Quality Aspects in Spatial Data Mining

Edited by
Alfred Stein
Wenzhong Shi
Wietske Bijker

CRC Press
Taylor & Francis Group

Quality Aspects in Spatial Data Mining

Edited by
Alfred Stein
Wenzhong Shi
Wietske Bijker

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

Version Date: 20121112

International Standard Book Number-13: 978-1-4200-6927-3 (eBook - PDF)

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

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

Qualitas est nobilior quantitate. Qualitas, non quantitas.

Sêneca, *Epistulae Morales* **17.4**

Quality in a product or service is not what the supplier puts in. It is what the customer gets out and is willing to pay for. A product is not quality because it is hard to make and costs a lot of money, as manufacturers typically believe. This is incompetence. Customers pay only for what is of use to them and gives them value. Nothing else constitutes quality.

Peter Drucker

It is not a question of how well each process works, the question is how well they all work together.

Lloyd Dobyns and Clare Crawford-Mason, *Thinking About Quality*

Contents

SECTION IV Applications

SECTION V Communication

Foreword

Quality Aspects in Spatial Data Mining, edited by Alfred Stein, Wenzhong Shi, and Wietske Bijker, and published by CRC Press is a highly impressive collection of chapters that address many of the problems that lie on the frontiers of spatial data mining, classification, and signal processing. The sections are authoritative and up to date. The coverage is broad, with subjects ranging from systems approaches to spatial data quality; quality of descriptions of socially constructed facts, especially legal data, in a GIS; and a multicriteria fusion approach for geographical data matching, to quality-aware and metadata-based decision-making support for environmental health, geostatistical texture classification of tropical rainforests in Indonesia, and formal languages for expressing data consistency rules and implications for reporting of quality metadata.

The wealth of concrete information in *Quality Aspects of Spatial Data Mining* makes it clear that in recent years substantial progress has been made toward the development of effective techniques for spatial information processing. However, there is an important point that has to be made.

Science deals not with reality but with models of reality. As we move further into the age of machine intelligence and automated reasoning, models of information systems, including spatial information systems, become more complex and harder to analyze. An issue that moves from the periphery to the center is that of dealing with information that is imprecise, uncertain, incomplete, and/or partially true. What is not widely recognized is that existing techniques, based as they are on classical, bivalent logic, are incapable of meeting the challenge. The problem is that bivalent logic is intrinsically unsuited for meeting the challenge because it is intolerant of imprecision and partiality of truth.

So what approach can be used to come to grips with information, including spatial information, that is contaminated with imprecision, uncertainty, incompleteness, and/or partiality of truth? A suggestion that I should like to offer is to explore the use of granular computing. Since granular computing is not a well-known mode of computation, I will take the liberty of sketching in the following its underlying ideas.

In conventional modes of computation, the objects of computation are values of variables. In granular computing, the objects of computation are not values of variables but the information about values of variables, with the information about values of variables referred to as granular values. When the information is described in a natural language (NL), granular computing reduces to NL computation. An example of granular values of age is young, middle-aged, and old. An example of a granular value of imprecisely known probability is not very low and not very high.

How can a granular probability described as "not very low and not very high" be computed? This is what granular computing is designed to do. In granular computing, the key to computation with granular values is the concept of a generalized constraint. The concept of a generalized constraint is the centerpiece of granular computing.

The concept of a constraint is a familiar one in science. But, in science, models of constraints tend to be oversimplified in relation to the complexity of real-world constraints. In particular, constraints are generally assumed to be hard, with no elasticity allowed. A case in point is the familiar sign "Checkout time is 1 p.m." This constraint appears to be hard and simple but in reality it has elasticity that is hard to define.

A fundamental thesis of granular computing is that information is, in effect, a generalized constraint. In this nontraditional view of information, the traditional statistical view of information is a special case.

The concept of a generalized constraint serves two basic functions: (a) representation of information and, in particular, representation of information that is imprecise, uncertain, incomplete, and/or partially true; and (b) computation/deduction with information represented as a system of generalized constraints. In granular computing, computation/deduction involves propagation and counterpropagation of generalized constraints. The principal rule of deduction is the so-called extension principle. A particularly important application area for granular computing is computation with imprecise probabilities. Standard probability theory does not offer effective techniques for this purpose.

What I said above about granular computing in no way detracts from the importance of the contributions in *Quality Aspects of Spatial Data Mining.* I have taken the liberty of digressing into a brief discussion of granular computing because of my perception that granular computing is a nascent methodology that has high potential relevance to spatial information processing — and especially to processing of information that is imprecise, uncertain, incomplete, and/or partially true — a kind of information that the spatial information systems community has to wrestle with much of the time.

In conclusion, *Quality Aspects of Spatial Data Mining* is an important work that advances the frontiers of spatial information systems. The contributors, the editors, and the publisher deserve our thanks and loud applause.

Lotfi Zadeh
Berkeley, California

Contributing Authors

Tamiru H. Alemseged
Department of Water Resources
ITC
Enschede, The Netherlands

Karl-Heinrich Anders
Institute of Cartography and Geoinformatics
Leibniz Universität Hannover
Hannover, Germany

Mohamed Bakillah
Département des Sciences Géomatiques
Centre de Recherche en Géomatique
Université Laval
Québec City, Québec, Canada

Yvan Bédard
Département des Sciences Géomatiques
Centre de Recherche en Géomatique
Université Laval
Québec City, Québec, Canada

Sarah A. Bekessy
School of Global Studies, Social Science and Planning
RMIT University
Melbourne, Australia

Wietske Bijker
Department of Earth Observation Science
ITC
Enschede, The Netherlands

Anna T. Boin
Department of Geomatics
Cooperative Research Centre for Spatial Information
University of Melbourne
Coburg, Australia

Jean Brodeur
Centre d'Information Topographique de Sherbrooke
Sherbrooke, Québec, Canada

Alan Brown
Countryside Council for Wales
Bangor, United Kingdom

James Brown
National Weather Service
NOAA
Silver Spring, Maryland, U.S.A.

Gilberto Câmara
Image Processing Division
National Institute for Spatial Research
São José dos Campos, Brazil

René R. Colditz
German Aerospace Center
German Remote Sensing Data Center
Wessling, Germany
and
Department of Geography
Remote Sensing Unit
University of Wuerzburg
Wuerzburg, Germany

Alexis J. Comber
Department of Geography
University of Leicester
Leicester, United Kingdom

Christopher Conrad
Department of Geography
Remote Sensing Unit
University of Wuerzburg
Wuerzburg, Germany

Robert Corner
Department of Spatial Sciences
Curtin University
Bentley, Western Australia

Sytze de Bruin
Centre for Geo-Information
Wageningen University
Wageningen, The Netherlands

Stefan Dech
German Aerospace Center
German Remote Sensing Data Center
Wessling, Germany
and
Department of Geography
Remote Sensing Unit
University of Wuerzburg
Wuerzburg, Germany

Eduardo S. Dias
SPINlab
Vrije Universiteit
Amsterdam, The Netherlands

Alistair J. Edwardes
Department of Geography
University of Zurich
Zurich, Switzerland

Pete F. Fisher
Department of Information Science
City giCentre
City University
London, United Kingdom

Andrew Frank
Institute of Geoinformation and Cartography
Technical University of Vienna
Vienna, Austria

Richard Gloaguen
Remote Sensing Group
Institute for Geology
TU-Bergakademie
Freiberg, Germany

Michael F. Goodchild
Department of Geography
National Center for Geographic Information and Analysis
University of California, Santa Barbara
Santa Barbara, California, U.S.A.

Abdelbasset Guemeida
Laboratoire Sciences et Ingénierie de l'Information et de l'Intelligence Stratégique
Université de Marne-la-Vallée
Marne-la-Vallée, France

Jan-Henrik Haunert
Institute of Cartography and Geoinformatics
Leibniz Universität Hannover
Hannover, Germany

Felix Hebeler
Department of Geography
University of Zurich
Zurich, Switzerland

Gerard Heuvelink
Wageningen University and Research Centre
Wageningen, The Netherlands
and
Alterra – Soil Science Centre
Wageningen, The Netherlands

Gary J. Hunter
Department of Geomatics
Cooperative Research Centre for Spatial Information
University of Melbourne
Parkville, Australia

Kerstin Huth
Faculty of Geomatics
Karlsruhe University of Applied Sciences
Karlsruhe, Germany

Robert Jeansoulin
Laboratoire d'Informatique de l'Institut Gaspard Monge
Université Paris-EST Marne-la-Vallée
Champs-sur-Marne, France

Simon D. Jones
School of Mathematical and Geospatial Sciences
RMIT University
Melbourne, Australia

Martin Knotters
Alterra – Soil Science Centre
Wageningen, The Netherlands

Karin Kollo
Department of Geodesy
Estonian Land Board
Tallinn, Estonia

Alex M. Lechner
School of Mathematical and Geospatial Sciences
RMIT University
Melbourne, Australia

Rodrigo Lilla Manzione
National Institute for Spatial Research
Image Processing Division
São José dos Campos, Brazil

Haixia Mao
Department of Land Surveying and Geo-Informatics
Advanced Research Centre for Spatial Information Technology
The Hong Kong Polytechnic University
Hong Kong SAR, China

Marco Marinelli
Department of Spatial Sciences
Curtin University
Bentley, Western Australia

Prashanth R. Marpu
Remote Sensing Group
Institute for Geology
Freiberg, Germany

Nick Mitchell
Faculty of Geomatics
Karlsruhe University of Applied Sciences
Karlsruhe, Germany

Mir Abolfazl Mostafavi
Département des Sciences Géomatiques
Centre de Recherche en Géomatique
Université Laval
Québec City, Québec, Canada

Gerhard Navratil
Institute for Geoinformation and Cartography
Vienna University of Technology
Vienna, Austria

Ana-Maria Olteanu
COGIT Laboratory
IGN/France
Paris, France

Alejandro Pauly
Sage Software
Alachua, Florida, U.S.A.

Ekaterina S. Podolskaya
Cartographic Faculty
Moscow State University of Geodesy and Cartography
Moscow, Russia

Ross S. Purves
Department of Geography
University of Zurich
Zurich, Switzerland

Tom H.M. Rientjes
Department of Water Resources
ITC
Enschede, The Netherlands

Gabriella Salzano
Laboratoire Sciences et Ingénierie de l'Information et de l'Intelligence Stratégique (S3IS)
Université de Marne-la-Vallée
Paris, France

Gertrud Schaab
Faculty of Geomatics
Karlsruhe University of Applied Sciences
Karlsruhe, Germany

Sven Schade
Institute for Geoinformatics
University of Münster
Münster, Germany

Michael Schmidt
German Aerospace Center
German Remote Sensing Data Center
Wessling, Germany
and
Remote Sensing Unit
Department of Geography
University of Wuerzburg
Wuerzburg, Germany

Markus Schneider
Department of Computer & Information Science & Engineering
University of Florida
Gainesville, Florida, U.S.A.

Monika Sester
Institute of Cartography and Geoinformatics
Leibniz Universität Hannover
Hannover, Germany

Wenzhong Shi
Department of Land Surveying and Geo-Informatics
Advanced Research Centre for Spatial Information Technology
The Hong Kong Polytechnic University
Hong Kong SAR, China

Alfred Stein
Department of Earth Observation Science
ITC
Enschede, The Netherlands

Rangsima Sunila
Department of Surveying
Laboratory of Geoinformation and Positioning Technology
Helsinki University of Technology
Espoo, Finland

Yan Tian
Department of Land Surveying and Geo-Informatics
Advanced Research Centre for Spatial Information Technology
The Hong Kong Polytechnic University
Hong Kong SAR, China
and
Department of Electronic and Information Engineering
Huazhong University of Science and Technology
Wuhan, China

Martin Vermeer
Department of Surveying
Helsinki University of Technology
Helsinki, Finland

Jos von Asmuth
Kiwa Water Research
Nieuwegein, The Netherlands

Paul Watson
1Spatial
Cambridge, United Kingdom

Thilo Wehrmann
German Aerospace Center
German Remote Sensing Data Center
Wessling, Germany

Arief Wijaya
Remote Sensing Group
Institute for Geology
TU-Bergakademie
Freiberg, Germany
and
Faculty of Agricultural Technology
Gadjah Mada University
Yogyakarta, Indonesia

Graeme Wright
Department of Spatial Sciences
Curtin University
Bentley, Western Australia

Lotfi A. Zadeh
University of California
Berkeley, California, U.S.A.

Introduction

ABOUT THIS BOOK

Spatial data mining, sometimes called image mining, is a rapidly emerging field in Earth observation studies. It aims at identification, modeling, tracking, prediction, and communication of objects on a single image, or on a series of images. All these steps have to deal with aspects of quality. For example, identification may concern uncertain (vague) objects, and modeling of objects relies, among other issues, on the quality of the identification. In turn, tracking and prediction depend on the quality of the model. Finally, communication of uncertain objects to stakeholders requires a careful selection of tools.

Quality of spatial data is both a source of concern for the users of spatial data and a source of inspiration for scientists. In fact, spatial data quality and uncertainty are two of the fundamental theoretical issues in geographic information science. In both groups, there is a keen interest to quantify, model, and visualize the accuracy of spatial data in more and more sophisticated ways. This interest was at the origin of the 1st International Symposium on Spatial Data Quality, which was held in Hong Kong in 1999, and still is the very reason for the 5th symposium, ISSDQ 2007, in Enschede, The Netherlands. The organizers of this symposium selected the best papers presented at the conference to be published in this book after peer-review and adaptation.

DATA QUALITY—A PERSPECTIVE

The quality of spatial data depends on "internal" quality, the producer's perception, and "external quality," or the perspective of the user. From the producer's point of view, quality of spatial data is determined by currency, geometric and semantic accuracy, genealogy, logical consistency, and the completeness of the data. The user's concern, on the other hand, is "fitness for use," or the level of fitness between the data and the needs of the users, defined in terms of accessibility, relevancy, completeness, timeliness, interpretability, ease of understanding, and costs (Mostafavi, Edwards, and Jeansoulin, 2004).

The field of spatial data quality has come a long way. Five hundred years ago, early mapmakers like Mercator worried already about adequate representation of sizes and shapes of seas and continents to allow vessel routing. Mercator's projection allowed representing vessel routes as straight lines, which made plotting of routes easier and with greater positional accuracy. Ever since, surveyors, cartographers, users, and producers of topographic data have struggled to quantify, model, and increase the quality of data, where accuracy went hand in hand with fitness for use. Next to navigation, description of property, from demarcation of countries to cadastre of individual property, became an important driving force behind the quality of

spatial data in general, with emphasis on positional accuracy and correct labeling of objects (e.g., ownership).

In the environmental sciences, the focus on aspects of quality of spatial data differed from the topographic sciences. Of course soils, forests, savannahs, ecosystems, and climate zones needed to be delineated accurately, but acceptable error margins were larger than in the topographic field. Attention was focused on the adequate and accurate description of the content. Well-structured, well-described legends became important, and statistical clustering techniques such as canonical analysis were used to group observations into classes. With a trend toward larger scales (higher spatial resolution), the positional accuracy became more important for the environmental sciences for adequate linking and analyzing of data of different sources, while the need for thematic accuracy and thematic detail increased in the topographic sciences. Thematic and positional accuracy became increasingly correlated.

For a long time scientists have realized that, in reality, objects weren't always defined by sharp boundaries and one class of soils or vegetation will change gradually into another in space as well as in time. Nevertheless, because of a lack of appropriate theory and appropriate tools, everything had to be made crisp for analysis and visualization. In the last decade or so, theories for dealing with vague objects and their relations have been developed (Dilo et al., 2005), such as fuzzy sets, the egg-yolk model (Cohn and Gotts, 1996), the cloud model (Cheng et al., 2005 citing Li et al., 1998) and uncertainty based on fuzzy topology (Shi and Liu, 2004).

The way we look at our world, and the way we define objects from observations, depend on the person, background, and purpose. One remotely sensed image, one set of spatial data, can be a source for many different interpretations. Of course there are a number of common perceptions in society that enable us to communicate spatial information. These common perceptions change with time as the challenges society faces change. A look at a series of land cover maps from the same area but from different decades clearly shows how thinking went from "exploration" and "conservation" to "multiple-use" and the legend and the spatial units changed accordingly, even where no changes happened on the ground. This is where ontology plays a role.

During times when spatial data were scarce, a limited number of producers produced data for a limited well-known market of knowledgeable users with whom they had contact. Now there are many producers of spatial data; some are experts, others are not. Users have easy access to spatial data. Maps and remote sensing images are available in hard copy and via the Internet in ever-growing quantities. Producers have no contact with all users of their data. Spatial data are also easily available to users for whom the data were not intended (fitness for use!) and to nonexpert users, who do not know all the ins and outs of the type of data. Not all producers of spatial data are experts either; yet, their products are freely available. A good example is Google Earth and Google Maps, where everyone with access to the Internet can add information to a specific location and share this with others. The increasing distance between producer and user of spatial data calls for adequate metadata, including adequate descriptions of data accuracy in terms that are relevant to both the user and producer of the data.

This book addresses quality aspects in spatial data mining for the whole flow from data acquisition to the user. A systematic approach for handling uncertainty

and data quality issues in spatial data and spatial analyses covers understanding the sources of uncertainty, and modeling positional, attribute, and temporal uncertainties and their integration in spatial data as well as modeling uncertainty relations and completeness errors in spatial data, in both object-based and field-based data sets. Such types of approaches can be found as Section I, "Systems Approaches to Spatial Data Quality." Besides modeling uncertainty for spatial data, modeling uncertainty for spatial models is another essential issue, such as accuracy in DEM. Section II, "Geostatistics and Spatial Data Quality for DEMs," deals specifically with this aspect of data quality. Uncertainties may be propagated or even amplified in spatial analysis processes, and, therefore, uncertainty propagation modeling in spatial analyses is another essential issue, which is treated in more detail in Section III, "Error Propagation."

Quality control for spatial data and spatial analyses should ensure the information can fulfill the needs of the end users. For inspiration to users and producers alike, practical applications of quality aspects of spatial data can be found in Section IV, "Applications." New concepts and approaches should prove their worth in practice. Questions from users trigger new scientific developments. Just like the need to represent routes by straight lines on maps inspired Mercator to develop a map projection, present-day users inspire scientists to answer their questions with innovative solutions, which in turn give rise to more advanced questions, which could not be asked previously.

From a known user, one can get specifications of the data quality that are needed. But what to do with the (yet) unknown users, who may use the data for unforeseen purposes, or the "non-users" or "not-yet users" (Pontikakis and Frank, 2004), from whom we would like to know why they are not using spatial information? Section V, "Communication," focuses on ways to communicate with users about their needs and the quality of spatial data.

ACKNOWLEDGMENTS

This book emerged from a symposium, consisting of presentations, proceedings, and a social program. Prior to the conference we organized a very careful review process for all the papers. At this stage, we thank the reviewers, who were indispensable to having this book reach the standard that it has at the moment: Rolf de By, Rodolphe Devillers, Pete Fisher, Andrew Frank, Michael Goodchild, Nick Hamm, Geoff Hennebry, Gerard Heuvelink, Gary Hunter, Robert Jeansoulin, Wu Lun, Martien Molenaar, Mir Abolfazl Mostavafi, and David Rossiter.

We realize very well that any symposium has its support. At this stage, we would like to thank the ITC International Institute for Geo-Information Science and Earth Observation for hosting this meeting and for all its support. In particular, we thank Saskia Tempelman, Rens Brinkman, Janneke Kalf, Harald Borkent, Frans Gollenbeek, and many others. Without their input the meeting would not have been possible. The International Society for Photogrammetry and Remote Sensing (ISPRS) actively participated in getting the symposium organized, and we thank them for the support given. Finally, we thank the sponsors of the meeting:

CRC Press/Taylor & Francis, the CTIT Research School at Twente University, the PE&RC Research School based at Wageningen University, and the Dutch Kadaster and Geoinformatics Netherlands.

Alfred Stein, Wenzhong Shi, and Wietske Bijker

REFERENCES

Cheng, T., Z. Li, M. Deng, and Z. Xu. 2005. Representing indeterminate spatial object by cloud theory. In: L. Wu, W. Shi, Y. Fang, and Q. Tong (Eds.), *Proceedings of the 4th International Symposium on Spatial Data Quality*, 25th to 26th August 2005, Beijing, China, The Hong Kong Polytechnic University.

Cohn, A. G. and N. M. Gotts. 1996. Geographic objects with indeterminate boundaries, chapter The "egg-yolk" representation of regions with indeterminate boundaries. In: Burrough and Frank (eds.), *GISDATA*, 171–187.

Dilo, A., R. A. de By, and A. Stein. 2005. A proposal for spatial relations between vague objects. In: L. Wu, W. Shi, Y. Fang, and Q. Tong (Eds.), *Proceedings of the 4th International Symposium on Spatial Data Quality*, 25th to 26th August 2005, Beijing, China. The Hong Kong Polytechnic University.

Li, D., D. Cheung, X. Shi, and D. Ng. 1998. Uncertainty reasoning based on cloud model in controllers. *Computers Math. Application*, 35, pp. 99–123.

Mostafavi, M. A., G. Edwards, and R. Jeansoulin. 2004. An ontology-based method for quality assessment of spatial databases. In: A. U. Frank and E. Grum (compilers), *Proceedings of the ISSDQ '04*, Vol. 1. Geo-Info 28a, pp. 49–66. Dept. for Geoinformation and Carthography, Vienna University of Technology.

Pontikakis, E. and A. Frank, 2004. Basic spatial data according to users' needs: Aspects of data quality. In: A. U. Frank and E. Grum (compilers), *Proceedings of the ISSDQ '04*, Vol 1. Geo-Info 28a, pp. 13–29. Dept. for Geoinformation and Carthography, Vienna University of Technology.

Shi, W. Z. and K. F. Liu, 2004. Modeling fuzzy topological relations between uncertain objects in GIS, *Photogrammetric Engineering and Remote Sensing*, *70*(8), pp. 921–929.

Section I

Systems Approaches to Spatial Data Quality

INTRODUCTION

Spatial data quality is a concept that is partly data- and object-driven and partly based on fitness for use. In order to integrate, the systems approach is likely to be useful. A systems approach is well known in geo-information science (one may think of the GEOSS initiative) as well as in several other fields of science, like agriculture, economy, and management sciences. Its approach thus serves as a guiding principle for spatial data quality aspects. For spatial data, geographical information systems found their place in the 1980s, and these systems are still potentially useful to serve the required purposes. But here the word "system" largely expresses the possibilities of storing, displaying, handling, and processing spatial data layers. This is not sufficient for the emerging field of spatial data quality, requiring in its current development a full systems approach. In fact, data can be different as compared to previously collected and analyzed data, and the objects will be inherently uncertain. As compared to the traditional GIS, a systems approach to spatial data quality should be able to deal with uncertainties. These uncertainties are usually expressed either by statistical measures, by membership functions of fuzzy sets, or they are captured by metadata.

A first and foremost challenge is thus to be able to extract, i.e., to query, vague spatial objects from databases. Common GIS, still seen as a spatial database with some specific functionalities, do not allow one to do so. This field is, at the moment, therefore, still very much an area of research rather than an issue of production. As concerns the data aspect, socially constructed facts are recognized as being important. This refers in part to social objects, but also to legal facts.

More recently, semantic issues have found their place in spatial research, thus acknowledging that the traditional fuzzy and statistical measures may fall short. Modern and prospective approaches toward spatial data quality are thus governed by semantic aspects of data and maps. In the frame of this section, semantic issues are approached along two lines. First, a conceptual framework for quality assessment is presented. Such a framework may be different from the ordinary conceptual frameworks, which did not include data quality aspects explicitly. In this sense, one chapter considers semantic mapping between ontologies. Next it is recognized that a semantic reference system should account for uncertainty. A requirement analysis is thus appropriate in that sense.

Section I of the book thus considers modern aspects of a systems approach to spatial data quality.

1 Querying Vague Spatial Objects in Databases with VASA

Alejandro Pauly and Markus Schneider

CONTENTS

1.1 INTRODUCTION

Many man-made spatial objects such as buildings, roads, pipelines, and political divisions have a clear boundary and extension. In contrast to these crisp spatial objects, most naturally occurring spatial objects have an inherent property of vagueness or indeterminacy of their extension or even of their existence. Point locations may not be exactly known; paths or trails might fade and become uncertain at intervals. The boundary of regions might not be certainly known or simply not be as sharp as that of a building or a highway. Examples are lakes (or rivers) whose extensions (or paths) depend on pluvial activity, or the locations of oil fields that in many cases can only be guessed. This inherent uncertainty brings to light the necessity of more adequate models that are able to cope with what we will refer to as *vague spatial objects.*

Existing implementations of geographic information systems (GIS) and spatial databases assume that all objects are crisply bounded. With the exception of a few domain-specific solutions, the problem of dealing with spatial vagueness has no widely accepted practical solution. Instead, different conceptual approaches exist for

which researchers have defined formal models that can deal with a closer approximation of reality where not all objects are crisp. For the treatment of vague spatial objects, our *vague spatial algebra* (*VASA*), which can be embedded into databases, encompasses data types for *vague points*, *vague lines*, and *vague regions* as well as for all operations and predicates required to appropriately handle objects of these data types. The central goal of the definition of VASA is to leverage existing models for crisp spatial objects, resulting in robust definitions of vague concepts derived from proven crisp concepts.

In order to fully exploit the power of VASA in a database context, users must be able to pose significant queries that will allow retrieval of data that are useful for analysis. In this chapter, we provide an overview of VASA and the capabilities it provides for handling vague spatial objects. Based on these capabilities, we describe how users can take full advantage of an implementation of VASA by proposing meaningful queries on vague spatial objects. We use sample scenarios to explain how the queries can be posed with a moderate extension of SQL.

This chapter starts in Section 1.2 by summarizing related work that covers relevant concepts from crisp spatial models as well as other concepts for handling spatial vagueness. In Section 1.3 we introduce the VASA concepts for data types, operations, and predicates. Section 1.4 shows how a simple extension to SQL will be of great benefit when querying vague spatial data. Finally, in Section 1.5 we derive conclusions and expose future work.

1.2 RELATED WORK

Existing concepts relevant to this work can be divided into two categories: (1) concepts that provide the foundation for the work presented in this chapter and (2) concepts that are defined with goals similar to those of the work in this chapter.

Related to the former, we are interested in crisp spatial concepts that define the crisp spatial data types for *points, lines,* and *regions* [25]. We are also interested in the relationships that can be identified between instances of these types. Topological relationships between spatial objects have been the focus of much research, and we concentrate on the concepts defined by the *9-intersection model* originally defined in [10] for simple regions, and later extended for simple regions with holes in [11]. The complete set of topological relationships for all type combinations of complex spatial objects is defined in [25] on the basis of the 9-intersection model.

We categorize available concepts for handling spatial vagueness by their mathematical foundation. Approaches that utilize existing exact (crisp) models for spatial objects include the *broad boundaries* approach [6, 7], the *egg-yolk* approach [8], and the *vague regions* concept [12]. These models extend the common assumption that boundaries of regions divide the plane into two sets (the set that belongs to the region, and the set that does not) with the notion of an intermediate set that is not known to certainly belong or not to the region. Thus we say that these models extend crisp models that operate on the Boolean logic (*true*, *false*) into models that handle uncertainty with a three-valued logic (*true*, *false*, *maybe*). *VASA,* our concept for handling spatial vagueness (Section 1.3), is based on exact models for crisp spatial objects. Although fundamentally different from the exact-based approaches, rough

set theory [22] provides tools for deriving concepts with a close relation to what can be achieved with exact models. Rough set theory–based approaches include early work by Worboys in [26], the concepts for deriving *quality measures* presented in [4], and the concept of *rough classification* in [1].

One of the advantages of fuzzy set theory is the ability to handle *blend-in* type boundaries (such as that between a mountain and a valley). Approaches in this category include earlier *fuzzy regions* [3]; the formal definition of *fuzzy points*, *fuzzy lines*, and *fuzzy regions* in [23]; and an extension of the rough classification from [1] to account for fuzzy regions [2]. A recent effort for the definition of a *spatial algebra* based on fuzzy sets is presented in [9]. Finally, probabilistic approaches [13] focus on an *expected* membership to an object that can be contrasted to the membership values of fuzzy sets that are objective in the sense that they can be computed formally or determined empirically.

Concepts even closer to that dealt with in this chapter, namely, querying with vagueness, are discussed in [17] where it is proposed that vagueness does not necessarily appear only in the data being queried but can also be part of the query itself. The work in [24] proposes classifications of membership values in order to group sets of values together (near fuzzy concepts). For example, a classification could assign the term "mostly" to high membership values (e.g., 0.95–0.98). In the context of databases in general, the approaches in [15, 16, 18, 19] all propose extensions to query languages on the basis of an operator that enables vague results under different circumstances. For example, in [15] the operator *similar-to* for QBE (Query-by-Example) is proposed alongside relational extensions so that related results can be provided in the event that no exact results match a query. In [18] the operator ~ is used in a similar way to the *similar-to* operator. All these approaches require additional information to be stored as extra relations and functions about distance that allow the query processor to compute close enough results. Although some of these approaches are extended to deal with fuzzy data, the general idea promotes the execution of vague queries over crisp data.

1.3 VASA

In this section we describe the concepts that compose our vague spatial algebra. The foundation of VASA is its data types, which we specify in Section 1.3.1. Spatial set operations and metric operations are introduced in Section 1.3.2. Finally, the concept of vague topological predicates is briefly introduced in Section 1.3.3.

1.3.1 Vague Spatial Data Types

An important goal of VASA (and of all approaches to handling spatial uncertainty that are based on exact models) is to leverage existing definitions of crisp spatial concepts. In VASA, we enable a generic vague spatial type constructor v that, when applied to any crisp spatial data type (i.e., *point*, *line*, *region*), renders a formal syntactic definition of its corresponding vague spatial data type. For any crisp spatial object x, we define its composition from three disjoint point sets, namely the interior (x°), the boundary (∂x) that surrounds the interior, and the exterior (x^-) [25]. We

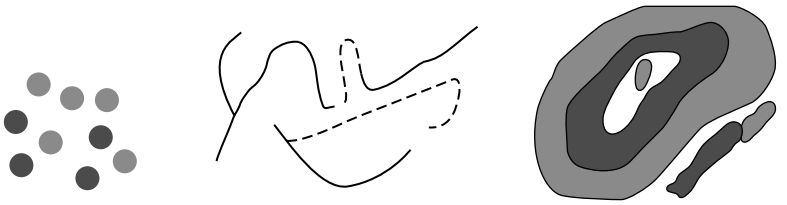

FIGURE 1.1 A vague point object (a), a vague line (b), and a vague region (c). Kernel parts are symbolized by dark gray points, straight lines, and dark gray areas. Conjecture parts are symbolized by light gray point, dashed lines, and light gray areas.

also assume a definition of the geometric set operations union ($\oplus$), intersection ($\otimes$), difference ($\ominus$), and complement ($\boxminus$) between crisp spatial objects such as that from [14].

Definition 1 Let $\alpha \in \{point, line, region\}$. A *vague spatial data type* is given by a type constructor v as a pair of equal crisp spatial data types α, i.e.,

$$v(\alpha) = \alpha \times \alpha$$

such that, for $w = (w_k, w_c) \in v(\alpha)$,

$$w_k^\circ \cap w_c^\circ = \varnothing$$

holds.

We call $w \in v(\alpha)$ a (two-dimensional) *vague spatial object* with *kernel part* w_k and *conjecture part* w_c. Further, we call $w_o := (w_k, w_c)$ the *outside part* of w. For $\alpha = point$, $v(point)$ is called a *vague point* object and denoted as *vpoint*. Correspondingly, for *line* and *region* we define $v(line)$ resulting in *vline* and $v(region)$ resulting in *vregion*.

Syntactically, a vague spatial object is represented by a pair of crisp spatial objects of the same type. Semantically, the first object denotes the kernel part that represents what certainly belongs to the object. The second object denotes the conjecture part that represents what is not certain to belong to the object. We require both underlying crisp objects to be disjoint from each other. More specifically, the constraint described above requires the interiors of the kernel part and the conjecture part to not intersect each other. Figure 1.1 illustrates instances of a vague point, a vague line, and a vague region as objects of the data types defined above.

1.3.2 Vague Spatial Operations

For the definition of the vague spatial set operations that compute the *union*, *intersection*, and *difference* between two vague spatial objects, we leverage crisp spatial set operations to reach a generic definition of vague spatial set operations.

TABLE 1.1
Components Resulting from Intersecting Kernel Parts, Conjecture Parts, and Outside Parts of Two Vague Spatial Objects with Each Other

union	***k***	***c***	***o***
k	*k*	*k*	*k*
c	*k*	*c*	*c*
o	*k*	*c*	*o*

intersection	***k***	***c***	***o***
k	*k*	*c*	*o*
c	*c*	*c*	*o*
o	*o*	*o*	*o*

difference	***k***	***c***	***o***
k	*o*	*c*	*k*
c	*o*	*c*	*c*
o	*o*	*o*	*o*

complement	***k***	***c***	***o***
k	*o*	*c*	*k*

We define the syntax of function $h \in$ [*intersection*, *union*, *difference*] as h: $v(\alpha) \times v(\alpha) \rightarrow v(\alpha)$. The complement operation is defined as *complement*: $v(\alpha) \rightarrow v(\alpha)$. Semantically, their generic (type-independent) definition is reached by considering the individual relationships between kernel parts, conjecture parts, and the outside part (i.e., everything that is not a kernel part or conjecture part) of the vague spatial objects involved in the operations. The result of each operation is computed using one of the tables in Table 1.1. For each operation, the rows denote the parts of one object and the columns the parts of another, and we label them *k*, *c*, and *o* to denote the kernel part, conjecture part, and outside part, respectively. Each entry of the table denotes the intersection of kernel parts, conjecture parts, and outside parts of both objects, and the label in each entry specifies whether the corresponding intersection belongs to the kernel part, conjecture part, or outside part of the operation's result object.

Each table from Table 1.1 can be used to generate an executable specification of the given crisp spatial operations. For each table, the specification operates on the kernel parts and conjecture parts to result in a definition of its corresponding vague spatial operation. Following are such definitions as executable specifications of geometric set operations over crisp spatial objects:

Definition 2 Let $u, w \in v(\alpha)$, and let u_k and w_k denote their kernel parts and u_c and w_c their conjecture parts. We define:

$$
\begin{aligned}
u \ \textbf{\textit{union}}\ w &:= (u_k \oplus w_k, (u_c \oplus w_c) \ominus (u_k \oplus w_k)) \\
u \ \textbf{\textit{intersection}}\ w &:= (u_k \otimes w_k, (u_c \otimes w_c) \oplus (u_k \otimes w_c) \oplus (u_c \otimes w_k)) \\
u \ \textbf{\textit{difference}}\ w &:= (u_k \otimes (\boxminus(w_k \oplus w_c)), (u_c \otimes w_c) \oplus (u_k \otimes w_c) \oplus u_c \otimes (\boxminus(w_k \oplus w_c)) \\
\textbf{\textit{complement}}\ u &:= (\boxminus(u_k \oplus u_c), u_c)
\end{aligned}
$$

1.3.3 Vague Topological Predicates

For the definition of topological predicates between vague spatial objects (*vague topological predicates*), it is our goal to continue leveraging existing definitions of crisp spatial concepts, in this case topological predicates between crisp spatial objects. Topological predicates are used to describe purely qualitative relationships such as *overlap* and *disjoint* that describe the relative position between two objects and are preserved under continuous transformations.

For two vague spatial objects $A \in v(\alpha)$ and $B \in v(\beta)$, and the set $T_{\alpha\beta}$ of all crisp topological predicates between objects of types α and β [25], the topological relationship between A and B is determined by the 4-tuple of crisp topological relationships (p,q,r,s) such that $p,q,r,s \in T_{\alpha\beta}$ and

$$p(A_k,B_k) \wedge q(A_k \oplus A_c,B_k) \wedge r(A_k,B_k \oplus B_c) \wedge s(A_k \oplus A_c,B_k \oplus B_c)$$

We define the set $V_{\alpha\beta}$ of all vague topological predicates between objects of types $v(\alpha)$ and $v(\beta)$. Due to inconsistencies that can exist between elements within each tuple, not all possible combinations result in 4-tuples that represent valid vague topological predicates in the set $V_{\alpha\beta}$. An example is the 4-tuple

$$(overlap(A_k, B_k), disjoint(A_k, B_k \oplus B_c), disjoint(A_k \oplus A_c, B_k), \\ disjoint(A_k \oplus A_c, B_k \oplus B_c))$$

In this example, the implications of $overlap(A_k, B_k) \Rightarrow A_k^\circ \cap B_k^\circ \neq \varnothing$ and $disjoint(A_k, B_k \oplus B_c) \Rightarrow A_k^\circ \cap (B_k \oplus B_c)^\circ = \varnothing$ clearly show a contradiction.

In [21], we present a method for identifying the complete set of vague topological predicates. At the heart of the method, each 4-tuple is modeled as a *binary spatial constraint network* (BSCN). Each BSCN is tested for *path-consistency*, which is used to check, via constraint propagation, that all original constraints are consistent; otherwise, the inconsistency indicates an invalid 4-tuple.

For each type combination of *vpoint*, *vline*, and *vregion*, possibly thousands of predicates are recognized. Sets of 4-tuples are created into clustered vague topological predicates. Clusters can be defined by the user who specifies three rules for each cluster: One rule is used to determine whether the clustered predicate certainly holds between the objects, the second to determine whether the cluster certainly does not hold, and the third to determine when the cluster maybe holds, but it is not possible to give a definite answer. This effectively symbolizes the three-valued logic that is central to our definition of vague spatial data types.

1.4 QUERYING WITH VASA

We propose two ways of enabling VASA within a database query language: The first, as presented in Section 1.4.1, works by adapting VASA to partially work with SQL, currently the most popular database query language. The second, presented in Section 1.4.2, extends SQL to enable handling of vague queries.

1.4.1 Crisp Queries of Vague Spatial Data

One of the advantages of being able to use VASA in conjunction with popular DBMSs is the availability of a database query language such as SQL. We focus on querying with SQL as it represents the most popular and widely available database query language. SQL queries can be used to retrieve data based on the evaluation of Boolean expressions. This obviously represents a problem when dealing with vague spatial objects because their vague topological predicates are based on a three-valued logic. On the other hand, the current definitions of numeric vague spatial operations do

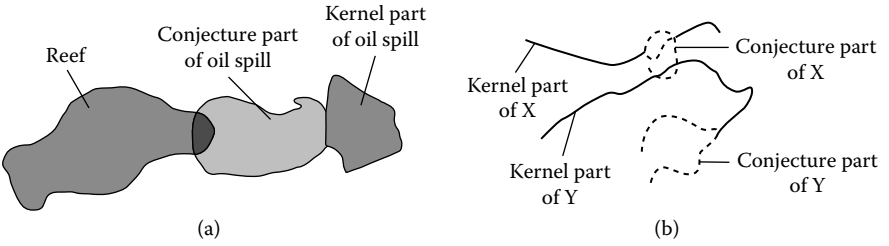

FIGURE 1.2 (a) A representation of an ecological scenario using vague regions. (b) Scenario illustrating the use of vague lines to represent routes of suspected terrorists *X* and *Y*.

not suffer from this issue because the operations return crisp values that are later interpreted by the user (e.g., the user posing a query must know that *min-length* returns the length associated with the kernel part of a vague line object). Thus, these concepts are already adapted to provide crisp results of vague data.

In the case of vague topological predicates, the first step in order to allow querying of vague spatial objects through SQL is to adapt the results of the predicates to a form understandable by the query language. The adaptation of the three-valued vague topological predicates to Boolean predicates can be done with the following six transformation predicates that are defined for each vague topological predicate *P* that can operate over vague spatial objects *A* and *B* (see Figure 1.2):

$$
\begin{aligned}
True\_P(A,B) = true &\Rightarrow P(A,B) = true \\
True\_P(A,B) = false &\Rightarrow P(A,B) = maybe \vee P(A,B) = false \\
Maybe\_P(A,B) = true &\Rightarrow P(A,B) = maybe \\
Maybe\_P(A,B) = false &\Rightarrow P(A,B) = true \vee P(A,B) = false \\
False\_P(A,B) = true &\Rightarrow P(A,B) = false \\
False\_P(A,B) = false &\Rightarrow P(A,B) = true \vee P(A,B) = maybe
\end{aligned}
$$

With this transformation in place, queries operating on vague spatial objects can include references to vague topological predicates and vague spatial operations. For example, for the purpose of storing scenarios such as that in Figure 1.2a, assume that we have a table *spills*(*id* : *INT , name* : *STRING, area* : *VREGION*) where the column representing oil spills is denoted by a vague region where the conjecture part represents the area where the spill may extend to. We also have a table *reefs*(*id* : *INT, name* : *STRING, area* : *VREGION*) with a column representing coral reefs as vague regions. We can pose an SQL query to retrieve all coral reefs that are in any danger of contamination from an oil spill. We must find all reefs that are not certainly *Disjoint* from the *Exxon-Valdez* oil spill:

```
SELECT r.name FROM reefs r, spills s
WHERE s.name = "Exxon-Valdez" and NOT True_Disjoint
(r.area,s.area);
```

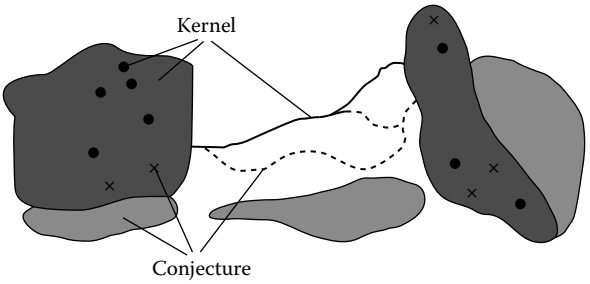

FIGURE 1.3 The vague spatial object representation of an animal's roaming areas, migration routes, and drinking spots.

Vague topological predicates can also be used to optimize query performance. Assume that, as illustrated in Figure 1.2b, we have data of terrorists' routes represented by vague lines in the table *terrorists*(*id : INT , name : STRING, route : VLINE*). We want to retrieve the minimum length of the intersections of all pairs of intersecting routes of terrorists. To do so, we choose to compute the intersection of only those pairs that are certainly not *Disjoint* and neglect the computation of the intersection of those pairs that have been determined to not certainly intersect:

```
SELECT a.name, b.name, min-length(intersection(a.route,
b.route))
FROM terrorists a, terrorists b WHERE False_Disjoint
(a.route,b.route);
```

Other queries can include retrieval based not only on spatial data but also based on common type data (i.e., numbers, characters) stored alongside the spatial objects. Being able to relate both data domains (spatial and nonspatial) in queries is one of the main advantages of providing VASA as an algebra that can extend current DBMSs that are well-proven to provide the necessary services for dealing with data of common types. We can provide such queries based on Figure 1.3, where the data can be stored in the table *animals*(*id* : *INT , name* : *STRING, roam area* : *VREGION , mig route* : *VLINE , drink spot* : *VPOINT*).

For example, we wish to retrieve all species of animals whose average weight is under 40 lbs. Their last count was under 100 and may have roaming areas completely contained within the roaming areas of carnivore animals whose average weight is above 80 lbs. This information might recognize animal species with low counts that could be extinct due to larger predators. The extinction of the smaller species can be catastrophic even for the larger species that depend on the smaller for nutrition. This retrieval uses data elements that are both spatial and nonspatial:

```
SELECT s.name FROM animals s, animals l
WHERE s.avgsize<40 AND l.avgsize>80 AND s.count<100;
```

Queries can also be posed to test elements from within single tuples in the database. For example, we would like to retrieve all animal species that do not have

drinking spots that are certainly lying inside their roaming areas. For any of these species, environmentalists must create artificial drinking spots where the animals can hydrate:

```
SELECT s.name,
FROM animals s
WHERE NOT False_Disjoint(s.drink_spot,s.roam_area);
```

1.4.2 A Vague Query Language Extension for Vague Queries on Vague Spatial Data

We analyze the approaches introduced in Section 1.2 and notice that, in the context of VASA, we are not trying to solve the problem of dealing with vague queries, but we need to query vague data. Thus, we propose to extend a common query language such as SQL with the operator ~. However, our data themselves are vague and thus we do not need the extra relations and functions of distance required by previous approaches. As a result, the semantics of the operator ~ is not the same as in the existing literature, where it allows for vague queries to be executed on crisp data. Instead, we will allow for the execution of vague queries over vague data.

Boolean predicates in SQL and in fact in many programming languages implicitly assume truth values unless otherwise noted, which is commonly done with the negation operators NOT or !. We propose ~ to operate syntactically similar to !, but semantically, instead of negating the result, it opens the possibility for uncertain results. For example, let us assume we have the table *tempzones*(*id* : *INT* , *name* : *STRING*, *shape* : *VREGION*) that contains information about different temperature zones, including their representation as vague regions in the column named *shape*. We pose the following query:

```
SELECT a.name, b.name
FROM tempzones a, tempzones b
WHERE Overlap(a.shape,b.shape);
```

This query will return only those regions that certainly overlap. But instead we want to include in the result all those regions that might overlap as well, so we pose the query again as

```
SELECT a.name, b.name
FROM tempzones a, tempzones b
WHERE ~Overlap(a.shape,b.shape);
```

In this case, the interpretation of ~ should allow the retrieval of all temperatures that may or may not overlap in addition to those that definitely overlap. For the use of numeric values in queries, the query processor should be able to handle number ranges as an atomic data type such that we can combine the minimum and maximum area operations on vague regions into one operator and pose the following query:

```
SELECT a.name
FROM tempzones a
WHERE a.shape.area()~300;
```

That is, the result of this query will include all temperature zones whose area range includes 300. The inclusion of this operator and the management of number ranges does not preclude the use of exact operators that would allow dealing with crisp spatial regions. Because crisp spatial objects represent simply a specific instance of vague spatial data types, a query such as the following can still be executed with the result set including all those temperature zones that were modeled as vague regions with no conjecture, thus representing crisp regions:

```
SELECT a.name
FROM tempzones a
WHERE a.shape.area()=300;
```

This extension, of course, would require actual re-implementation of the query language within the DBMS in order to enable the handling of numeric ranges and three-valued logic operations.

1.5 CONCLUSIONS AND FUTURE WORK

The conceptual design of VASA that we have presented in this chapter shows the clear goal of leveraging existing crisp concepts. There is more than one reason behind this goal. The first reason is to take advantage of existing robust concepts for handling crisp spatial objects. Second, at the conceptual level, the correctness of the definitions for vague concepts largely rests on the correctness of the defined crisp concepts; thus, we reduce the chance of errors in our definitions. As an example, see Definition 2, where vague spatial operations are defined as an executable specification on the basis of crisp spatial operations. Third, the executable specification translates easily to the implementation level. Having an existing correct implementation of crisp spatial data types, their operations, and predicates, we can implement VASA by instantiating existing crisp spatial data types and executing operations on them.

In Section 1.2, we mentioned current approaches to handling spatial vagueness and imprecision. VASA's concepts feed from all these and thrive in providing a complete type system that includes a systematic approach to vague spatial operations, and most importantly to vague topological predicates. The main advantages of VASA include conceptual simplicity, robustness derived from existing robust crisp concepts, and viability of implementation. In contrast, VASA's main disadvantage is its inability to effectively deal with situations that would seem appropriate for fuzzy set–based systems. Nonetheless, we believe that future work can be directed toward more general definitions based on exact models that would be more near to the capabilities of fuzzy set–based systems but that can take advantage of existing crisp concepts.

Based on these concepts, we have proposed ideas for database querying of objects from VASA. While these ideas are simple, they are able to fully exploit the

capabilities of VASA and allow the user to pose significant queries that can handle spatial vagueness. The proposed language extension and transformation mechanisms further reassure the advantages of defining VASA on the basis of existing exact spatial models. These advantages include robustness of formal concepts that can directly transfer into an implementation that also benefits from simplicity.

Other future work related to VASA stems in at least two directions that are worth following: The first involves enabling similar querying ideas to systems that attempt to handle vagueness with a higher precision, such as fuzzy set theory–based systems or even systems with finite multivalued logics (i.e., more than three values). The other direction involves the performance aspect of implementing indexes that can operate on vague spatial objects and whether it is possible to extend current indexing concepts for crisp spatial objects, thus following the design of VASA.

ACKNOWLEDGMENT

This work was partially supported by the National Science Foundation under grant number NSF-CAREER-IIS-0347574.

REFERENCES

[1] O. Ahlqvist, J. Keukelaar, and K. Oukbir. Rough Classifcation and Accuracy Assessment. *Int. Journal of Geographical Information Science*, 14:475–496, 2000.

[2] O. Ahlqvist, J. Keukelaar, and K. Oukbir. Rough and Fuzzy Geographical Data Integration. *Int. Journal of Geographical Information Science*, 17:223–234, 2003.

[3] D. Altman. Fuzzy Set Theoretic Approaches for Handling Imprecision in Spatial Analysis. *Int. Journal of Geographical Information Systems*, 8(3):271–289, 1994.

[4] T. Beaubouef and F. Petry. A Rough Set Foundation for Spatial Data Mining Involving Vague Regions. In *IEEE Int. Conf. on Fuzzy Systems*, pp. 767–772, 2002.

[5] P. A. Burrough and A. U. Frank, editors. *Geographic Objects with Indeterminate Boundaries*. Taylor & Francis, Boca Raton, FL, 1996.

[6] E. Clementini and P. Di Felice. A Spatial Model for Complex Objects with a Broad Boundary Supporting Queries on Uncertain Data. *Data & Knowledge Engineering*, Vol. 37, Issue 3, pp. 285–305, 2001.

[7] E. Clementini and P. Di Felice. An Algebraic Model for Spatial Objects with Indeterminate Boundaries. In Burrough and Frank, editors, *Geographic Objects with Indeterminate Boundaries*. Taylor & Francis, Boca Raton, FL, 1996, pp. 153–169.

[8] A. G. Cohn and N. M. Gotts. The "Egg-Yolk" Representation of Regions with Indeterminate Boundaries. In Burrough and Frank, editors, *Geographic Objects with Indeterminate Boundaries*. Taylor & Francis, Boca Raton, FL, 1996, pp. 171–187.

[9] A. Dilo, R., A. de By, and A. Stein. A System of Types and Operators for Handling Vague Spatial Objects. *Int. Journal of Geographical Information Science*, 21(4):397–426.

[10] M. J. Egenhofer. A Formal Definition of Binary Topological Relationships. In *Int. Conf. on Foundations of Data Organization and Algorithms*. Springer-Verlag, New York, 1989, pp. 457–472.

[11] M. J. Egenhofer, E. Clementini, and P. Di Felice. Topological Relations between Regions with Holes. *Int. Journal of Geographical Information Systems*, 8:128–142, 1994.

[12] M. Erwig and M. Schneider. Vague Regions. In *5th Int. Symp. on Advances in Spatial Databases*. Springer-Verlag, New York, 1997, pp. 298–320.

[13] J. T. Finn. Use of the Average Mutual Information Index in Evaluating Classification Error and Consistency. *Int. Journal of Geographical Information Systems*, 7(4):349–366, 1993.

[14] R. H. Guting and M. Schneider. Realm-Based Spatial Data Types: The ROSE Algebra. *VLDB Journal*, 4:100–143, 1995.

[15] T. Ichikawa and M. Hirakawa. ARES: A Relational Database with the Capability of Performing Flexible Interpretation of Queries. *IEEE Trans. on Software Engineering*, 12:624–634, 1986.

[16] J. Kung and J. Palkoska. Vague Joins — An Extension of the Vague Query System VQS. In *9th Int. Workshop on Database and Expert Systems Applications*, IEEE Computer Society, Los Alamitos, Ca, USA, 1998, pp. 997–1001.

[17] D. H. Lee and M. H. Kim. Accommodating Subjective Vagueness through a Fuzzy Extension to the Relational Data Model. *Information Systems*, 18:363–374, 1993.

[18] A. Motro. VAGUE: A User Interface to Relational Databases That Permits Vague Queries. *ACM Trans. on Information Systems*, 6:187–214, 1988.

[19] J. Palkoska and J. Kung. VQS — A Vague Query System Prototype. In *Int. Workshop on Database and Expert Systems Applications*, IEEE Computer Society, Los Alamitos, CA, USA, 1997, pp. 614–618.

[20] A. Pauly and M. Schneider. Vague Spatial Data Types, Set Operations, and Predicates. In *East-European Conf. on Advances in Databases and Information Systems*, Springer-Verlag, Berlin/Heidelberg, 2004, pp. 379–392.

[21] A. Pauly and M. Schneider. Topological Reasoning for Identifying a Complete Set of Topological Predicates between Vague Spatial Objects. In *FLAIRS Conference*, AAAI Press, Menlo Park, CA, USA, 2006, pp. 731–736.

[22] Z. Pawlak. Rough Sets. *Int. Journal of Computer and Information Sciences*, pp. 341–356, 1982.

[23] M. Schneider. Uncertainty Management for Spatial Data in Databases: Fuzzy Spatial Data Types. In *Int. Symp. on Advances in Spatial Databases.* Springer-Verlag, New York, 1999, pp. 330–351.

[24] M. Schneider. Fuzzy Topological Predicates, Their Properties, and Their Integration into Query Languages. In *ACM Symp. on Geographic Information Systems.* ACM Press, New York, 2001, pp. 9–14.

[25] M. Schneider and T. Behr. Topological Relationships between Complex Spatial Objects. *ACM Trans. on Database Systems (TODS)*, 31:39–81, 2006.

[26] M. Worboys. Imprecision in Finite Resolution Spatial Data. *GeoInformatica*, 2(3):257–279, 1998.

2 Assessing the Quality of Data with a Decision Model

Andrew Frank

CONTENTS

2.1 INTRODUCTION

Research in data quality is hindered by a lack of understanding of what quality for data means. The slogan "data quality is 'fitness for use'" is not giving an answer because it leaves open the question to what use the data should be fit. Data, especially GIS data, can be used in many ways; remember that a precursor of GIS was called a "multi-purpose cadastre" (Arentze et al., 1992; Harvey, 1997)! Data are used to improve decisions; decisions can be made without pertinent information (case of "null" information, e.g., none, inappropriate), and decisions are not necessarily

changed after data are needed—only confidence is increased (Frank, to appear 2007). GIS data can be used to improve many decisions, from ordinary, everyday decisions in wayfinding (left or right here?) to complex decisions about the location of a new nuclear power plant or a new factory or the violation of an international treaty (Abushady and Frank, 2005).

The quality of the information influences the decision—it must be assessed with respect to the decision-making process: Can it be used to make this decision? Does the lack of quality influence the outcome?

The diversity of the decisions GIS data are used for makes it difficult to understand how the quality of the data affects the decision. This is further complicated by the psychological complexity of how people actually make decisions. A number of studies have shown how data quality propagates from the data stored to data derived from a GIS to help make decisions (Karssenberg and De Jong, 2005). De Bruin et al. (2001, 2003) investigated whether acquiring better data for a particular decision is worthwhile.

Schneider (1999) and Frank (2007) have been able to reduce decisions as they are made by engineers when designing technical artifacts to a statistical test. Once the engineer has selected the model and parameters to include, the decision itself can be reduced to a comparison of two desired quantities. This approach is generalized here to as broad a range of decisions as possible.

This approach to data quality from the perspective of a user is different from describing data quality from the perspective of the data producer working with a specification, which typically emphasizes precision of location (Timpf et al., 1996). Unfortunately, such quality descriptions from the producer perspective are seldom relevant for users of the data (Shyllon and Hunter, 2004).

In this chapter I briefly review in Section 2.2 the model for engineering decisions as proposed before (Frank, 2007). In Section 2.3 different types of decisions are analyzed. Twaroch and Achatschitz (2005) investigate how the user's situation can be captured separately in an interactive process; the models their work produces can be used to assess the propagation of data quality to decision quality as described here. Ignoring the psychological complexity of decision, especially if made in a group, a similar reduction to a comparison of values devised from the data stored can be achieved. Section 2.4 then generalizes the model for random errors in the data, and Section 2.5 discusses the propagation of different data quality aspects from stored data to desired quantities.

As a result, the chapter shows a reduced model of decision making, which separates the psychological complexities of taking a decision into a first phase in which the "problem" is conceptualized into a decision test and a model selected. This process is in most decisions not consciously performed or verbalized. In the second phase, the decision is computed according to the model selected. It is possible to construct the model used "after the facts," when the decision is made and one can reconstruct the process. This reconstructed model can then be used to assess how data quality has influenced the decision, which makes the method described not only of theoretical interest but also practically applicable.

With this division of a complex decision into two steps the propagation of data quality can be computed, because error propagation affects only the second one

and can be formalized. This chapter identifies the processing steps for which the propagation of imperfections is necessary and points to the research needed to give general rules for the ones not currently well understood.

A note on terminology: I prefer to speak of imperfections of the data (Frank, 2007) and to characterize these. This is focusing on the effects such imperfections have on the (imperfect) result, and I avoid statements like "low data quality" or "lack of data quality." All data contain imperfections, and it seems conceptually simpler to address these imperfections, rather than talk about data quality, which describes the degree of absence of imperfections.

2.2 ENGINEERING DESIGN DECISIONS

Engineering design, for example, for buildings, bridges, sewage systems, etc., is based on physical observations that are combined in formulas. The results are used to decide if a design satisfies the requirements and is acceptable or not. Error propagation is applicable here, and one can ask how much every value computed is influenced by the error in the data. Schneider has analyzed the influence of assumptions about load, strength of materials, or required safety levels (Schneider, 1999).

In engineering design, decisions can be abstracted to a comparison between the load on a system S compared with the resistance of the system R as designed. A design is acceptable if the resistance is larger than the load: $R > S$ resp. $R - S > 0$.

For a bridge, this means that the resistance R of the structure (i.e., maximum capacity) must be higher than the maximally expected load S (e.g., assumed maximum high water event). For a more environmental example, the opening under a bridge is sufficient and inundation upstream is avoided when the maximally possible flow R under the bridge is more than the maximal amount of water S expected from rainfall on the watershed above the bridge. To assess the influence of data quality on the decision, one computes the error on $(R - S)$ using the law of error propagation and applies test statistics to conclude whether the value is lager than zero with probability p (e.g., 95%).

The law of error propagation for a formula

$$r = f(a, b, \ldots)$$

for random uncorrelated errors e_a, e_b, e_c on values a, b, c, … was given by C. F. Gauss as

$$e_r^2 = \left(\frac{\partial f}{\partial a}\right)^2 e_a^2 + \left(\frac{\partial f}{\partial b}\right)^2 e_b^2 + \ldots \tag{2.1}$$

where e_i is the standard deviation of value i. If the observations are correlated, the correlation must be included (Ghilani and Wolf, 2006). The test on $R - S > 0$ is then

$$\frac{R - S}{\sqrt{\sigma_R{}^2 + \sigma_S{}^2}} > C \tag{2.2}$$

where C is determined by the desired significance, e.g., for 95%, $C = 1.65$.