Introduction

THE EDITORS

A recurrent theme throughout the first part of this book is that the inherent complexity of the geographical world makes it virtually impossible for any digital representation to be complete, however limited its scope. Although some exceptions exist (we can, for example, create a perfect digital representation of the latitude of the Equator, or a line on the Earth's surface that is by definition straight), there will otherwise be differences between the database contents and the phenomena they represent. Various terms are used to describe these differences, depending on the context. Differences can exist because of errors of measurement, while the term 'uncertainty' seems more appropriate if the digital representation is simply incomplete. More generally, one might simply refer to the 'quality' of the representation.

If data quality is an important property of almost all geographical data, then it must affect the decisions made with those data. In general, the poorer the quality of the data, the poorer the decision. Bad decisions can have severe consequences, as when an ambulance is sent to the wrong location, or a school is inadvertently built over an abandoned storage facility for hazardous waste. Geographical data are often used for regulatory purposes, or to resolve disputes: the custodians of such data are clearly exposed to potential liability if the data are shown to be in error.

Despite what appear to be obvious arguments in favour of explicit treatment of data quality in GIS, and despite substantial research into appropriate methods, much GIS practice continues to proceed as if data were perfect. Results of GIS analysis – whether in the form of tables, maps, or displays – rarely show estimates of confidence, or other indicators of the effects of data quality. In part, such attitudes have been inherited from cartographic practice, since it is often difficult to determine the quality of mapped information. In part, they may also reflect a general tendency to give computers more credit than they deserve – to believe that because numbers or maps have emerged as if by magic from digital black boxes, they must necessarily be reliable.

This section contains four chapters that together represent the state of the geographical data quality art, or, more accurately, science. Howard Veregin presents in Chapter 12 an overview of the components of data quality; their interactions and dependencies; and the efforts that have been made in recent years to embed them in standards. From the perspective of the data producer, quality refers to the difference between the actual characteristics of the product and the relevant specifications that define it, or the claims made about it. Information on quality is immensely useful in managing the production process, particularly if the results of quality analysis point back to suspect sources. On the other hand, details of the production process may be of only marginal interest to a potential user of the data, who is concerned solely with whether the data meet particular requirements. Data quality can thus vary from user to user, depending on respective needs; and the effective measurement and documentation of data quality against needs that are often poorly defined can be an immensely complex and frustrating process.

The problems of determining data quality have been further complicated in recent years by the growth of new communication technologies. These have made it far easier for data to be found, accessed, and shared. The user of a geographical dataset may now be many steps removed from the producer. User and producer may be from entirely different backgrounds, with very little in the way of shared terminology or culture. Even if the data are well documented, the lack of effective systems for documentation, in the form of metadata, may leave the user with an incomplete or incorrect understanding of the meaning of the data. For example, if the units of measurement of a variable are not documented, or if the documentation is not transferred to the user, then from the perspective of the user the data are now subject to a further source of inaccuracy. To the user at the end of a long chain of communication, data quality is most

appropriately defined as a measure of the difference between the database's contents and the user's understanding of their true values. The same collection of bits can have different levels of quality, both increasing and decreasing, as it passes from one custodian to another.

In Chapter 13, Peter Fisher discusses alternative models of uncertainty. The traditional scientific concept of measurement error, which accounts for differences between observers or measuring instruments, turns out to be far too simple as a framework for understanding quality in geographical data. Many geographical concepts are incompletely specified – as for example when population density at a point is reported, without specifying the area over which the density was measured – and such incompleteness of specification is an appropriate component of data quality. Many concepts are poorly defined, leading to understandable disagreement between observers. In this context it is useful to distinguish, as Fisher does, between such terms as 'vague', 'fuzzy', and 'probable'. Both fuzzy set theory and probability theory have been found to be useful in modelling uncertainty in geographical data, although their axioms differ in several key respects.

If agreement can be reached on how to measure and express data quality, then such information should be made available to users, preferably by storing it as part of, or in conjunction with, the database. Quality measures that are true of the entire contents are conveniently stored as part of the database metadata, the digital equivalent of documentation. But other quality measures may be true only of parts of the database, such as classes of objects, or individual objects, or even parts of objects, or regions of the study area. In such cases, it is necessary to have 'slots' in the database available to store data quality information in appropriate, meaningful ways. Such slots might take the form of additional data quality attributes of objects, or components of an object class's description; or it might even be necessary to create a complete map of data quality, showing how quality varies across the study area. Thus data quality becomes a significant part of the representation itself.

With adequate information available on data quality, it is possible to determine its effects on the results of GIS analysis, and for decisions made with GIS to reflect the uncertainties present in the base data. This topic of error propagation is the subject of Chapter 14, by Gerard Heuvelink. Several general strategies for error propagation are proposed, at least one of which will be valid in any context. Typically the error propagation is hidden from the user, who sees only a standard GIS function, such as 'compute slope', but is presented with results that include both the requested estimates of slope, and measures of confidence or uncertainty in the results. Software for error propagation is increasingly available in the GIS world, often in the form of specialised 'add-ons' written in a GIS's macro or scripting language.

In the final chapter in this section, Kate Beard and Barbara Buttenfield review techniques that have been developed for visualising uncertainty, issues raised by their use, and problems requiring additional research. Traditional cartographic practice includes remarkably few methods for visualising uncertainty; whether this is because it is difficult to do so within the constraints of map-making, or whether it reflects a human desire to see the world as simpler than it really is, remains a subject of debate. What is beyond doubt, however, is that the continuation of such practices in the world of GIS is both technically and ethically indefensible. The digital world is far more flexible, and Beard and Buttenfield illustrate many of the methods that have been proposed and implemented by the research community. There have been experiments with sound, animation, and use of the third dimension, each with attendant advantages and disadvantages.

Despite such progress on the research front, the issues of dealing with uncertainty remain. GIS has been adopted by individuals and agencies who see its benefits in terms that often include increased accuracy compared to previous methods; yet the data stored in GIS are in most cases no more accurate. Suppose, for example, that the research community were to suggest, on theoretically defensible grounds, that the only effective method for visualising uncertainty would be to present the user with several equally likely versions of how the world might actually look. Uncertainty in soil mapping could be presented by showing ten alternative, equally likely maps of the same area. While this might make perfect sense from the perspective of error theory, it would be almost completely alien to a culture raised on single, apparently exact maps. The problems of coping with uncertainty, and of introducing its effective treatment from a managerial perspective, are the subject of Chapter 45 by Gary Hunter.