PREFACE

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 One, 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 thirty 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 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 whatever 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 (Shortridge and Goodchild, 2002). 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 Hunter and Goodchild, 1997). 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 (Goodchild, 2002), 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; and 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 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

REFERENCES

M.F. Goodchild (2002) Measurement-based GIS. In W. Shi, P.F. Fisher, and M.F. Goodchild, editors, *Spatial Data Quality*. New York: Taylor and Francis, pp. 5–17.

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.

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.