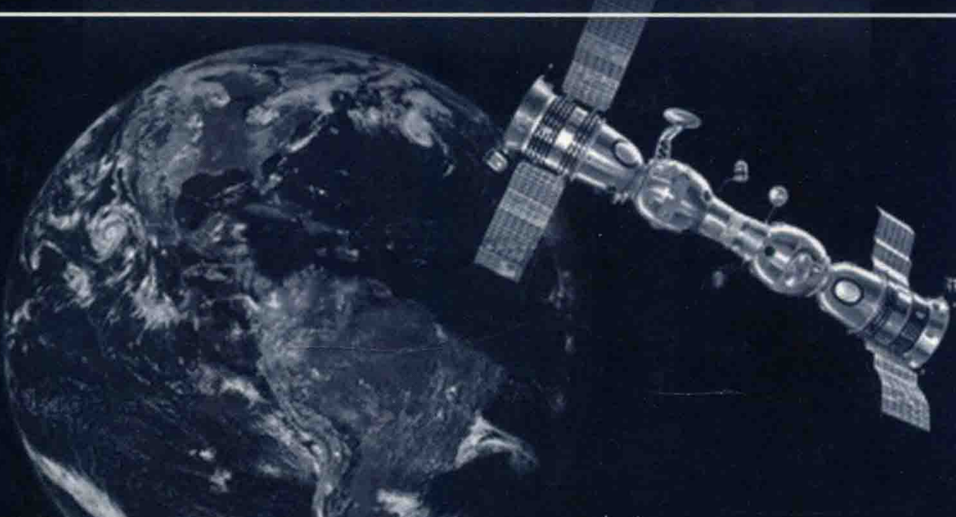
Principles of Modeling Uncertainties in Spatial Data and Spatial Analyses



Wenzhong Shi

CRC Press
Taylor & Francis Group

Principles of Modeling Uncertainties in Spatial Data and Spatial Analyses

Wenzhong Shi

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number: 978-1-4200-5927-4 (Hardback)

Library of Congress Cataloging-in-Publication Data

Shi, Wenzhong.
Principles of modeling uncertainties in spatial data and spatial analysis / Wenzhong Shi.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-4200-5927-4 (alk. paper)
1. Geographic information systems--Data processing. 2. Geographic information systems--Mathematical models. I. Title.

G70.212.S545 2008
910.285--dc22 2008005717

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Principles of Modeling Uncertainties in Spatial Data and Spatial Analyses

Foreword

Michael F. Goodchild

On February 3, 1998 a military aircraft of the U.S. Marine Corps struck and severed the cables supporting a gondola in Cavalese, Italy, leading to the deaths of 20 people. The plane had wing and tail damage but was able to return to its base. It subsequently became clear that the cable-car was not shown on the maps being used by the pilot, and in the subsequent trial the pilot claimed also that the aircraft's height-measuring system had malfunctioned. Errors, omissions, and uncertainties in geographic information do not often result in international incidents of such significance, but it is an inescapable fact that any knowledge of the Earth's surface is subject to uncertainty of some kind, whether it be in the positions of features, their existence, or their description. It is impossible to know the exact location of anything on the Earth's surface, since our methods of measurement are always subject to error; uncertainties creep into our maps, databases, and written records through a wide range of additional mechanisms. Many of the classification schemes used to map aspects of the Earth's surface, from soils to land use, are inherently uncertain, with the result that two people mapping the same area will never produce identical maps. Despite centuries of progress in mapping technology, the creation of geographic information remains as much an art as a science.

From this perspective it is important for users of geographic information to have some awareness of the uncertainties that are likely to influence their decisions—to know what the database does *not* tell them about the real world, and about the reliability of what it *does* tell them. As geographic information technologies have developed over the past few decades it has become clear that one ignores issues of uncertainty at one's peril. For example, regulations in many countries now make it difficult to construct in areas classified as wetland. Decisions are made daily about the uses to which private land can be put. But such decisions are clearly open to legal challenge if they can be shown to have been based on maps that are inherently uncertain.

Research on the description and modeling of uncertainty in geographic information began in earnest in the late 1980s, and has accelerated over the past two decades. Today, a large literature describes successful efforts to address the issue, and tools are increasingly available to allow the effects of uncertainty to be propagated, so that uncertainties can be associated with the results of analysis as well as with the inputs.

John Shi has been one of the leaders in this research area. He has made very significant contributions, particularly in the modeling of uncertainties in geographic features of complex geometry, and has also made a very valuable contribution as the organizer of a series of conferences on spatial data quality, the International Symposia on Spatial Data Quality, that have occurred every 2 years since 1999. The conferences provide a forum for a very broad-based discussion of recent research on uncertainty that is not limited to any single paradigm or theoretical framework, because experience over the past two decades has shown that this problem of uncertainty is so pervasive, and so multidimensional, that no single approach can possibly address it.

The contents of this book provide an excellent introduction to this multidimensional problem. Early research tended to focus on error and accuracy, on the grounds that the creation of geographic information was similar to any problem of scientific measurement, and could be addressed through the application of the theory of errors. Although this is conceptually simple, in reality geographic information tends to have some very awkward properties that complicate the approach enormously. Maps are not collections of independent measurements, but instead represent the culmination of a long and complex process that induces very strong correlations in errors. All positions will be subject to error, but nearby positions will have more similar errors than distant positions—in other words, relative errors of position over short distances tend to be much less than absolute errors. To handle this, models of uncertainty need to incorporate strong spatial autocorrelations and to require comparatively advanced mathematics.

By the mid-1990s, however, it was clear that some aspects of the problem of uncertainty derived in part from the inherent vagueness of definitions, and that these could be handled much more effectively using the theoretical constructs of fuzzy sets. For example, we may not know exactly what is meant by wetland, but nevertheless it may make sense to be able to say that this area is more like wetland than that area. The approach was immediately attractive to many users of geographic information, who found it more intuitive and accessible than the statistical approach.

Both statistical and fuzzy frameworks are covered in this book, which provides a comprehensive overview of the current state of the field. At the same time, it strongly reflects John Shi's own approaches, and the very significant contributions he has made. It should be indispensable reading for anyone interested or actively engaged in this research area, and desirable reading for anyone using geographic information to solve real problems. We have made much progress in the past two decades, and today few users of geographic information systems are willing to assume that their outputs are exact and correct *just because they came from a computer.* Additionally, we are still some distance from the goal of placing a plus or minus on every output, and of incorporating uncertainty into every decision made with geographic information. But this book may help us get closer to that goal.

Preface

This book presents four major theoretical breakthroughs in uncertainty modeling, which have resulted from the writer's investigation of uncertainty modeling in spatial data and spatial analysis. They are (a) advances in spatial object representation, from determinedness to *uncertainty*-based representation of geographic objects in geographic information science (GISci); (b) from uncertainty modeling for *static* spatial data to *dynamic* spatial analyses; (c) from uncertainty modeling for spatial *data* to spatial *models*; and (d) from error *description* of spatial data to spatial data quality *control*.

GISci, compared to classical sciences such as mathematics or physics, is a newly emerged science. Its theoretical foundations have been recently formulated and are in a vital position of readiness for further development. This book intends to contribute to the progress of one fundamental area of GISci—data quality and uncertainty modeling for spatial data and spatial analyses.

GISci covers the following major fundamental areas: (a) space and time, (b) data quality and uncertainty, and (c) spatial analysis. Of these three areas, data quality and uncertainty, area (b) is considered both independently and also in relation to areas (a) and (c).

There are two types of geographic entities and phenomena in the real world: the precise and the uncertain. Cognition and representation of space and time, area (a), is considered not only in relation to precise spatial objects, but also in relation to uncertain spatial objects. Uncertainty in spatial representation is unavoidable, as it is not yet possible to represent the infinite natural world in a finite way. All measurements contain a certain degree of error, no matter how high the accuracy of modern measurement technologies. Thus, it is to be expected that, given the infancy of GISci, the currently used determined approaches must have the potential for further development to encompass uncertain-based approaches in cognition and representation of space and time.

Uncertainty modeling and quality control are two key issues in area (b). Uncertainty modeling depicts the error in spatial data, and uncovers the nature of uncertainty distribution. Quality control aims to control the overall quality of spatial data, with the possibility of reducing the spatial data error to a desired level.

Spatial analysis in GISci area (c) is, likewise, not uncertainty free. Uncertainty in source data and limitations of spatial analysis models may propagate, or any original uncertainties may become amplified. The quality of a decision, based on a spatial analysis, is affected by the quality of the original data, the quality of the spatial analysis model, and the degree of uncertainty that is propagated or generated in the spatial analysis.

A framework for handling uncertainties in spatial data and spatial analyses is outlined in this book. Covered are uncertainties in the real and natural worlds, uncertainties in the cognition of natural objects, modeling errors in spatial object measurement, modeling uncertainty in spatial models, uncertainty propagation in spatial analyses, quality control for spatial data, and finally a presentation of uncertainty information. Theories and methods for handling these uncertainties are given and, as such, form principles for modeling uncertainties in spatial data and spatial analyses.

In line with the logic of uncertainty generation and handling outlined above, this book is organized in seven sections: (I) overview, (II) modeling uncertainties in spatial data, (III) modeling uncertainties in spatial model, (IV) modeling uncertainties in spatial analyses, (V) quality control of spatial data, (VI) presentation of data quality information, and (VII) epilogue. Within each section are several chapters. The outline of the contents of the seven parts, listed above, is as follows.

Section I (Chapters 1–3) provides an overview of the principles of modeling uncertainty of spatial data and spatial analysis, with the principles of uncertainty in general together with the concepts of uncertainty in spatial data and analyses introduced in Chapter 1. Various uncertainty sources are summarized in Chapter 2. They include uncertainty inherited in the natural and real worlds, uncertainty due to human cognition, uncertainty introduced in data capture processes, and uncertainty arising from spatial data processing and spatial analyses. Mathematical foundations of uncertainty modeling in spatial data and analyses are summarized in Chapter 3. They include probability theory, statistics, evidence theory, fuzzy set and fuzzy topology, and rough set theory, as well as information theory and entropy.

The developed methods for handling uncertainties in spatial data are introduced in Section II (Chapters 4 to 6). Uncertainties in spatial data can be classified into the following five types: (a) positional uncertainty, (b) attribute uncertainty, (c) temporal uncertainty, (d) incompleteness, and (e) logic inconsistency. Since temporal information can be regarded as a form of attribute information, the methods for handling attribute uncertainty are potentially applicable to handling temporal uncertainty. Positional uncertainty modeling is the focus of Chapter 4 while attribute uncertainty modeling is introduced in Chapter 5. Uncertainty modeling for integrated positional and attribute uncertainty, an issue which is critical in multisource data integration, is introduced in Chapter 6.

Error models for positional uncertainty have been comprehensively studied and are presented systematically, as uncertainty models for points, line segments, polylines, curves, polygons, and area objects in Chapter 4. Line feature errors are modeled through (a) the confidence region, (b) error distribution, and (c) error of points on the lines with correlated and independent cases. Of these error models, the confidence region model of a line segment forms the corner stone of positional error modeling of objects in object-based GIS.

As previously indicated, geographic phenomena can fall into two classes: the determine-based and the uncertain-based. Previous solutions have mainly provided determine-based representation of space and time in Euclidean space, where determined points, lines, areas, and volumes are described. Uncertainty modeling of spatial data presented in Chapter 4 forms a basis for uncertainty-based representation of the geographic entities covering points, line segments, polylines, curve lines, and polygons.

The properties of attribute uncertainty and the methods to model the attribute uncertainties in GIS and remote sensing are introduced in Chapter 5. These methods include the rate of defect model, the use of which is straightforward for practical applications; the probability vector models, which can indicate uncertainty spatial distribution; and the commonly used error matrix method.

The method for integrating positional and attribute uncertainties is introduced in Chapter 6. It should be noted that modeling the integrated error is a critical issue in multisource data integration. The theoretical model "S-band" and the two solutions of the model are given: (a) probability theory-based solution and certainty factor-based solution.

Modeling uncertainties in spatial models is introduced in Section III (Chapters 7 and 8). Modeling uncertain topological relationships is covered in Chapter 7. The concepts for topological relationships between objects in GIS and modeling the topological relationships are introduced. In addition, fuzzy topology is adapted to model uncertain topological relationships between spatial objects in GIS. Geographic objects can, in fact, be represented by either spatial data or spatial models. Two typical examples of a spatial model are the digital elevation model (DEM) and curve functions in GIS. Natural terrains in the natural world can be represented by digital elevation models. In this case, the model can be either regular, such as a regular tessellation like a square, or irregular, such as a triangulated irregular network (TIN). A curve entity in the natural world—for instance, a curved portion of a road— can be represented by either a regular curve, such as circular curve, or an irregular curve, approximated by the third-order spline curve function.

Since error models for a curve are introduced in Chapter 4 (Section II), Chapter 8 concentrates mainly on handling model error in DEM. Two advances have been made along the lines of DEM model accuracy estimation: (a) a formula to estimate the average model accuracy of a TIN, and (b) accuracy estimation of the bicubic interpolation model.

Modeling uncertainties in spatial analyses is the focus of Section IV (Chapters 9 to 11). Here, spatial analyses mainly refer to GIS spatial analyses, such as overlay analysis, buffer analysis, and line simplification analysis. Each analysis is a transformation, based on one or more original spatial data set(s). Uncertainty inherited from the original data set(s) are further propagated or even amplified through such a spatial analysis. In many cases, new uncertainties are also generated. In addition to modeling uncertainty in static spatial data, a logical further step is to model uncertainty in spatial analyses.

Covered in Chapter 9 is modeling uncertainties in overlay spatial analyses where analytical and simulation approaches for determining uncertainty propagated through an overlay analysis are given. Modeling uncertainty in buffer spatial analysis is covered in Chapter 10. A solution is provided for quantifying uncertainty propagation through a buffer spatial analysis, where four error indicators are proposed for quantifying the uncertainties. In Chapter 11, the newly proposed uncertainty modeling method for line simplification spatial analysis is given.

Focus in the first chapters is on descriptions of uncertainties, including descriptions of uncertainty in spatial data, descriptions of uncertainty in spatial analysis, and descriptions of uncertainty in spatial models. To ensure an understanding of handling uncertainty, description is a necessary first step; a further step is to control or even reduce the uncertainties in the spatial data, analyses, or models, if possible.

Therefore, Section V introduces another theoretical breakthrough in handling uncertainties in spatial data. This is from uncertainty description to spatial data quality control. In this regard, the quality control for spatial data is introduced. An explanation of quality control for object-based spatial data is given in Chapter 12. The aim is to control the overall geometric quality of vector spatial data by the least squares adjustment method with an example for vector cadastral data. Quality control for field-based spatial data—the aim being to geometrically rectify high resolution satellite imageries by point- and line-based transformation models—is presented in Chapter 13. Quality control for digital elevation models, designed to improve DEM accuracy with the newly proposed hybrid interpolation model and bidirectional model, is described in Chapter 14.

The methods for presenting data quality information are described in Section VI (Chapters 15 to 17). Visualization techniques for uncertainty information are provided in Chapter 15. Examples given are the error ellipse, arrow-based approach, grayscale map-based, color map, symbol-based approach, and the three-dimensional visualization approach. An object-oriented metadata system for managing metadata on uncertainty information is introduced in Chapter 16. A Web service-based data quality information solution for disseminating uncertainty information on the Internet is introduced in Chapter 17.

Finally, in Section VII, (Chapter 18), a summary and comments on principles of modeling uncertainties in spatial data and spatial analyses are given. Suggestions for further development in the field of modeling uncertainties in spatial data and analyses are outlined.

Wenzhong Shi
Hong Kong

Acknowledgments

I would like to thank my family, Wendy, Julia, and Daniel, for their support in writing this book. I would like also extend my thanks to my colleagues, Tracy, Yan, Kim, Elaine, Jacqueline, Eryong, Xiaohna, Yumin, and many others for their assistance.

The work described in this book was supported by the Young Professor Scheme of The Hong Kong Polytechnic University; the Distinguished Young Scholar Fund (B) from the National Natural Science Foundation, China (No. 40629001), the CRC Program of Hong Kong Research Grants Council (No. 3_ZB40), and the Chang Jiang Scholars Program, Ministry of Education, China.

Wenzhong Shi
Hong Kong

Contents

SECTION I Overview

SECTION II Modeling Uncertainties in Spatial Data

SECTION III Modeling Uncertainties in Spatial Model