundaries between phenomena do not always exist. Chapter 6 presents statistical methods used to measure and manipulate uncertainties related to data. Finally, Chapter 7 presents several methods of quantitative and qualitative reasoning that allow the manipulation of uncertain data. These concepts are illustrated through an example of land use.

The third part, entitled “Internal Quality of Vector Data: Production, Evaluation and Documentation”, presents the problems of quality at different stages of producing vector data. Chapter 8 addresses the concerns of a geographic vector data producer for quality related issues (for example, quality control and quality assurance). Chapter 9 presents a specific approach to improve internal data quality during production, based on the definition of rules that verify whether the objects have possible relationships according to their semantic. Chapter 10 presents how to describe the quality of data produced, as well as national and international standards on the description of quality and its documentation in the form of metadata. Finally, Chapter 11 complements Chapter 10, by presenting an evaluation of quality from a conceptual point of view and then addressing several methods that can be used to measure quality.

The fourth and final part, entitled “External Quality: Communication and Usage”, takes the point of view of the user, who bases his or her decisions on the data. Chapter 12 addresses the communication of information on quality to users and the use of information by users. It discusses questions such as the management and visual representation of quality information. Chapter 13 presents a formal approach based on ontologies to evaluate the fitness of use of data for a given purpose (that is, external quality). Chapter 14 addresses the relationship between data quality and the decision-making process based on these data. Finally, Chapter 15 deals with legal considerations, such as civil liability associated with data quality.

In conclusion, we hope that this book will help to answer many questions regularly asked regarding the quality of geographic information:

– Everyone says that data quality is important; is that true? Yes. It is essential. Failure to consider data quality may result in serious consequences.

– Everyone says that quality is complex? Yes, but there are a growing number of methods to overcome this complexity.

– Is there a better way to consider quality? No. It is impossible to answer, *a priori* and in a unique way, the question of whether or not the data are good, even in the case of a precise application defined in advance. Solutions exist however, that clarify the use made of data, for a given purpose, according to known information on their quality. The world is not deterministic. The ultimate decision is subjective and depends on the user.

Bibliography

[FOO 03] FOODIE G.M. and ATKINSON P.M. (eds), *Uncertainty in Remote Sensing and GIS*, 2003, New York, John Wiley & Sons, p 326.

[GOO 98] GOODCHILD M. and JEANSOULIN R. (eds), *Data Quality in Geographic Information: From Error to Uncertainty*, 1998, Paris, Hermès Science Publications.

[GUP 95] GUPTILL S.C. and MORRISON J.L. (eds), *Elements of Spatial Data Quality*, 1995, Oxford, UK, Pergamon Press, p 250.

[SHI 02] SHI W., GOODCHILD M.F. and FISHER P.F. (eds), *Spatial Data Quality*, 2002, London, Taylor and Francis, p 336.

[ZHA 02] ZHANG J. and GOODCHILD M.F., *Uncertainty in Geographical Information*, 2002, London, Taylor and Francis, p 192.