

LECTURE NOTES IN GEOINFORMATION AND CARTOGRAPHY

LNG&C

Gerhard Navratil (Editor)

Research Trends in Geographic Information Science



Springer

Lecture Notes in Geoinformation and Cartography

Series Editors: William Cartwright, Georg Gartner, Liqiu Meng,
Michael P. Peterson

Gerhard Navratil (Ed.)

Research Trends in Geographic Information Science

 Springer

Editor

Dr. Gerhard Navratil
TU Wien
Institute of Geoinformation and Cartography
Gußhausstr. 27–29 E127
1040 Wien
Austria

ISSN 1863-2246 e-ISSN 1863-2351
ISBN 978-3-540-88243-5 e-ISBN 978-3-540-88244-2
DOI 10.1007/978-3-540-88244-2
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2009930471

© Springer-Verlag Berlin Heidelberg 2009

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Cover design: deblik, Berlin

Cover image: The picture for the cover was taken by Andrew U. Frank and changes by Gerhard Navratil

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

In February 2008 Andrew U. Frank, head of the Institute for Geoinformation and Cartography of the Vienna University of Technology, celebrated his 60th birthday. This anniversary gave reason to organize a scientific meeting in Vienna on June 30th and July 1st 2008. We invited six renowned key researchers and asked others to propose presentations. The topic of the whole meeting was defined by Andrew U. Frank's main interest: Upcoming fields of research.

The final program consisted of sixteen presentations: six keynote presentations held by Helen Couclelis, Max J. Egenhofer, Mike F. Goodchild, Werner Kuhn, David M. Mark, and Lotfi A. Zadeh and ten short presentations. Approximately 60 researchers from 5 continents used the opportunity to get in contact with new research ideas. A major issue, when creating the schedule, was to put in enough time for discussion. This proved successful and everybody left the meeting with a bunch of new ideas and full of enthusiasm to explore them further.

The only blemish of this wonderful meeting was that there were no adequate proceedings. Although a paperback collection of the papers were printed, they were no match for the quality of the presentations and the discussions. Therefore, we asked Springer if they would be interested in publishing a book containing most of the presented material. Springer agreed and we asked the presenters of the meeting to submit extended versions of their papers. We also informed several other key scientists, who could not participate in the meeting, about the possibility to contribute to the book. Unfortunately, some of the participants of the meeting could not submit their papers because they were already submitted elsewhere. However, we received a fair number of papers and after a review process eighteen papers were selected for publication.

The result of this process lies before you. The eighteen papers are arranged in three groups:

- philosophical background and semantics (chapters 1 to 5),
- mathematical methods (chapters 6 to 13), and
- tools and application (chapters 14 to 18).

In some cases it was difficult where to put the paper. Other editors would have moved some papers to other parts of the book. A reason for this is that papers do not concentrate on a few core topics. The topics are distributed over the wide field of geoinformation science and thus no crisp groups can be determined. I tried to group papers with similar topics and to order the material in a suitable way.

Many people helped were responsible for the success of the meeting and the completion of this book. I want to thank everybody who was involved. Special thanks go to our secretary Edith Unterweger for executing all administrative tasks and to our publication editor Christian Gruber for both his help with the printed results and his performance as a magician at the meeting. The whole team of the Institute for Geoinformation and Cartography supported me in a fantastic way during the process and many details would not have worked without their help. Without the presenters of the meeting and the authors of the chapters in this book I would have had nothing to edit. Thus they earn a big ‘thank you’! Finally, I have to thank the reviewers, who gave invaluable input to all authors and thus helped to guarantee high quality of the final book (in alphabetic order):

Arnold Bregt
Marco Cassanova
Helen Couclelis
Michael Drmota
Andrew U. Frank
Barbara Hofer
Gary J. Hunter
Farid Karimipour
Marinos Kavouras
Gerhard Muggenhuber
Takeshi Shirabe
Alfred Stein
Kathleen Stewart
Erik Stubkjaer
Richard Wadsworth
Stephan Winter

I hope the book inspires young and experienced researchers alike and leads to new ideas. Please keep in mind the motto of the meeting:

If everybody shares his ideas then everybody gains and nobody loses.

Gerhard Navratil

Vienna, Austria, April 2009

Table of Contents

Section I: Philosophical Background and Semantics 1

Helen Couclelis

Ontology, Epistemology, Teleology: Triangulating Geographic

Information Science.....3

- 1 Introduction 4
- 2 Ontologies, Artificial Worlds, and Cognitive Semantics 6
- 3 Information, Information Objects, and Purposes: Sketch for
an Ontological Framework 8
- 4 Triangulating Geographic Information Science 11
 - 4.1 Ontology-Epistemology..... 11
 - 4.2 Ontology-Teleology..... 12
 - 4.3 Teleology-Epistemology 13
- 5 Conclusion: From Triangles to Tetrahedra (perhaps) – and Beyond.. 13

Marinos Kavouras

Geonoemata Elicited: Concepts, Objects, and Other Uncertain

Geographic Things..... 17

- 1 Introduction 17
- 2 Obstacles and Challenges 20
- 3 Corpus of Geographic Knowledge 21
 - 3.1 Essence in Geographic Concepts..... 21
 - 3.2 Principal Ontologies 22
 - 3.3 Light vs. Specialised Ontologies 22
 - 3.4 Conceptual Instruments 23
 - 3.5 Geonoemata from Natural Language..... 23
- 4 Conclusions 23

Nancy J. Obermeyer

Virtue Ethics for GIS Professionals 27

- 1 Introduction 27
- 2 Development of Ethics for GIS Practitioners 28
- 3 Virtue Ethics 30

4 Linking Virtue Ethics to Professionalism..... 32
 5 Promoting Virtue among GIS Professionals..... 33

Andrew U. Frank

Why Is Scale an Effective Descriptor for Data Quality? The Physical and Ontological Rationale for Imprecision and Level of Detail..... 39

1 Introduction 40
 2 Tiered Ontology..... 41
 2.1 Tier 1: Point Observations 42
 2.2 Tier 2: Objects 42
 2.3 Tier 3: Social Constructions 43
 3 Information Processes..... 44
 3.1 Observations of Physical Properties at Point..... 45
 3.2 Object Formation (Granulation) 46
 3.3 Boundary Identification..... 47
 3.4 Determination of Descriptive Summary Data 47
 3.5 Classification 48
 3.6 Constructions 49
 4 Random Effects on the Observations 49
 4.1 Influence on Object Formation and Summary Values 49
 4.2 Classifications..... 51
 4.3 Qualitative Description..... 51
 5 Finite Observation Devices 52
 5.1 Effects of Size of Sensor 52
 5.2 Effects of Finite Number of Observations..... 53
 5.3 Effects of Finite Representation of Observations..... 55
 5.4 Effects of Scales on Object Formation 55
 5.5 Qualitative Descriptors 56
 6 Scale as a Summary Description 56
 7 Conclusions 57

Werner Kuhn

Semantic Engineering..... 63

1 Introduction 63
 2 An Engineering View of Concepts 65
 2.1 A Triadic Notion of Concepts 65
 2.2 The Semantic Triangle Revisited 66
 3 Ontologies as Networks of Constraints 68
 4 Grounding Constraint Networks..... 69
 5 Integrating Folksonomies with Ontologies..... 71
 6 Accommodating Uncertainty..... 72
 7 Conclusions 72

Section II: Mathematical Methods.....77**Christopher Gold****A Common Spatial Model for GIS.....79**

1 Introduction	79
2 Geographic Data	80
3 Models of Space	81
3.1 Objects and Fields	82
3.2 Objects: Spatial Extension.....	84
4 Data Structures, Algorithms and Applications	85
4.1 Classes of Algorithms.....	85
4.2 Incremental Algorithms and Applications.....	86
4.3 Dynamic Algorithms and Applications	89
4.4 Kinetic Algorithms and Applications	89
5 Conclusions	92

Lotfi A. Zadeh**Computation with Imprecise Probabilities.....95**

Extended Abstract.....	95
------------------------	----

Gary J. Hunter et al.**Spatial Data Quality: Problems and Prospects.....101**

1 Introduction	101
2 Problems	104
2.1 Poor Quality Reporting.....	104
2.2 Incomplete Quality Descriptions	105
2.3 Barriers to Communicating Quality.....	107
2.4 Keeping Track of Error.....	109
2.5 Application of Data Quality Information	109
3 Future Prospects	110
3.1 Enhanced Quality Reporting	110
3.2 Improved Quality Descriptions	111
3.3 More Effective Quality Communication	112
3.4 Better Error Tracking.....	114
3.5 Complete Utilization of Quality Information	116
3.6 Final Remarks.....	117
4 Conclusion	118

Richard A. Wadsworth et al.**Latent Analysis as a Potential Method for Integrating Spatial****Data Concepts** 123

1 Introduction	123
2 Estimating Semantic Consistency.....	124

3 Methods 125
 4 Results 126
 4.1 Number of Latent Variables in a Data Set..... 126
 4.2 Latent Variables Uncovered by Latent Analysis 127
 5 Discussion and Conclusions 130

Alfred Stein et al.

Stereology for Multitemporal Images with an Application to Flooding 135
 1 Introduction 135
 2 Stereology..... 137
 2.1 Basics..... 137
 2.2 Identifying 2D Objects from Images 139
 2.3 Remote Sensing Images 140
 3 Discussion..... 147
 4 Concluding Remarks 148

Kathleen Stewart Hornsby, Naicong Li

Modeling Spatiotemporal Paths for Single Moving Objects..... 151
 1 Introduction 151
 2 Modeling Spatiotemporal Paths of Moving Objects..... 153
 3 Modeling Spatiotemporal Paths of Movement:
 Open and Closed Paths 155
 4 Possible Path Patterns for Single Moving Objects 159
 5 Summary and Future Work 164

Markus Schneider

Moving Objects in Databases and GIS: State-of-the-Art and Open Problems..... 169
 1 Introduction 169
 2 Moving Objects in Unconstrained Environments..... 170
 2.1 Historical Moving Objects..... 171
 2.2 Predictive Moving Objects 173
 3 Moving Objects in Constrained Environments..... 177
 3.1 Spatial Networks 177
 3.2 Moving Objects in Spatial Networks..... 181
 4 Conclusions 184

Michael Drmota

The Degree Distribution of Random Planar Graphs..... 189
 1 Introduction 189
 2 Planar Graphs 191
 3 Random Planar Maps..... 196

4 The Random Graph Model of Erdős and Rényi	197
5 Conclusions	198

Section III: Tools and Applications.....201

Gilberto Câmara et al.

Geographical Information Engineering in the 21st Century	203
1 Introduction	203
2 From GIS-20 to GIS-21	205
3 Change, Cognition, and Semantics: Three Critical Issues	207
3.1 Change	207
3.2 Semantics	209
3.3 Cognition	209
4 Building New Tools to Model Change: An Engineering View	210
5 A Problem and a Possible GIEngineering Solution: A Global Forest Information System	211
6 Final Remarks: GI Engineers and GI Scientists Need to Cooperate	214

Robert Laurini

Towards Visual Summaries of Geographic Databases Based

on Chorems.....	219
1 Introduction	219
2 What are Chorems?	220
2.1 From Conventional Cartography to Chorem Maps	220
2.2 Results of a Study of Existing Manually-Made Chorem Maps	224
2.3 Towards New Concepts for Geographic Databases	224
3 Architecture of the System	225
3.1 Spatial Pattern Discovery	226
3.2 Chorem Layout	227
3.3 ChorML	229
4 Final Remarks	231

Stephan Winter, Yunhui Wu

Intelligent Spatial Communication	235
1 Introduction	235
2 A Criterion for Intelligent Spatial Communication	237
3 A Framework for the Requirements Analysis	241
3.1 The Phases of Wayfinding Communication and Their Tasks	241
3.2 The Spatiotemporal Context of Wayfinding Communication	244

3.3 Representing an Intelligent Agent in Wayfinding Communication	245
4 Conclusions	247
 Marcelo G. Metello, Marco A. Casanova	
Training Games and GIS	251
1 Introduction	251
2 Requirements for Geospatial Training Games	252
2.1 Realistic User Experience.....	252
2.2 Interoperability with Existing GIS	252
2.3 Time Flow Control	253
2.4 Player Performance Evaluation	253
3 An Architecture for Geospatial Training Games.....	254
4 Process Modeling	256
4.1 Discrete Process Modeling	256
4.2 Continuous Process Modeling.....	258
4.3 Workflows	259
4.4 The Event Framework	259
5 An Example of a Geospatial Training Game.....	260
5.1 Emergency Plans	260
5.2 Example of Running the Emergency Training Game.....	261
5.3 Benefits of Using the Emergency Training Game.....	263
6 Conclusions	263
 Erik Stubkjær	
Cadastre and Economic Development	265
1 Introduction	265
1.1 A Research Brief on Cadastre and Economic Development	265
1.2 “Why Isn’t the Whole World Developed?”	267
2 Statistical Analyses of the Causes of Economic Growth.....	268
2.1 Technology and Belief	269
2.2 Formal Education	269
2.3 The Spread of News	271
2.4 Limits to Statistical Analyses of Causes of Economic Growth	272
3 The Quality of Institutions and the Role of the State	273
3.1 Available Indices of the Protection of Property Rights	273
3.2 The Management of Institutional Change for Growth	275
4 Creating Capital with Covered Bonds	276
5 Conclusion	277
 Index	 281

Ontology, Epistemology, Teleology: Triangulating Geographic Information Science

Helen Couclelis

University of California, Santa Barbara, California, USA
cook@geog.ucsb.edu

Abstract

For the past several years ontology has enjoyed a robust regard within the geographic information science community. Ontology is however only one apex of a triangle of knowledge that also involves epistemology and the (long discredited) notion of teleology. Without epistemology we lack a systematic understanding of the nature of the correspondence between ontologies and the general or specific domain of inquiry each of them represents. Without teleology we miss the crucial distinction – essential especially in ontologies of change – between the outcomes of causal processes on the one hand, and the results of purposeful action by sentient actors or machines on the other. This paper argues that connecting current conceptions of ontology with these two other, complementary perspectives would allow new kinds of scientific questions to be addressed as well as to expand the scope of ontology itself in geographic information science. I briefly outline an ontological framework that builds on the epistemological notion of *information* and is guided by the teleological notion of *purpose*, and based on this sketch I suggest a possible way of completing the golden Greek triangle of geographic information science.

Was wir als Wirklichkeit wahrnehmen, ist unsere Erfindung!
(Heinz von Foerster)

¹ What we perceive as reality is our own invention.
<http://www.univie.ac.at/constructivism/HvF.htm>

1 Introduction

Of the three pillars of Aristotle's philosophy, ontology alone enjoys a robust regard within geographic information science. What was not long ago the esoteric pursuit of a handful of philosophically inclined researchers has now become a mainstream subfield yielding tangible results. Ontology is however only one apex of a triangle of knowledge that also includes epistemology and teleology. While ontology deals with what exists in a given world, epistemology is concerned with the nature and scope of knowledge, and teleology with the *reasons* (not causes) why the world is as it is and why and how it is changing. Reasons derive from the beliefs, thoughts, hopes and desires that lead people to strive towards particular purposes or goals, whereas causes have effects regardless of any reference to human intentionality (Lyons 1995). The distinction may often boil down to a difference in perspective. You may correctly think that the rain caused me to open my umbrella, though from my perspective it is my desire not to get wet, along with my belief that a personally unpleasant state will result if I do get wet, that leads me to open my umbrella. In many (most?) cases the first, simpler account may be good enough, but clearly something is lost by leaving out the teleological explanation. This paper argues that linking current conceptions of ontology in geographic information science with teleology as well as with epistemology would allow new kinds of scientific questions to be addressed would expand the scope of ontology itself towards novel and fruitful directions.

In developing the argument for epistemology and teleology this paper explores the implications of basing ontology construction in geographic information science on the dual principles of *information* and *purpose*. Information, a fundamental epistemological notion, seems a natural choice for an *information science*, but there are additional advantages relevant to the task at hand. One advantage is that information is a relational rather than an absolute concept, expressing an intrinsic relationship between information source and recipient (Williamson 1994; Huchard et al. 2007). This establishes a basis for taking into account the interests of the user and thus for forging a link between ontology and teleology. Another advantage of using information as the central notion rather than concepts or linguistic terms more directly associated with the empirical world is precisely that it begs the question of the relationship between that empirical world and the ontology under consideration, thus making epistemology inescapable. Finally, since information comes in quanta, it facilitates a constructivist approach to ontology development, an approach that drives the framework briefly outlined in section 3.

Purpose, the hallmark of teleology, is not a notion emphasized in existing ontologies of geographic information science, and yet every geo-

graphic representation is developed for some purpose. In addition, many if not most of the empirical entities represented in these ontologies have also been created or modified by humans with specific purposes in mind: these are the ‘artificial’ entities that Simon (1969) writes about in his famous book *‘The Sciences of the Artificial’*. This double observation provides an answer to the geospatial ontology builder’s basic challenge, which is how best to represent geographic phenomena. In the case of socially produced geographic phenomena such as roads, land parcels and campgrounds the challenge is augmented by the necessity to draw a line between these tangible, physical entities and the rest of the social world. *Purpose* may be seen as the interface between observable geographical entities on the one hand, and social needs and wants on the other. Purpose is like a permeable membrane enveloping the geographical world that permits socially conditioned questions to traverse it one way, and socially meaningful interpretations to emerge in the other direction.

The ontological framework outlined below, which is anchored by the notions of information at one end and of purpose at the other, is populated by an ordered sequence of discrete domains of *information objects* that correspond to – but are logically different from – the more familiar types of empirical entities or concepts that are the focus of most other ontologies. Details of the framework summarized here are presented in a forthcoming article. This paper focuses on the triangle of knowledge of geographic information science that has ontology, epistemology and teleology as its apexes, and of which the interior is occupied by the foundational notions of information, (information) object, and purpose. The sides of that triangle are also meaningful, suggesting a number of research questions that are either new or are novel perspectives on familiar geographic information science questions. For example:

- *Ontology-Epistemology*: Given two arbitrary ontologies, how can we tell a priori to what extent these may be compatible? Given a specific ontology, how can we tell whether it is complete and internally consistent?
- *Ontology-Teleology*: How can we best represent artificial objects and purposeful change within the same ontological framework as natural objects and non-purposeful change? How can we seamlessly integrate the natural and the artificial in both static descriptions and representations of change?
- *Teleology-Epistemology*: How do we know that a proposed ontology is appropriate for particular user purposes? What is the role of intentional stance in helping decide among competing ontologies?

Ontology and epistemology have been considered together before from the perspective of geographic information science (Frank 2001) but the lowly

status of teleology in traditional scientific thinking must have been a major reason why that third essential perspective on knowledge has not been a more prominent part of the field's agenda. Yet teleology has undergone a quiet renaissance since the beginning of the computer era in connection with the realization that some kinds of advanced machines are characterized by purposeful behavior (Rosenblueth et al. 1943). More recently the international conference series DEON (Conference on Deontic Logic in Computer Science, <http://deon2008.uni.lu/cfp.htm>) has brought teleology into mainstream computer science, highlighting the host of contemporary domains of application that can benefit from implementations of telic thinking. In my view geographic information science should be one of these domains.

The next section briefly discusses current notions of ontology in connection with related research in geographic information science. I then sketch out a hierarchical ontological framework based on the dual notions of information and purpose, arguing that (a), the systematic relationships between levels of that framework suggest answers to certain significant epistemological questions and (b), the relevance of teleology becomes evident at the highest levels of that hierarchy. Only the static case is discussed here, covering time-slice extensions but not processes, actions and dynamic events. Since the purpose of this paper is primarily to pose questions and to stimulate discussion, the conclusion is brief and speculative.

2 Ontologies, Artificial Worlds, and Cognitive Semantics

Small -'o' ontology is the description of a world – not 'The' World. The ontology of a domain of inquiry is the formal description of an artificial world (one of an infinity of possible small-'w' worlds) that gives rise to legitimate *representations* or models constructed within that domain. This view is in agreement with a widely accepted definition according to which an ontology is a formal and explicit specification of a shared conceptualization (Gruber 1993; Borst et al. 1999). Shared conceptualizations and their kin, descriptions, representations and models, are just as likely to be dynamic as static, yet existing geographic information ontologies tend to be strongly biased towards static categorizations and classifications. Beyond internalizing dynamics and process this definition also stresses the intersubjective and cognitive nature of ontologies. It is also compatible with a perspective on models from computer science, proposing that a model represents a microworld consisting not only of contents (things and their relationships) but also of spatial structure, temporal structure, 'physics' (processes allowed or rules of interaction and behavior), and rules of inference or logic (Smyth 1998; Couclelis 2002; see also Zeigler et al.

2001). To the extent that ontologies are models this view applies to them also, emphasizing their contingent nature which extends beyond contents and structure to process and to the fabric of space and time themselves.

The relevant ontologies for the framework sketched in this paper are the foundation ontologies, which purport to describe fundamental concepts and relations that are valid across domains. This is in contrast to the more specialized domain ontologies that are specifically tailored to the needs of particular areas of inquiry or application. Foundation ontologies may focus on real-world entities, on concepts or linguistic entities, or they may be mixed, integrating both 'external' and 'internal' representations (see Agarwal 2005 for a review). Still other ontologies strive to encompass the entire spectrum of empirical reality as seen from a spatial standpoint, including abstract aspects such as the social and experiential. These are usually hierarchical, consisting of sequences of 'levels', 'realities', 'worlds' or 'spaces', and they tend to be quite similar in principle. For example, Couclelis (1992) suggests a hierarchy consisting of mathematical space, physical space, socioeconomic space, behavioral space, and experiential space; Guarino's (1999) ontology of particulars includes a physical level, functional level, biological level, intentional level, and social level; and Frank's (2003) proposed tiers are: physical reality, observable reality, object world, social reality, and cognitive agents. These and several other empirically based ontologies may all be useful and intuitively plausible but they tend to lack a convincing justification for the designation, order and contents of the levels, as well as a systematic procedure for moving up and down the hierarchy.

Kuhn et al. (2007, p 7–8) propose a list of potential benefits for ontology engineering to be gained from cognitive semantics. These include:

- “*Grounding ontologies*, that is, establishing primitives that are both meaningful and suitable as building blocks for ontologies”.
- “Moving space and time from their current status as application domains to become *foundational* aspects of ontology”.
- “Reconciling *meaning and truth*”.
- “Allowing for *perspectivalism* without giving in to relativism”.
- “A cognitively plausible ...understanding and formalization of *conceptual mappings*”.
- “*Personalizing* geospatial services” by taking into account “situational and personal context”.

These desiderata are clearly epistemological while the call for perspectivalism and personalized geospatial services also hint at teleology in that they bring user needs, perspectives and context into the picture. While not explicitly emphasizing cognitive semantics, the framework outlined in the

next section indicates one possible way of addressing these points in ontology engineering. One important question that this framework raises is precisely the connection between information, user purpose, and cognitive semantics. The very partial and tentative answer provided here will hopefully contribute to a much needed broader discussion on this issue.

3 Information, Information Objects, and Purposes: Sketch for an Ontological Framework

The ontological framework adumbrated here consists of an ordered sequence of seven systematically related hierarchical levels. The levels are differentiated by their degree of semantic richness, ranging from minimal to maximal semantic complexity. Like most hierarchies this framework may be approached from either end, from the bottom up or from the top down. A very important point concerns the distinction between the mode of generation and the mode of use of this hierarchy. While it is interpreted and used from the highest level down (from purpose to minimal necessary information: the *intentional* direction), it is more logically presented from the bottom up (from minimal elementary information building up to purpose: the *generative* direction). The generative direction, which helps explain the systematic procedure by which the levels are derived from each other, will not be discussed in this paper.

The idea underlying the construction of the hierarchy is the following. Like all information sciences, geographic information science is about *representations* of entities, not directly about the entities themselves. Representations are made up of *information* selected and organized for some *purpose*. Now, the possible ways of selecting and organizing information to represent any non-trivial geographical phenomenon are in principle indefinitely many. The reason why we come up with models that are 'use'-ful is that this process of selecting and organizing information is implicitly or explicitly guided by some practical purpose. The purpose of the framework outlined here is thus to present a systematic model of how representations of geographic entities relate to available information on the one hand, and to the purpose(s) for which such representations are constructed, on the other. Seven different semantic domains, corresponding to the seven levels of the hierarchy, are distinguished in this framework, though no claim can be made that seven levels are either necessary or sufficient. These semantic domains range from the most complex, populated by representations that fully correspond to their intended purposes, to the most sparse, where only the potential existence of information suitable for constructing a specific representation of interest is ascertained. In between the

two ends lie five more levels that can be derived from each other either by adding purposefully selected semantic content (bottom up), or, conversely, by subtracting suitably selected semantic content (top down), so that the representations ('information objects') left behind are still meaningful though increasingly semantically impoverished.

The following is a brief outline of the levels of the hierarchy, while Table 1 provides an illustrative example of how the 'same' phenomenon is specified differently across the levels depending on purpose. Teleology thus motivates the construction of this entire ontological framework. At the same time, the suggested recursive decomposition procedure clarifies the epistemological relations among levels. Taking the top-down view, the levels are derived from each other by subtracting at each step suitably defined domains of semantic content until there is hardly anything left behind. Imagine an originally fully conscious intentional agent (a person, a group, or a society) becoming more and more semantically challenged as it descends the hierarchy— or perhaps 'Hal' the *Space Odyssey* computer as its modules are gradually being stripped away by the spaceship's frantic survivors (Clarke 1968).

Level 7: Purpose. Purpose is not itself a geospatial concept, but as mentioned earlier, it is the interface between the world of geospatial entities on the one hand, and the social world of intentional agents on the other. Purpose determines what spatial functions need to be represented, what distinct spatial entities belong together to form a complex object, how simple objects are named and categorized, what spatial patterns and measurable properties correspond to the entities of interest and how these should be analyzed, what sort of information is relevant, and finally, what spatio-temporal framework must underlie the representations appropriate for the purpose in question. Purposes thus *select* suitable information subsets out of a comprehensive domain of possible data and *construct* out of these the semantically appropriate (to that purpose) information objects.

Earlier we distinguished two different kinds of purposes relating to geographic representations: (i) their intrinsic purposes *qua* models built with an end use in mind (e.g., the representation of a weather front intended for navigation rather than for presentation on television), and (ii) the extrinsic (to the representation itself) purpose of any artificial entities (e.g., the purpose of a school or a bridge) that may need to be reflected in the data model. The illustrative example in Table 1 combines these two cases: here a map of roads *qua* representation serves two different purposes, in the first of which the main objects represented (roads) are approached as artificial entities endowed with their own purpose (transportation), whereas in the second case (ecological study) the purpose of the roads is irrelevant and only their emerging function as dangerous physical barriers to wildlife movements is of interest.

Level 6: Function. (i) Every representation is designed to function cognitively in particular ways so as to support the purposes for which it was developed. Moreover, (ii) in artificial entities and natural entities adapted for human purposes, function is the geospatial realization of these purposes.

Level 5: Complex objects. Entities made up of discrete or inhomogeneous parts are recognized as single objects to the extent required by the function(s) necessary to meet specific purposes.

Level 4: Simple objects. Spatially connected, homogeneous objects are categorized and named depending on their role in the context of complex objects or directly on their function. This is the lowest level at which information objects are identified as specific real-world entities.

Level 3: Classes. At this level spatio-temporal patterns and object attributes are analyzed and classified based on their measurable properties (e.g., as in automatic classification), though the available information is no longer sufficient to identify the resulting information objects with specific empirical entities.

Level 2: Observables. Crude information objects at this level only allow the qualitative knowledge that distinct kinds of relevant information exist at specific space-time points.

Level 1: Existence. By now the framework has been drained of all semantic content except for the notion that specific points of the spatio-temporal plenum are associated with information appropriate for the purposes specified at level 7.

4 Triangulating Geographic Information Science

We may now return to the questions posed in the introduction of this paper and sketch out some answers suggested by the ontological framework outlined in the previous section.

Table 1. Purpose in representations versus the purposes of artificial entities: contrasting examples.

	A road map of region X	A map of roads in region X
7 Purpose	Facilitate vehicular travel planning and navigation	Identify and mitigate barriers to wildlife movements
6 Function	Represent possible routes from place A to place B	Represent the locations where wildlife corridors intersect with roads
5 Complex objects	A road network	A wildlife corridor network
4 Simple objects	Places, freeways, arterials, collectors, intersections, ramps,	Roads, wildlife corridor segments, underpasses, culverts,

3 Classes	roundabouts,... Information objects associated through location, geometry, topology, directionality, surface, flow etc. attributes	high-conflict intersections,... Information objects associated through data on incident frequency, barrier permeability, height above ground, width, density per unit area, soil characteristics, etc.
2 Observables	Hard, rough, green, brown, wet,...	Open, blocked, green, hard, kill, dry, wet...
1 Existence	“Road-map relevant information exists here now at such-and-such appropriate granularity”	“Wildlife-corridor relevant information exists here now at such-and-such appropriate granularity”

4.1 Ontology-Epistemology

Given two arbitrary ontologies, how can we tell a priori to what extent these may be compatible?

The contents of the different levels of the semantic hierarchy and especially the relations holding among levels suggest how alternative ontologies may be mapped into that structure. Category theory provides the tools for effectuating such mappings in a computer science as well as in a mathematical context (Peirce 1991). Most object-based ontologies will map primarily into Levels 4 (simple objects) and 5 (complex objects) though they will also have parts extending to the lower (usually) and higher (rarely) levels of the hierarchy. Given any two ontologies, the extent of overlap of these mappings may be interpreted as the degree to which the corresponding ontologies are compatible (interoperable).

Given a specific ontology, how can we tell whether it is complete and internally consistent?

Mapping a given ontology into the hierarchical structure may also be used to suggest internal inconsistencies and gaps as well as opportunities for broadening the scope of that ontology. Basically, a consistent ontology will extend across a number of consecutive levels (e.g., spanning levels 3, 4 and 5 is consistent but not 2, 4 and 5 since the latter involves a jump from raw observations to objects, bypassing issues of measurement). Conversely, an ontology whereby the information objects corresponding to a given level are less developed or numerous than the ones below manifests gaps that may be filled by exploiting the full range of available lower-level objects.

4.2 Ontology-Teleology

How can we best represent artificial objects and purposeful change within the same ontological framework as natural objects and causal change? How can we seamlessly integrate the natural and the artificial or purposeful in both static descriptions and representations of change?

Artificial objects have many material properties similar to those of natural objects yet their existence and structure are incomprehensible when viewed strictly from the perspective of natural science (Simon 1969). This is because the ontological essence of artificial objects is very different from that of natural objects, the former existing only as a result of human purpose, having been designed and built to serve a specific function or goal. Natural objects can also serve a purpose to the extent that they afford a needed function, e.g. while on a walk in the country I can sit on a rock of the right size and shape for the purpose of resting. Rocks and chairs have very different origins and geometrical properties (no natural process could produce an office chair, and no natural object has the clean geometry of an IKEA piece of furniture) yet seen from Level 7 of the ontological framework these differences are minimized through a potential common purpose of interest to the user. While the treatment of change is not discussed in this paper, an analogous distinction holds between natural processes and purposeful actions, in the sense that many natural processes may be harnessed and modified for a purpose and thus become agents in a purposeful process (e.g., wind-generated electricity).

4.3 Teleology-Epistemology

How do we know that a proposed domain ontology is appropriate for particular user purposes?

Indirectly albeit insistently, this paper has stressed the related issues of user needs and fitness for use of data, which are more fully treated elsewhere in this volume (Hunter et al. forthcoming). The present paper suggests a possible approach to these questions focusing on ontology rather than on database design. Indeed, an ontological framework culminating at the level of purpose immediately suggests how such questions might be tackled. One must first identify the information objects at the top levels of the hierarchy that correspond to the purpose or purposes of interest. Then, having mapped the domain ontology in question into the hierarchical framework, one follows the relevant branching paths downward to the highest level at which elements of the ontology under consideration can be found. If the branching paths from purpose to function to complex objects etc. intersect these elements, and the ontology is internally consistent (see

above), then the ontology is appropriate for the purpose. This also suggests how the next, related question may be approached:

What is the role of cognitive stance in helping decide among competing ontologies?

We may venture that several choices are possible depending on the informational depth desired for, or required by specific applications. It may thus be sufficient in some cases to choose an ontology that describes what Bibby and Shepherd (2000) call the 'brute, unproblematic GIS objects' thoroughly and well, (Levels 4 and 5), while in other cases ontologies spanning the higher levels of semantic complexity at which function and purpose are made explicit may be necessary. This too is a purpose-oriented decision, and so, one way or another, teleology reaffirms itself.

5 Conclusion: From Triangles to Tetrahedra (perhaps) – and Beyond

The term 'information' in geographic information science makes it clear that the field is about constructing useful representations – models – of a particular kind of real-world phenomena, not about studying the phenomena themselves. Unlike geography (or, for that matter, geology, medicine, biology, chemistry, economics, and so on) geographic information science is not about describing and explaining the world, but rather about representing it in ways that are most useful to those who do the actual describing and explaining. This makes geographic information science a 'meta-' discipline, and like all information sciences dealing with representations it cannot escape philosophy (just think of the entanglement of computer science with the philosophy of mind, or the philosophical problems raised by the Church-Turing thesis). In this paper philosophy has manifested itself through the three basic questions: what (ontology), how (epistemology) and why (teleology) for geographic information science. The close relationships between these three broad fields are implicit in the manner the framework outlined in these pages was constructed, and the significance of these relationships for geographic information science was hinted at above. This short paper could not get into implementation issues but the framework generation process described is qualitatively similar to recursive decomposition and should thus be amenable to computational treatment.

To close, one definite conclusion that may be drawn from all the preceding is that there is significant foundational work yet to be done in the area of geographical ontologies. Recent emphases in ontology development that tend to privilege implementation, visualization, data mining and other more applied issues over conceptual exploration seem to suggest that the

theoretician's job is now over and that we are squarely in the era of applying and refining what we have learned. I disagree. I think that there is still tremendous scope for the investigation of the more abstract and philosophical aspects of the subject, and that both the theoretical and applied lines of ontology research will greatly benefit from being more closely intertwined with each other. This essay focused on two rather neglected perspectives, arguing for the potential contributions of epistemology and teleology to geographic ontology development. There are many more such angles, that of cognitive semantics mentioned earlier being one. Constructivism also comes to mind, a philosophical tradition reflected in the framework presented in this paper and one with close connections to the University of Vienna (see the opening quote by Heinz von Foerster), to name just one more. There would be no better way to honor Dr. Andrew Frank, a pioneer in thinking about such issues, than for the papers in this volume to reaffirm their authors' commitment to pursuing further the fundamental questions relating to the representation of the geographical world.

References

- Agarwal P (2005) Ontological considerations in GIScience. *International Journal of Geographical Information Science (IJGIS)* 19(5): 501–536
- Bibby P, Shepherd J (2000) GIS, land use, and representation. *Environment and Planning B: Planning and Design* 27(4): 583–98
- Borst WN, Akkermans JM, Top JL (1997) Engineering Ontologies. *International Journal of Human-Computer Studies* 46: 365–406
- Carton L (2007) Map making and map use in a multi-actor context: Spatial visualizations and frame conflicts in regional policymaking in the Netherlands. PhD Dissertation, Delft Technical University, The Netherlands
- Clarke AC (1968) 2001: A Space Odyssey, New American Library, New York
- Cohn AG (2008) Mereotopology. *Encyclopedia of GIS* 2008: 652
- Couclelis H (1992) Location, place, region, and space. In: Abler RF, Marcus MG and Olson JM (eds) *Geography's Inner Worlds*. Rutgers University Press, New Brunswick, NJ, pp 215-233
- Couclelis H (2002) Modeling frameworks, paradigms, and approaches. In: Clarke KC, Parks BE and Crane MP (eds) *Geographic information systems and environmental modeling*, Longman & Co, New York, pp 34–48
- Frank AU (2001) The rationality of epistemology and the rationality of ontology. Smith B and Brogaard, B (eds) *Rationality and Irrationality: Proceedings of the 23rd International Ludwig Wittgenstein Symposium*, Holder-Pichler-Tempsky, Vienna, pp 667–679
- Frank AU (2003) Ontology for spatio-temporal databases. In: Koubarakis M, Selis T, Frank AU, Grumbach S, Güting RH, Jensen CS, Lorentzos N, Manolopoulos Y, Nardelli E, Pernici B, Schek H-J, Scholl M, Theodoulidis B,

- Tryfona N (eds) *Spatiotemporal databases: The chorochronos approach*, Lecture Notes in Computer Science 2520, Springer, Berlin Heidelberg New York
- Goodchild MF, Yuan M, Cova T (2007) Towards a general theory of geographic representation in GIS. *International Journal of Geographic Information Science (IJGIS)* 21(3): 239–60
- Gruber TR (1993) A Translation Approach to Portable Ontology Specifications, *Knowledge Acquisition* 5: 199–220
- Guarino N (1999) The role of identity conditions in ontology design. In: Freksa C and Mark DM (eds) *Spatial information theory: a theoretical basis for GIS*, Proceedings of the International conference COSIT '99, Stade, Germany, Springer, Berlin Heidelberg New York, 221–34
- Howarth JT (2008) *Landscape and Purpose: modeling the functional and spatial organization of the land*. PhD Dissertation, Department of Geography, University of California, Santa Barbara, USA
- Huchard M, Rouane Hacene M, Roume C and Valtchev P (2007) Relational concept discovery in structured datasets. *Annals of Mathematics and Artificial Intelligence* 49(1-4): 39–76
- Hunter M, Bregt AK, Heuvelink GBM, De Bruin S, Virrantaus K (forthcoming) *Spatial data quality: problems and prospects*
- Kuhn W (2003) Semantic reference systems. *International Journal of Geographical Information Science (IJGIS)* 17(5): 405–409
- Kuhn W, Raubal M, Gärdenfors P (2007) Editorial: Cognitive semantics and spatio-temporal ontologies. *Spatial Cognition and Computation* 7(1): 3–12
- Peirce B (1991) *Basic Category Theory for Computer Scientists*. MIT Press, Cambridge
- Probst F (2007) *Semantic Reference Systems for Observations and Measurements*. PhD Thesis, University of Münster, Germany
- Rosenblueth A, Wiener N, Bigelow J (1943) Behavior, Purpose and Teleology. *Philosophy of Science* 10: 18–24
- Simon HA (1969) *The Sciences of the Artificial*. MIT Press, Cambridge
- Smyth CS (1998) A representational framework for geographic modeling. In: Egenhofer MJ and Golledge RG (eds) *Spatial and temporal reasoning in geographic information systems*, Oxford University Press, New York, pp 191–213
- Williamson T (1994) *Vagueness*. Routledge, London
- Zeigler BP, Elzas MS, G, Klir GJ, Oren TI (eds) (2001) *Methodology in Systems Modelling and Simulation*. North-Holland, Amsterdam

Geonoemata Elicited: Concepts, Objects, and Other Uncertain Geographic Things

Marinos Kavouras

School of Rural and Surveying Engineering, National Technical University of Athens, 9, H. Polytechniou Str., 157 80 Zografos Campus, Athens, Greece, mkav@mail.ntua.gr

Abstract

Almost thirty years after the beginning of geographic information science (GIScience) as an interdisciplinary but distinct scientific field, new and deeper research questions have arisen, questions which make us return back to the fundamental issues of geographic concepts, knowledge representation, and semantically-aware approaches. The questions are very difficult to answer, yet this should not prevent us from always pursuing the very nature of geographic meaning. It is evident that meanings and understandings in the geospatial domain (thereinafter called “geonoemata”) pivot around the connection between the central representational notions of concepts and objects. The use and application of these notions can be accounted for most problems in interoperability, non-universality of approaches, misinterpretation, and semantic conflicts. In this section, an attempt is made to identify a number of open and promising research questions in great need for progress.

1 Introduction

It is always fascinating to retrospect and realize the long distance covered by Geographic Information Systems and Science in the last 30 years at least. The path has not been an easy one, the objective was neither self-evident

nor always clear, and some prominent scientists have throughout the years (and their life) significantly contributed to what has become a renowned field today. Having appreciated this long contribution, it would be interesting but also fruitful to examine what was at times the research agenda, which questions have either found a sufficient answer or present no interest anymore, and what went right or wrong in these 30 years. But, even more importantly, it would be crucial to identify which issues remain open and long-standing, which questions still have not found a sufficient answer, and which theories will provide an essential core and make GIScience not just survive but thrive in the years to come.

Indeed, being more knowledgeable now, and after having resolved or left behind us many issues especially technical ones, it is not difficult to realize that the real problem was not the initial lack of geodata or standards. Nor it was how data are formalized, structured, or organized in a database. And clearly it was not how fast algorithms perform and can be optimized. Without diminishing their role, what really kept GIScience going and not being absorbed by other technological fields, were the open (and difficult to answer) theoretical questions. The interdisciplinary but also distinct character of GIScience has attracted scientists from many other seemingly diverse scientific disciplines.

The truth is that many of these theoretical/fundamental issues are not completely new. They were addressed early enough by some envisaged scientists, but maybe too-early to be appreciated by the large geographic information community. Two reasons changed the picture.

The first reason was the expansion of “traditional” spatial science to other disciplines: Philosophy, Epistemology, Linguistics, Cognitive Science, Psychology, Knowledge Science and Engineering, etc. Parenthesis: It could be claimed that this expansion completely reversed the classic question “What is special about spatial information?” to “How much information is left that cannot be treated as spatial?”

The second reason that helped change the picture was (a) the progress of systems, applications, sensors, and communications (including the web), and (b) the enormous acquisition of geospatial¹ data. The natural next step was to associate and integrate this information, which revealed the serious issue of incompatibility, especially that of semantic conflicts. All attempts to standardise information failed at some level, simply because you cannot completely standardize the way people observe, perceive, think, and form concepts.

In other words, what proves to be essential is the way reality is carved and represented, and how the resulting differences in meaning and knowl-

¹ Throughout the Section, the terms “geographic” and “geospatial” are used interchangeably.

edge about geospace, i.e., *geonoemata*, are elicited. It is a necessity to understand this duality relation between the representational notions of *concepts* and *objects*. This duality is met in different fields under different names, but presents a strong similarity. Whether known as <concept ↔ object>, <concept ↔ symbol>, <signifier ↔ signified>, <territory ↔ map>, <word ↔ idea>, <sign ↔ thing>, (Frege 1892; Ogden and Richards 1923; Korzybski 1933) or other, it all boils down to the fact that a representation is a substitute. It is neither the concept formed when observing a real thing (geographic reality in our case), nor the thing itself. In the same context, Frank (2000) uses a multi-agent system as a model for map production, communication, and use. The model consists of two disjoint tiers: reality and beliefs the first tier represents the environment and the second the agents' understanding and knowledge of the environment.

Semantic-aware approaches became thus a critical issue to geospatial research. The different and non-complementary representations of geographic reality (lets call it "poems") are the expression of views, conceptions, and meanings of their creators ("the poets"), which at the other end are received diversely by the different users ("poetry readers") creating different views, conceptions, and meanings. With the dramatic increase of internet users, the problem escalates. Data and its intended meaning are not confined to domain experts or the community that created it, but to a wider and a-priori unknown community of users. This often prevents the development of universal approaches, while it often leads to data misinterpretation and misuse. Without resorting to Husserl's Phenomenology and deeper uses of "noema/noesis", the systematic clarification of "geonoemata" and semantic conflicts is considered to be of vital importance towards a proper elicitation of geo-knowledge. This inevitably entails significant progress in ontological research.

The deeper questions are indeed very difficult to answer, yet this should not prevent us from always pursuing the very nature of geographic meaning. The world functions for centuries (although admittedly not well) without having resolved these notions, by making assumptions and by using axioms, limited theories, vague knowledge, heuristics, rules of thumb, trial and error, etc. Probably, the methodological instruments of GIScience need to take such approaches into consideration, be aware of all the assumptions made, and proceed to pragmatic issues.

In this section, an attempt is made to identify a number of open research questions in great need for progress. Naturally, the list itself is also open - it could not be otherwise.

2 Obstacles and Challenges

Although research on geosemantics and ontologies has tremendously increased in the last ten years, its practical results are still limited. This can be accounted to a number of reasons or obstacles (Kavouras and Kokla, 2008). The same obstacles on the other hand, present open research areas and important challenges. Three main areas are the following:

1. There is no clear understanding (not to call it lack of awareness or even confusion) first about the meaning of the notions, and secondly about the objectives set by users. That is to say, semantic conflicts are not only something between concepts, but they are also a problem at a meta level, between notions very loosely used, such as *concepts*, *objects*, *ontologies*, *semantics*, *integration*, *similarity*, *context*, etc. As a result, it is not easy to design appropriate solutions to problems.
2. Besides the significant research accomplishments in GIScience, there has not been a remarkable progress towards developing a concrete and universal *corpus of geographic knowledge*. This would require a synthesis of all independent theories and methods into a lattice of meta-ontologies which after thorough analysis and reviewing would be valuable to geo-ontology engineering.
3. The third obstacle is the lack of well established *formal instruments*, to deal with the highly demanding knowledge representation/engineering needs. Traditional methodologies, algorithms and data structures cannot play this role anymore. Given a task, the user needs to know which conceptual structure is more expressive or appropriate, how conceptual structure are converted or integrated, etc. Also, the user needs to know what options exist for the elicitation of semantic knowledge from different sources including textual descriptions.

The first two obstacles, and especially the second one, appear to be the most persistent. They involve and depend on domain (geospatial) knowledge, entailing thus research topics of long term objective but also of large value. The third obstacle does not solely refer to geospatial knowledge, and could generally advance independently. Nevertheless, geospatial knowledge may present additional requirements which general purpose conceptual instruments are unable to deal with. Finally, explication and implementation issues have been purposely left out; for they are easier to be dealt with once theoretical-conceptual issues have been clarified.

Given the above framework, in the rest of the section a number of important and promising research questions are put forward. All of them, theoretical approaches as well as formal instruments, could fall in or contribute to a main objective, that is, geo-ontologies, concepts, and semantics

as well as the development of a *corpus of geographic knowledge*. The list is only indicative and could not be closed. Some questions appear to be difficult to deal with rigorously in the foreseeable future. In these cases, there is no other way than adopting a workable solution, and working constantly towards its improvement.

3 Corpus of Geographic Knowledge

Ontological research in GIScience, as in many other disciplines, has grown substantially (see review by Agarwal 2005), yet there is still an open field of challenge. Especially, as the emphasis is put on deeper (semantic) information (Kavouras and Kokla 2008), it becomes necessary to move from the explication level higher to the ontological level. In order to do so, it is essential to build an ontological corpus of geographic knowledge.

3.1 Essence in Geographic Concepts

Concept formation is a central issue in all disciplines, GIScience not excluded. Metaphysics, Cognitive Science, and Epistemology have the potential to contribute to this objective. It is highly unlikely however that an overall encompassing theory will ever be in the position to describe different, partial, yet complimentary perspectives of geographic reality. It is nevertheless important that the essence of geographic concepts is approached at a pragmatic level.

Since concepts are the basic conceptual units of an ontology, a set of principal semantic dimensions can be set to which each concept can be projected. This set of semantic dimensions is similar to what has become known as a *semantic reference system* (Kuhn 2003). The properties, relations, and other signified elements (Kavouras and Kokla 2008) in describing a concept can prove very useful, and can be based on the universal distinctions of *top-level ontologies* (Sowa 2000). The determination of such a universal set of essential properties/relations for geoconcepts would make a great research accomplishment of immediate practical use. Several prominent initiatives for the development of a top-level ontology exist, such as IEEE Standard Upper Ontology, Upper Cyc Ontology, Basic Formal Ontology, etc. However, there is not yet an established top-level ontology which would be used as a common basis for the evolvement and association of more specialized domain-dependent ontologies. Furthermore, Agarwal (2005) argues that it is debatable whether a unified approach to geospatial ontology may exist, taking into account the interdisciplinary

nature of geographic information research, and the different conceptualizations and terminology used for the same geospatial concepts.

3.2 Principal Ontologies

This research direction focuses on the development of a set of interoperable ontologies covering the most basic/common (*principal*) concepts and relations of the geospatial domain, also called *the core*. Despite the different views on the general issue of core ontologies (Borgo and Gangemi 2004), they constitute a vital area of active research. The core should probably rely on the general structure of *top-level ontologies* and include different *views* of the domain. The existing amount of work, which otherwise could be considered significant, has not reached this stage yet. Exemplary principal ontologies, without any particular order, cover the following: *views and context, location, boundaries, spatiotemporal relations, essential properties, affordances, spatial operators, change, uncertainty, etc.* The benefit of establishing such a corpus, even an imperfect one, is manifold:

- First, it would help geospatial research move forwards, without running constantly in circles.
- Secondly, it would provide a solid basis/reference for comparing different views and resolve conflicts.
- Thirdly, it could systematise, compare and harmonise GIScience education and curricula.

3.3 Light vs. Specialised Ontologies

Traditionally, the analysis and processing of geographic information were of concern to the experts. Experts, appreciating the intricacy of the world phenomena, naturally attempt to deal with the problems as they are taught to, that is, in a complex scientific way. On the other hand, nowadays in the internet era more and more non-specialists attempt to deal with geospace and its information in a much simpler way based on “common sense”. The ontologies employed by this type of users are much lighter than those used by specialists. Research-wise, it is important to semantically strengthen these approaches. In this area, worth mentioning is the work of *Semantic Geospatial Web* (Egenhofer 2002).

The problem of supporting cognitive, common-sense, naïve approaches may be quite difficult (and thus challenging), because “common-sense” may prove to be not that “common”.

3.4 Conceptual Instruments

Besides the number of advanced conceptual structures, formalisms, and instruments developed, and their application to various fields, including GIScience, neither their suitability, nor their applicability have been sufficiently studied so far (Kavouras and Kokla 2008). Major instruments have proven to be *Formal Concept Analysis - FCA* (Ganter and Wille 1999), *Conceptual Graphs – CG* (Sowa 2000), *Chanel Theory and Information Flow - IF* (Barwise and Seligman 1997), and others. Another important challenge here is the effective synthesis of such advanced structures in frameworks. Along these lines, the Information Flow Framework (IFF) (Kent 2004) is an attempt to unify IF with FCA, while, Wille (1997) has attempted to join FCA and CGs.

3.5 Geonoemata from Natural Language

Though research on linguistic aspects of geographic space has been around for sometime (Mark and Frank 1991), traditional geospatial research has paid only limited attention to geographic knowledge and meaning elicited from natural language. Geonoemata can be systematically extracted from linguistic descriptions and formalised. This is a very exciting research direction, not only because it can explore a wealth of knowledge such as textual descriptions, but also because it has the potential to establish a “natural” and semantically rich communication between humans and machines in the future.

Advances in natural language processing/understanding (NLP/NLU), as far as geospatial concepts are involved, are absolutely necessary (Kokla 2008). An analogy with more traditional geospatial “languages” such as cartography may also prove to be useful. Some other promising NLP techniques are (Kavouras and Kokla 2008): (a) text summarization (Mani and Maybury 1999), and (b) controlled “natural” languages (Sowa 2004; Kavouras and Kontaxaki 2005).

4 Conclusions

GIScience needs not just to endure but also to establish a core identity as a real science. Developing a corpus of geographic knowledge is of utmost importance. Towards this objective, probably the most critical questions are (a) how concepts are formed, (b) how representations fulfil their role, and (c) how geonoemata are elicited. In this context, a short indicative list of challenging research areas has been introduced.

The relation their progress may have to other research areas is obvious. Such progress affects work on semantic similarity and interoperability. It also affects the way we deal with vague or imperfect geographic knowledge. The development of universal methods and tools, almost regardless to when dealing with spatial, thematic or temporal information, or when dealing with visual, cartographic, or textual descriptions of space, heavily depends on the successful tackling of the issues of the previous paragraph.

The fact that some problems are not likely to find a final solution in the near future, for they involve amongst others some deep yet unanswered philosophical issues, should not inhibit us from pursuing at least a pragmatic approach. After all, the endless effort to develop a multi-faceted yet coherent representation of geographic reality shall determine conflicts between the different geoconcepts we form, and significantly improve our communication and our (admittedly partial) understanding of this world.

References

- Agarwal P (2005) Ontological considerations in GIScience. *International Journal of Geographical Information Science (IJGIS)* 19(5): 501–535
- Barwise KJ, Seligman J (1997) *Information Flow: the Logic of Distributed Systems*, Cambridge Transactions in Theoretical Computer Science, Cambridge University Press, Cambridge (UK)
- Borgo S, Gangemi A (eds) (2004) Preface. In: *Core Ontologies in Ontology Engineering 2004 - (Un)Successful cases and best practices for ontology engineering: reusing well-founded ontologies for domain content specification*, Proceedings of the EKAW*04 Workshop on Core Ontologies in Ontology Engineering, Northamptonshire (UK)
- Egenhofer JM (2002) Toward the Semantic Geospatial Web. In: *Proceedings of the Tenth ACM International Symposium on Advances in Geographic Information Systems*, McLean, Virginia
- Frank AU (2000) Spatial Communication with Maps: Defining the Correctness of Maps Using a Multi-Agent Simulation. In: Freksa C, Brauer W, Habel C, Wender KF (eds) *Spatial Cognition II (International Workshop on Maps and Diagrammatical Representations of the Environment, Hamburg, August 1999, Lecture Notes in Artificial Intelligence, Vol. 1849, Springer, Berlin Heidelberg New York*, pp 80–99
- Frege G (1892) On Concept and Object, originally published as *Über Begriff und Gegenstand*. *Vierteljahresschrift für wissenschaftliche Philosophie* 16: 192–205, translated in Geach, Black (1952), pp 42–55
- Ganter B, Wille R (1999) *Formal Concept Analysis - Mathematical Foundations*. Springer, Berlin Heidelberg New York
- Kavouras M, Kokla M (2008) *Theories of Geographic Concepts: Ontological Approaches to Semantic Integration*, CRC Press, Taylor & Francis Group, Boca Raton, FL, USA

-
- Kavouras M, Kontaxaki S (2005) Spatial knowledge extraction from geographical databases: an approach based on the controlled English query language Geo-Q and conceptual graphs. In: Proceedings of GIS Planet 2005, Estoril, Portugal
- Kent RE (2004) The IFF Foundation for ontological knowledge organization. In: Williamson NJ, Beghtol C (eds) Knowledge Organization and Classification in International Information Retrieval, Cataloging and Classification Quarterly, The Haworth Press Inc., Binghamton, New York
- Kokla M (2008) GeoNLP: A tool for the extraction of semantic information from definitions. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XXXVII, Part B2, Beijing 2008, pp 691–696
- Korzybski A (1933) Science and sanity: An Introduction to Non-Aristotelian Systems and General Semantics.
- Kuhn W (2003) Semantic Reference Systems. *International Journal of Geographical Information Science* 17(5): 405–409
- Mani I, Maybury MT (1999) *Advances in Automatic Text Summarization*, MIT Press
- Mark DM, Frank AU (eds) (1991) *Cognitive and Linguistic Aspects of Geographic Space*, Proceedings of the NATO Advanced Study Institute, Series: NATO Science Series D, Vol. 63, Kluwer, Dordrecht
- Ogden CK, Richards IA (1923) *The Meaning of Meaning*, Routledge & Kegan Paul, London
- Sowa JF (2000) *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks Cole Publishing Co., Pacific Grove, CA
- Sowa JF (2004) Common Logic Controlled English, <http://www.jfsowa.com/clce/specs.htm>, last date accessed 01.2009
- Wille R (1997) Conceptual graphs and formal concept analysis. In: Lukose D, Delugach HS, Keeler M, Searle L, Sowa JF (eds) *Proceedings of International Conference on Conceptual Structures ICCS'97*, Lecture Notes in Computer Science, vol. 1257, Springer, Berlin Heidelberg New York, pp 290–303

Virtue Ethics for GIS Professionals

Nancy J. Obermeyer

Indiana State University

Abstract

The coalescence of the GIS profession in the U.S. became a reality with the establishment of the GIS Certification Institute (GISCI) and its formalization of a code of ethics in 2004. Since then, progress has continued, with the crafting of rules of conduct for GIS professionals, and procedures for bringing charges of unethical behavior against GISPs. Development of additional measures to encourage ethical performance by GIS practitioners continues, as case studies are currently in the works. This paper recommends that one more element be added to the ethics suite: virtue ethics. The paper links the idea of virtue ethics to the Weberian concept of professionalism and advocates a formal integration of virtue ethics in the education of GIS professionals.

1 Introduction

The coalescence of the GIS profession became official in the U.S. with the creation of the “GIS Professional” or “GISP” designation and the concurrent formalization of a code of ethics in 2004 by the GIS Certification Institute (GISCI). Similar groups in Canada, Australia, and South Africa have made similar progress. Since then, efforts to encourage ethical behavior among GIS practitioners have continued, with GISCI’s crafting rules of conduct for GIS professionals and establishing a process for handling allegations of ethical transgressions by GISPs. Additional means to encourage ethical performance include efforts to write case studies of ethical dilemmas,

which are currently in the works. This chapter recommends that the GIS community explicitly explore the concept of virtue ethics as a complement to codes of ethics and encourage its adoption among GIS practitioners.

The chapter begins with a discussion of the development of ethics within the GIS community, focusing on the activities of the GIS Certification Institute (GISCI). This section also includes a discussion of GISCI's code of ethics and rules of conduct for GISPs. Next, the chapter introduces the idea of virtue ethics, tracing its roots to Socrates, Plato, and Aristotle, and discussing its implementation in the business and public administration communities. It continues by linking virtue ethics to one of the key elements of a profession, the social ideal, as defined by Weber (1947) and Pugh (1989), and concludes by discussing the challenge of teaching and promoting virtue ethics among GIS professionals, providing recommendations based on strategies employed outside of the GIS community.

2 Development of Ethics for GIS Practitioners

The general trend in professional ethics has been toward the development and implementation of codes of ethics and rules of conduct by both professional organizations and individual corporations (Arjoon 2000). The driving forces behind these initiatives, according to Arjoon, are "...consumer pressure, the need to provide quality service, and changing social values" (2000, p 159). The GIS community has been following this trend.

As already noted, the GIS Certification Institute (GISCI), established a code of ethics when it came into existence in 2004. It is not alone, however. The Canadian Institute of Geomatics (<http://www.cig-acsg.ca/page.asp>) also has a code of ethics, as do the Spatial Sciences Institute (SSI) of Australia and New Zealand (Connolly 2007) and the Geo-Information Society of South Africa (<http://www.gissa.org.za/>). The Association for Geographic Information in the U.K. has also discussed ethics for GIS professionals (Blakemore and Longhorn 2004). As well, the International Federation of Surveyors (FIG) also has developed a code of ethics (FIG 1998).

Codes of ethics and rules of conduct are undoubtedly useful, and as Craig (1993) noted, "The GIS profession needs a code of ethics" (1993, p 13). A significant reason that this need exists is the fact that GIS is so widely used and that the potential damage from misuse of the technology and data has far-reaching ramifications, especially in public sector applications (Esnard 1998). Among other benefits, a code of ethics provides goals of behavior to which its members should aspire (Craig 1993; Onsrud 1995). As Marshall Kaplan advised in his keynote address to the Urban

and Regional Information Systems Association (URISA) in 1986: “Consider the impact of your work” (Craig 1993).

This advice raises the question of where the impacts of one’s work are felt. Both Craig (2004) and Obermeyer (1998) agree with the International Federation of Surveyors (1998) that professionals have responsibility to (1) professional colleagues and their profession; (2) employers, clients, and other funders; and (3) society, while Craig adds (4) individuals at large and Obermeyer adds (5) students. Ultimately, the GISCI code of ethics specifically details the GIS professional’s obligations to society, employers and funders, colleagues and the profession, and to individuals and society. FIG also suggests that because of the long-term reliance on land surveys (and by extension, the GIS products that use them), professionals in the field also have a responsibility to future generations (FIG 1998).

The GIS Certification Institute’s Rules of Conduct identify specific ethical violations that are drawn from the GISCI Code of Ethics. Violation of these rules may result in charges being filed with the GIS Certification Institute and set in motion a hearing which could lead to censure, revocation of GISP status, or other similar penalty, if a hearing process determines that the charges are true and significant. This linkage between action and consequences thus gives teeth to GISCI’s code of ethics. (GISCI’s first ever charges of violation of Rules of Conduct by a GISP were pending in spring of 2009, as this chapter was written.)

While having significant value, codes of ethics and rules of conduct have limitations, too. Onsrud (1995) notes that one of the shortcomings of codes of ethics is that they are devised through the efforts of the members of the profession, but rarely include input from their client groups or the general public; this produces a bias in favor of the professional group itself, thus mitigating some of the value of the code of ethics. Recognizing that many codes of ethics reflect a concern for the professional’s duties to individuals in society and to society itself (as GISCI’s Code does), the potential conflict becomes clearer. Add to this the extensive use of GIS in public sector applications and the much-circulated statistic that 80% of all government data are spatially referenced, and the absence of input from the public, or even from public officials, rises in significance.

A second critique of codes of ethics is that most codes merely “...act as a regulatory force to prevent serious lapses in ethical judgment” (Bright et al. 2006, p 250). It is virtually impossible for those who develop codes of ethics and rules of conduct to predict or anticipate every situation that professionals will face (Arjoon 2000; Allred 2002; Chun 2005; Bright et al. 2006). Professionals, especially those newly integrated into the GIS profession, may well find some guidance in achieving competent and ethical performance of their duties from a code of ethics or rules of conduct, but if new or novel situations arise, they will be on their own. In light of the relative

youth of the GIS profession, its rapid expansion, as well as the evolving field of information technology itself, with its wide-ranging implementation and the many unanswered ethical questions, this critique has clear resonance. Codes of ethics and rules of conduct are no match for the wisdom that comes from experience (Macaulay and Lawton 2006).

Another important and related critique is that the presence of a code of ethics as the sole form of ethical guidance may leave the impression that "...morality is essentially rule-following, and that the ... task is to choose the right set of rules" (Anderson 1997, p 288). Similarly, Hartman (2007, p 313) notes that "Ethics is not a science," and therefore is not easily boiled down into a few memorable commandments. Others have noted that codes of ethics tend to define a "minimum threshold for performance and decision making" (Bright et al. 2006, p 250). Macaulay and Lawton (2006, p 708) go even further and suggest that the regulatory nature of codes of conduct has "...sublimated the need for virtuous conduct," while Arjoon (2000, p 160) worries that codes of ethics could "...simply become another procedure in the mountain of bureaucracy." Consequently, codes of ethics and rules of conduct do little to inspire those who must follow them (Anderson 1997; Arjoon 2000; Chun 2005).

In the end, "...each GIS practitioner is ultimately responsible for his or her own judgments and actions" (Esnard 1998), which brings us to virtue ethics.

3 Virtue Ethics

Virtue ethics has a long history, tracing its roots back to Socrates, Plato, and Aristotle. As its name implies, virtue ethics emphasizes moral character and virtuous behavior (Hursthouse 2007). According to these enduring traditions, ethics and ethical behavior are rooted in a desire for a "good life" (Devettere 2002). Desire in this sense is stronger than a mere wish and is so strong that it leads one to initiate the behaviors needed to make this "good life" a reality. Desires, however, may be rational or nonrational (Devettere 2002).

Rational desire comes from cognitive ability which, in turn, comes from maturity. Nonrational desires, on the other hand, involve no reflection on the part of the desirer, but rather covet whatever is pleasant or appears good. Socrates believed that humans have only rational desires, and that immoral behavior is the result of a cognitive mistake rather than a deliberate choice. Plato identified three types of desires: (1)appetitive, which are associated with the pleasure drive; (2)spiritual, which are emotional and cause us to be drawn to what appears to be good; and (3)rational desires, which draw us to that which is truly good. Furthermore, he believed that

even mature humans could have nonrational desires. Like Plato, Aristotle distinguished three types of pleasure: (1) appetite (similar to Plato's appetitive); (2) emotion, which includes anger, pride, and shame; and (3) rational desire, the will to achieve what we believe is truly good (Devettere 2002).

Virtue ethics relies on three central concepts: (1) virtue; (2) practical wisdom (phronesis); and eudaimonia (happiness or flourishing). Virtue is a character trait that is deeply entrenched in the disposition of an individual, reaching to his or her core. As such, virtue transcends intellect, and calls as well on emotions and emotional reactions, choices, values, desires, perceptions, attitudes, interests, expectations and sensibilities. Moreover, virtuous people embrace the concept of virtue as its own reward (Hursthouse 2007).

Practical wisdom (phronesis) is the knowledge or understanding that enables its possessor to "do the right thing" as through having internalized good intentions that lead to good and appropriate actions. In order to apply practical wisdom to an actual situation, the individual must first develop the capacity to recognize features of situations that are morally salient. This includes recognizing that some features of a situation are more important than others. Life experience itself helps us develop practical wisdom (Hursthouse 2007).

The third premise of virtue, eudaimonia, is usually translated as "happiness" or "flourishing," but this is a value-laden notion of happiness, a true, deep happiness or feeling of well-being (not to be confused with self-satisfaction) that comes from the knowledge that one is doing his or her best to lead a "good" life. Consequently, living one's life in accordance with virtue is crucial to achieving the deep feeling of flourishing that defines eudaimonia (Hursthouse 2007).

Therefore, "Virtue ethics is about desire and not about duty, about what we want to do and not what we ought to do, about personal happiness and not the greatest happiness of all" (Devettere 2002). Similarly, Johansson (2008) suggests that the ontology of desires matters, with alteristic desires of greater importance than egoistic desires.

Bright et al. (2006, p 250) explain the link between rules of conduct and virtuousness as being on a continuum, with unethical behavior at the left side of the continuum. Codes of ethics or rules of conduct would be in the middle, roughly equivalent to "what is merely expected in ethical conduct." At the other end of the spectrum are "virtuous-driven behaviors." Toward the left side of this spectrum is a focus on the "prevention of wrong," while the right side approach seeks the "promotion of good." To accomplish this goal, it is necessary for professions or organizations to "broaden the information to which they pay attention and use more of their cognitive abilities" when making decisions with ethical implications (Bright et al. 2006, p 250).

Thus, a number of advocates of virtue ethics explicitly identify and praise its value as a complement to codes of ethics (Bright et al. 2006; Chun 2005; Arjoon 2000). Whereas codes of ethics and rules of conduct cannot foresee all potential ethical challenges, the practical wisdom that epitomizes virtue can enable professionals to make judgments individually when novel and unexpected situations arise. By developing “moral insight” (Anderson 1997), we transcend merely following rules. By analogy, if the provision of codes of ethics and rules of conduct is equivalent to giving someone a fish, then developing “moral insight” or virtue is equivalent to knowing how to fish.

Virtue ethics has been criticized on several counts. First, virtue ethics are not readily codified because they rely on the individual’s inherent nature. (Hence, the consistent advocacy of virtue ethics as a complement to codes of ethics.) Second, all cultures do not share the same set of virtues and vices. For example, some cultures value the submissiveness of women, while others encourage women to express their independence. Third, some situations have no clear ethical answer. For example, is it ethical to remove a feeding tube from a loved one who is brain dead?

Some critics have also noted a “justification problem,” related to the presence or absence of a deity. Is the individual’s ethical behavior internal and a matter of his or her nature, or is such behavior responsive to external forces and carried out because one’s deity expects it? Does it matter? The final major criticism of virtue ethics is that it seems to rely on egoism as a driving force (Devettere 2002). Virtue ethics goes beyond a written code of ethics because it does not tell professionals “what to do,” but rather seeks to guide them regarding “what to be” (Hursthouse 2007).

4 Linking Virtue Ethics to Professionalism

Linking virtue ethics to professionalism is readily accomplished through identifying the core characteristics of a profession. Weber’s seminal discussion (1947 translation) of professions paves the way for later discussions of professions within specific fields, such as that offered by Pugh (1989) and modified for the GIS community by Obermeyer (1992, 1994) and Goodchild and Kemp (1992). The key elements of a profession are (1)body of knowledge; (2)professional culture and social ideal (including a “hall of fame”); (3)professional organization, (4)shared language; and (5)a code of ethics. It is professional culture, and specifically a social ideal that conform to the notion of virtue ethics.

As Pugh (1989) defines the social ideal, it is internally driven, just like virtue ethics. The social ideal to which professionals should aspire begins with being knowledgeable, proficient and responsible. But the professional

must also be humane and dedicated to service. Weber (1947) describes the significance of the social ideal: "...an inner devotion to the task, and that alone, should lift the scientist to the height and dignity of the subject he pretends to serve." Weber's definition of the social ideal is thus consistent with virtue.

Whereas codes of ethics identify principles to follow, virtue ethics "...looks to motivate aspirational values and seeks to answer the question, 'what kind of [professionals] should we be?'" to paraphrase Chun (2005, p 269), Craig's and Onsrud's assertion that codes of ethics provide a standard to which professionals should aspire notwithstanding. Allred (2002, p 6) suggests that a key component of professionalism is "to ensure that the public interest always remains paramount," while Evetts (2003) discusses professionalism as a form of social control. In any case, by promoting an "ethos of virtuousness" (Bright et al. 2006) organizations and professions can go beyond obeying rules and toward more enlightened and far-reaching insights and wisdom in the performance of professional duties and thus be better prepared to provide ethical responses in novel and unexpected circumstances.

5 Promoting Virtue among GIS Professionals

The GISCI Code of Ethics and Rules of Conduct tell GIS professionals "what to do;" and thus help to prevent wrongdoing. Yet, this code and these rules cannot possibly identify every eventuality and circumstance that a GIS professional may encounter over the course of his or her career. Nor do they promote virtue. Virtue ethics can play a crucial role as a complement to codified rules for GIS professionals, providing guidance on "what to be" when questions arise that the rules of conduct cannot answer specifically. Thus, one of the chief criticisms of virtue ethics (that it cannot be codified) becomes an asset when a code of conduct is already in place.

The GIS profession has an important role to play in helping to promote virtue ethics among its practitioners. There is general consensus that virtue, "conceived of as a type of knowledge, or skill, can be taught" (Begley 2006). One of the most valuable means for teaching moral issues is the narrative or case study (Hartman 2008; Dawson 2005). Anderson (1997, p 287) identifies the ability of narratives to go beyond intellect to stimulate imagination and "basic-level experiences of pain, pleasure, and well-being" as a crucial reason for their value as a teaching tool. When narratives and case studies are followed up with dialogue, discussion and debate, their didactic value increases (Pass and Willingham 2009). The International Federation of Surveyors has found the strategy of case studies to be valuable, too (Greenway 2002).

Too often, however, moral education is separated from intellectual learning (Hanson 2007). While specific courses on ethics are valuable, integrating discussions of ethics into instruction on the intellectual content is even more useful. Moreover, "...teachers teach ethics all the time by modeling acceptable behavior" to their students (Pass and Willingham 2009). Ethics education is most readily achieved when it is continuous (Lynch and Lynch 2002). Thus, integrating values into GIS education is the ideal.

Only a few universities in the U.S. have begun to include ethics courses in their GIS programs; this includes Penn State University, the University of Maine, Oregon State University and a few others. Professional organizations for GIS practitioners (GISCI, URISA and others) are also playing a role in providing education in ethics through workshops offered in conjunction with professional meetings. These organizations are also in a position to publish articles highlighting ethical dilemmas that GIS professionals may face and offering guidance. For example, GISCI is currently beginning work on case study materials highlighting ethical issues for GIS practitioners (the author is a member of GISCI's Ethics committee). These efforts must continue and expand to other institutions.

A second strategy for promoting ethics among GIS professionals is through communities and organizations. Both Egan (2005) and Dawson (2005) point to the importance of communities and organizations in building a "culture of integrity" (Egan 2005). The presence of leaders and other group role models who exhibit virtuous behavior enables others to witness, imitate, and learn (Dawson 2005; Moberg 2000). Similarly, Crockett (2005, p 205) recognizes "... the pivotal role that ... leaders hold for establishing a moral legacy ... that outlives its founders in an ethos of excellence."

According to Arjoon (2000, p 171), "Ethics... is the central task of leadership, in fact, true leadership is ethical leadership." This puts the onus on leaders to promote ethical implementations of GIS. Leaders of organizations using GIS must "inspire and "mobilize" others to promote the common good (Arjoon 2000, p 170). There must be a focus not just on specific decisions in particular cases, but rather on character traits (Chun 2005), including wisdom. And while we may recognize the value of "a few virtuous individual heroes" (Chun 2005) or "white knights" (Craig 2005), it is important that individual virtue inspires "shared and distinctive organizational virtuous characteristics" (Chun 2005, p 270).

In addition to providing individual leadership, such as Chun and Craig have described, professional organizations such as the GIS Certification Institute (GISCI) and others like it, such as the International Federation of Surveyors (FIG), have a key role to play. In part, the work of these organizations in implementing codes of ethics provides a foundation and a forum

for encouraging virtue among certified GISPs. But they must go further in the area of promoting virtuousness among GIS Professionals.

Consistent with Moberg's (2000) recognition of the value of moral exemplars, and communities promoting ethics, GISCI's parent organization, the Urban and Regional Information Systems Association (URISA) has established a "GIS Hall of Fame," to recognize leadership in the GIS community. To date, eight individuals (including Ian McHarg, Roger Tomlinson, and Jack Dangermond) and one organization (Harvard Lab for Computer Graphics) have achieved this honor. GISCI has a lifetime achievement award, but to date only Roger Tomlinson has been named to this honor.

Personal examples of experienced GIS practitioners can play a role in promoting virtue ethics among those who are at an earlier stage of their careers. As Will Craig put it, after listening to the original presentation that this paper expands, perhaps we should ask, "What would Mike Goodchild do?" (One could also be secure in asking what would Will Craig do? or what would any one of a number of other luminaries of the profession do?) The practical wisdom (Hursthouse 2007) of the GIS field's luminaries is exceedingly valuable to the profession and to its practitioners and should be acknowledged, appreciated, and shared to the extent possible.

Ironically, until the coalescence of the GIS profession, virtue ethics formed the backbone of ethical training in GIS. At that time, the GIS community was far smaller, and many newcomers had the opportunity to interact with our field's luminaries first hand, learning by observation what that luminary would do in a specific situation. While the profession has gained much by the codification of ethical behavior, it would be a shame to lose the education in virtue ethics made possible by personal contact with those who have developed the practical wisdom that comes only with experience.

Honoring the ethical behavior of GIS luminaries is one way to reinforce its value. A "hall of fame" is one way to do this; a book or some other publication drawing on the practical wisdom of the field's leaders is another.

One major challenge to helping GIS professionals to learn and assimilate virtue ethics is the great variation in how and where people learn GIS. While many GIS practitioners learn GIS in a post-secondary school setting, this is not the only avenue for gaining GIS expertise. Many GIS professionals enroll in short courses offered by the developers of the software that they use on the job. Still others teach themselves by using manuals and tutorials. In all likelihood, a large proportion of GIS practitioners are also outside the reach of professional organizations as well.

Instilling ethical behavior among GIS professionals is an ongoing responsibility. While the code of ethics is a good start, we must also promote a deeper understanding through the use and teaching of virtue ethics. This

will be an ongoing task, but one that must be done for the good of the GIS profession.

References

- Allred GK (Ken) (2002) The Professional Association – Guardian of the Public Interest, http://www.fig.net/pub/fig_2002/Ts1-4/TS1_4_allred.pdf
- Anderson J (1997) What cognitive science tells us about ethics and the teaching of ethics. *Journal of Business Ethics* 16: 279–291
- Arjoon S (2000) Virtue theory as a dynamic theory of business. *Journal of Business Ethics* 28: 159–178
- Begley AM (2006) Facilitating the development of moral insight in practice: teaching ethics and teaching virtue. *Nursing Philosophy* 7: 257–265
- Blakemore M and Longhorn R (2004) Ethics and GIS: The Practitioner’s Dilemma. In: AGI 2004 Conference Workshop on GIS Ethics, London, England, October 14, 2004
- Bright DS, Cameron KS, Caza A (2006) The amplifying and buffering effects of virtuousness in downsized organizations. *Journal of Business Ethics* 64: 249–269
- Chun R (2005) Ethical character and virtue of organizations: an empirical assessment and strategic implications. *Journal of Business Ethics* 57: 269–284
- Connolly J (2007) The Place of Ethics in the Spatial Information Sciences: an Australasian Perspective. In: GIS Development August 2007 (<http://www.gisdevelopment.net/magazine/global/2007/august/26.htm>).
- Craig WJ (1993) A GIS code of ethics: what can we learn from other organizations. *Journal of the Urban and Regional Information Systems Association* 5(2): 13–16
- Craig WJ (2002) Crafting a Code of GIS Ethics. *Geospatial Solutions* November: www.geospatial-online.com
- Craig WJ (2005) White knights of spatial data infrastructure: the role and motivation of key individuals. *Journal of the Urban and Regional Information Systems Association* 16(2): 5–13
- Crockett C (2005) The cultural paradigm of virtue. *Journal of Business Ethics* 62: 191–208
- Dawson D (2005) Applying stories of the environment to business: what business people can learn from the virtues in environmental narratives. *Journal of Business Ethics* 58: 37–49
- Devettere RJ (2002) *Introduction to Virtue Ethics: Insights of the Ancient Greeks*, Georgetown University Press
- Egan MD (2005) Integrity: intention or action? *Insurance Advocate*. January 31: 23–24
- Eisenstadt SN (1968) *Max Weber: on charisma and institution building*, The University of Chicago Press
- Esnard A-M (1998) Cities, GIS, and Ethics. *Journal of Urban Technology* 5(3): 33–45

- Evetts J (2003) The construction of professionalism in new and existing occupational contexts: promoting and facilitating occupational change. *International Journal of Sociology and Social Policy* 23(4/5): 22–35
- International Federation of Surveyors (FIG) (1998) Publication No. 17: Statement of Ethical Principles and Model Code of Professional Conduct <http://www.fig.net/pub/figpub/pub17/figpub17.htm>
- Goodchild MF and Kemp KK (1992) GIS accreditation: what are the options? *ACS Bulletin* (140)
- Greenway I (2002) FIG Publication No. 29: Business Matters For Professionals: A Guide to support professionals in the task of business management. (United Kingdom: International Federation of Surveyors, FIG), <http://www.fig.net/pub/figpub/pub29/figpub29.htm>
- Hansen DT (2007) John Dewey and a curriculum of moral knowledge. *Curriculum and Teaching Dialogue* 9(1 and 2): 173–181
- Hartman EM (2008) Socratic questions and Aristotelian answers: a virtue-based approach to business ethics. *Journal of Business Ethics* 78: 313–328
- Hursthouse Rosalind (2007) Virtue ethics, *Stanford Encyclopedia of Philosophy*, available at <http://plato.stanford.edu/search/searcher.py?query=Virtue+Ethics> (7/18/2044-47.03, updated 7/18/2007)
- Huxhold WE and Craig WJ (2003) Certification and ethics in the GIS profession. *Journal of the Urban and Regional Information Systems Association* 15(1): 51–64
- Johansson I (2008) How philosophy and science may interact: a case study of works by John Searle and Hernando de Soto. In: Smith B, Mark DM, Ehrlich I (eds) *The Mystery of Capital and the Construction of Social Reality*, Open Court, Chicago
- Lynch TD and Lynch CE (2002) Virtue ethics: a policy recommendation. *Public Administration Quarterly* Winter: 462–497
- Macaulay M and Lawton A (2006) From Virtue to Competence: Changing the Principles of Public Service. *Public Administration Review* 66(5): 702–710
- Moberg DJ (2000) Role models and moral exemplars: how do employees acquire virtues by observing others? *Business Ethics Quarterly* 10(3): 675–696
- Obermeyer NJ (1992) GIS: a new profession? In: Abstracts of the 1992 meeting of the Association of American Geographers, San Diego, CA, April 15-19, 1992
- Obermeyer NJ (1994) GIS: a new profession? *The Professional Geographer* 46(4): 498–503
- Obermeyer NJ (1998) Professional responsibility and ethics in the spatial sciences. In: Taylor DRF (ed) *Policy Issues in Modern Cartography*, Elsevier, Oxford, U.K., pp 215–232
- Onsrud HJ (1995) Identifying unethical conduct in the use of GIS. *Cartography and Geographic Information Systems* 22(1): 90–97
- Pass S and Willingham W (2009) Teaching ethics to high school students. *The Social Studies* January/February: 23–30
- Pugh DL (1989) Professionalism in Public Administration: Problems, Perspectives, and the Role of ASPA, *Public Administration Review* 49(1): 1–8
- Weber M (translated by Gerth HH and Wright Mills C) (1947) *From Max Weber: essays in sociology*, Oxford University Press

Why Is Scale an Effective Descriptor for Data Quality? The Physical and Ontological Rationale for Imprecision and Level of Detail

Andrew U. Frank

Department of Geoinformation and Cartography
Technical University Vienna
Gusshausstrasse 27-29/E127
A-1040 Vienna, Austria
frank@geoinfo.tuwien.ac.at

Abstract

Observations and processing of data create data and their quality. Quantitative descriptors of data quality must be justified by the properties of the observation process. In this contribution two unavoidable sources of imperfections in the observation of physical properties are identified and their influences on data collections analyzed. These are, firstly, the *random noise* disturbing precise measurements; secondly, *finiteness of observations*—only a finite number of observations is possible and each of it averages properties over an extended area.

These two unavoidable imperfections of the data collection process determine data quality. Rational data quality measures must be derived from them: Precision is the effect of noise in the measurement. The finiteness of observations leads to a novel formalized and quantifiable approach to level of detail.

The customary description of a geographic data set by ‘scale’ seems to relate these two sources of imperfection in a single characteristic; the theory described here justifies this approach for static representation of geographic space and shows how to extend it for spatio-temporal data.

1 Introduction

Digital geographic data comes in different qualities and applications have different requirements for the quality of their inputs. In order to advance the use of digital geographic data, qualitative descriptions of the quality provided or required is necessary. Traditionally *map scale* is used to describe summarily the quality of static geographic data when cartographically represented. The reduction in size, expressed as proportional scale, causes a reduction in precision and detail. Users of maps have learned which map scales are suitable for which task: orienteering uses map in the scale range 1:10.000 to 1:25.000, for driving by car from city to city maps 1:250.000 to 1:500.000 are sufficient, etc. Repeated experiences taught us these practical guidelines and we follow them without asking for a theory.

In the age of digital data, the traditional definition of map scale, as the proportion between distances on the map and in reality, lost its justification: locations are expressed with coordinates and distances computed are in real world units. Only when preparing a graphical display, a proportional scale is used. The concept of scale in a digital world has been critically commented, but no solution suggested (Lam et al. 1992; Goodchild et al. 1997; Reitsma et al. 2003).

Data quality research needs a quantitative, theory based approach. The theory must relate to the physical characteristics of the observation process, where the imperfections in the data originate. Data quality measures, which are not related to universal properties of observation remain specific for some data collection technologies (Timpf et al. 1996) and impede the assessment of results from integration with other datasets obtained by other methods and with incompatible data quality descriptions.

In this paper I explore the process of geographic data collection and show how scale is introduced originally when making observations. It must be carried forward as a quality indicator with the dataset. The same “scale” quality measure is later used when considering whether a dataset can be used effectively in a decision situation (Frank 2008a). The tiered ontology, previously described in a number of articles (Frank 2001, 2003) is used as foundation for the analysis of the data collection process and reviewed here briefly in section 2. The tiered ontology commits to a physical reality in a space time continuum that can be observed. In a second tier physical objects are formed and a third tier includes the conventional, socially constructed objects (Searle 1995).

In this article the focus is on data quality describing physical objects. Section 3 describes the processes that are used to transform information between the tiers. The ontological approach distinguishes point observations from descriptions of objects and their attributes. The point observations are simpler than the prototypical measurements of measurement

theory (Krantz et al. 1971); this reduction to a more primitive type of observation allows to include imperfections in the theory, which classical measurement theory could not deliver (Orth 1974). The analysis leads to a quantitatively assessment of the imperfections introduced by each process (Frank 2007)—a goal desired since the data quality discussion in the mid 1980s (Chrisman 1987; Robinson et al. 1987), and the development of measurement theory (Krantz et al. 1971) but never comprehensively, systematically, and operationally achieved.

An analysis of the properties of real (physical) observation processes reveals that the limitations in the observation processes introduce two types of unavoidable imperfections: *random noise dilutes the precision* of the observation (section 4), and the *finiteness of the sensor limits the level of detail* (section 5). In a well-designed sensor these two effects are comprehensively characterized by scale. Map scale, in this definition, is therefore not an artifact of cartography but originates in the physical observation process itself; every observation introduces necessarily a scale to the data, independent of cartographic rendering.

The novel contributions of this paper are:

1. An ontology based analysis reveals universal limitations of all physical observation processes. These limitations are random noise and finiteness of sensors; quantitative descriptors of data quality must originate in these universal sources of imperfections.
2. A theory of data quality grounded in universal properties of the observation process and thus independent of technology, usable to integrate data from different sources and assess the quality of the result.
3. Introducing scale as a property of data resulting from the observation process (and not an artefact of cartographic rendering).
4. Scale is a justifiable summary description for data quality of static physical geographic datasets for observation processes that are with balanced precision and resolution. It can be extended from the spatial to the temporal dimension.

2 Tiered Ontology

An ontology describes the conceptualization of the world used in a particular context (Guarino 1995; Gruber 2005). The ontology clarifies the concepts and communicates the semantics intended by data collectors and data managers to persons making decisions with the data. Clarification of semantics is equally important for the semantics of data quality description.

Therefore, the description of data quality must be included in the ontology (Frank 2008b).

The tiered ontology used here (Frank 2001, 2003) starts with tier O, which is the physical reality, that “what is”, independent of human interaction with it. Tier O is the Ontology proper in the philosophical sense (Husserl 1900/01; Heidegger 1927, reprint 1993). The ordinary space-time continuum is assumed as the structure of physical reality.

2.1 Tier 1: Point Observations

Reality is observable by humans and other cognitive agents (e.g., robots, animals). Physical observation mechanisms produce data values from the properties found at a point in space and time. $v = p(x, y, z, t)$. A value v is the result of an observation process p of physical reality found at point (x, y, z) and time t .

Tier 1 consists of the data resulting from observations at specific locations and times (termed “point observation”); philosophers sometimes speak of “sense data” (Stanford Encyclopedia of Philosophy <http://plato.stanford.edu/>). In GIS such observations are often realized as raster data resulting from remote sensing, similarly to our retina that performs such point observations in parallel. Sensors and sensor networks (Stefanidis et al. 2005) in general produce point observations as well, but of a different kind, as will be seen in section 5.

2.2 Tier 2: Objects

The second tier is a description of the world in terms of physical objects. Objects are regions of space that have uniformity in some property. The object representation reduces the amount of data, if the subdivision of the world into objects is such that most properties of the objects remain invariant in time (McCarthy et al. 1969). For example, most properties of a taxi cab remain the same for hours, days or even longer, and they need not be observed and processed repeatedly. Only location and occupancy of the taxi cab change often.

The formation of objects—what Zadeh calls granulation (Zadeh 2002)—is a complex process of (1) determining the boundaries of objects (2) summarizing some properties for the delimited regions and finally (3) determine the type of the object (classification). For objects on a tabletop, object formation is dominated by spatial cohesion, which moves as a single piece is an object: a cup, a saucer, and a spoon (Fig. 1).



Fig. 1. Simple physical objects on a tabletop: cup, saucer, spoon

Geographic space does not lead itself to such a single, dominant, subdivision as objects typically do not move. Multiple aspects are used to form regions of uniform properties, leading to different objects overlapping in the same space (Fig. 2). Watersheds, areas above some height or regions of uniform soil, uniform land management, etc. can be identified and they all overlap (Couclelis 1992). Object classification is optimized to classify objects suitable for certain operations (hunting, planting crops, grazing cattle, etc.).



Fig. 2. Fields in a valley: multiple overlapping subdivisions in objects are possible.

2.3 Tier 3: Social Constructions

Tier 3 consists of constructs combining and relating physical objects to abstract constructs. This includes constructions like money (Fig. 3), legal marriage, ownership of land, etc. Constructed reality links a physical

object X to mean the constructed object Y in the context Z. “X counts as Y in context Z” (Searle 1995).



Fig. 3. Some pieces of metal and a piece of paper counting as money in the Czech Republic

Social constructions give meaning to physical objects or processes. Socially constructed objects can alternatively be constructed from other constructed objects, but all constructed objects are eventually grounded in physical objects. No “freestanding Y terms”, contrary to (Zaibert et al. 2004).

The present article focuses on physical observations and objects; the generalization of results to social construction is left to future work.

3 Information Processes

Any ontology for an information system that separates different aspects of reality must not only conceptualize the objects and processes in reality but must also describe the information processes that link the different conceptualizations and transform between them. This is of particular importance for an ontology that divides conceptualization of reality in tiers (Frank 2001; Smith and Grenon 2004). This transformation process introduces imperfections and is therefore responsible for the data quality.

Information processes transform information obtained at a lower tier to a higher tier (Fig. 4). All human knowledge is directly or indirectly the result of observations, transformed in chains of information processes. The processes that connect the tiers of ontology are described in this section before analyzing the limitations that produce the imperfections in the observation in the next two sections. All imperfections in data must be the result of some aspect of an information process. As a consequence, all theory

of data quality and error modeling has to be related to empirically justified properties of the information processes.

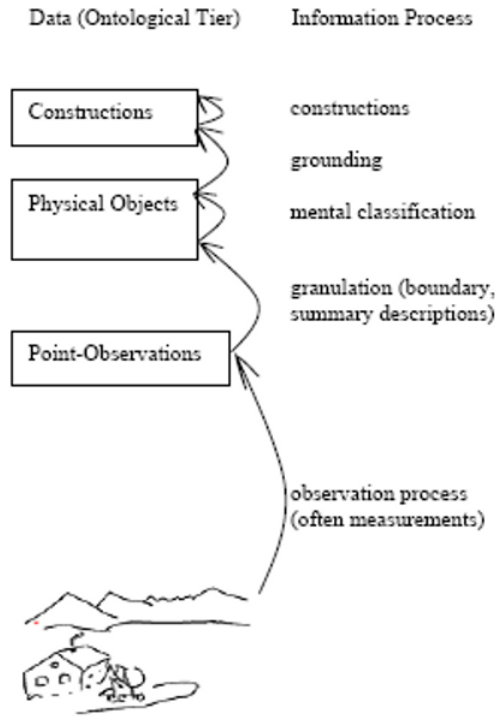


Fig. 4. Tiers of ontology and information processes transforming data between them

3.1 Observations of Physical Properties at Point

The physical process that links tier 0 to tier 1 is the observations of physical properties at a specific point. Observations are imperfect for two causes

- random noise disturbs the value produced and reduces precision, and
- finiteness of sensors force the observation not at a point but over an extended area and limits resolution.

A systematic bias can be included in the model of the sensor and be corrected by a function and is not further considered. Noise reducing precision is the focus of sections 4 and the finiteness of the sensor limiting resolution is discussed in section 5 (Fig. 5).

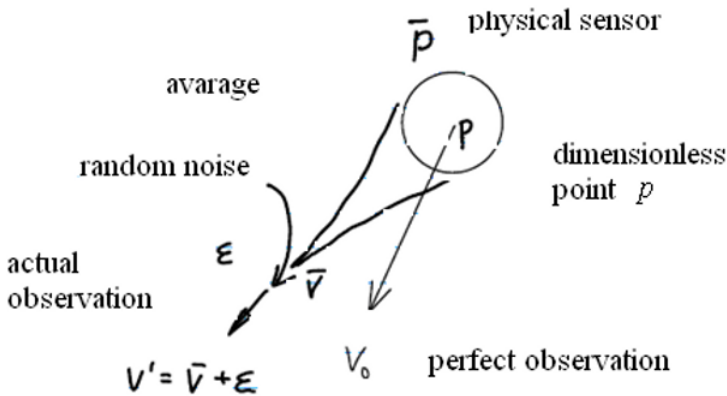


Fig. 5. Imperfections in point observations

Observations must be represented with finite length symbols. This discretization introduces yet another imperfection sometimes expressed as ‘dynamic range’. It is of minor importance because observation systems can be constructed such that this influence is negligible compared to the influences of noise (see also subsection 5.3).

3.2 Object Formation (Granulation)

Human cognition focuses on objects and object properties. We are not aware that our eyes (but also other sensors in and at the surface of our body) report point observations. For example, the individual sensors in the eye’s retina give a pixel-like observation, but the eye seems to report about size, color, and location of objects around us (Marr 1982). The observations are, immediately and without the person being conscious about the processes involved, converted to object data connecting tier 1 point observation to tier 2 object properties. Such processes are found not only in humans but higher animals also form mental representations of objects as well.

Object formation increases the imperfection of data—instead of having detailed knowledge about each individual pixel only a summary description of, for example, the object “middle wheat field” in Fig. 2 is retained. Reporting information with respect to objects results in a substantial reduction in size of the data. For example in Fig. 2, the area for the field includes approx. 1.5 million pixels each of which has 8 bits in three color changes. The compact representation as a region requires few points for

the boundary and a few bytes to describe the average color of the region. This is a computer model and not necessarily representative for processes in a human brain but gives nevertheless a general idea of the million fold compression achieved by object formation.

Object formation consists of three information processes

- boundary identification,
- computing summary descriptions, and
- classification,

which will be sketched in the following three subsections (more details in (Frank draft 2005)).

3.3 Boundary Identification

Objects are formed as regions in two-dimensional space (or 3D, 2D + T, 3D + T, etc.) that are uniform in some aspect. An object boundary is determined by first selecting a property and a property value that should be uniform across the object, similar to the well-known procedure for regionalization of 2D images. Tabletop objects in Fig. 1 are uniform in the material coherence and in their movement (Reitsma et al. 2003) the field in Fig. 2 is uniform in its color. Object formation exploits the strong correlation found in the real world; human life would not be possible, if not for most properties and places, nearby values are very similar. Which properties must be uniform to form an object is determined by the interactions intended and the situation. The focus of this article excludes a detailed discussion of processes and how they depend on properties of the object involved relating properties, object boundaries, and affordance of processes (Gibson 1986; Raubal 2002; Kuhn 2007).

3.4 Determination of Descriptive Summary Data

Descriptive values, summarize the properties of the object limited by a boundary. The computation is typically an integral (or similar summary function) that determines the sum, maximum, minimum, or average over the region, e.g., total weight of a movable object, amount of rainfall on a watershed, maximum height in country (Tomlin 1983; Egenhofer et al. 1986).

$$a_n = \iiint_{F_n} v(x, y, z) dx dy dz \quad (1)$$

where the attribute value for attribute a and object n is the integral over the 3D region F_n of the object n for the property value at $v(x, y, z)$.

3.5 Classification

Objects once identified are mentally classified. On the tabletop (Fig. 1), we see a cup, spoon, and saucer; in a landscape (Fig. 2) forest, fields, and streams are identified. Classes—similar to types in computer languages (Cardelli 1997)—indicates which operation can be performed with an object. Gibson introduced the term affordance (Gibson 1986; Raubal 2002).



Fig. 6. Pouring requires two container objects and one liquid object

Mental classification relates the objects identified by granulation processes to operations, i.e., interactions of the cognitive agent with the world. To perform an action, e.g., to pour water from a pitcher into a glass (Fig. 6) requires a number of properties of the objects involved: the pitcher and the glass must be containers, i.e., having the affordance to contain a liquid, the object poured must be a liquid, etc. I have used the term distinction for the differentiation between objects that fulfill a condition and those that do not (Frank 2006). Distinctions are partially ordered: a distinction can be finer than another one (e.g., drinkable is a subtaxon of liquid); distinctions form a taxonomic lattice (Frank 2006).

3.6 Constructions

The tier 3 contains constructed objects and actions, which are linked through granulation and mental classification to the physical reality of

physical objects and operations. They are directly dependent on the information processes described above, but details are not consequential for present purposes.

4 Random Effects on the Observations

Physical sensors are influenced by random processes that produce perturbations of the observations. The unpredictable disturbance is typically modeled by a probability distribution. For most sensor a normal (Gaussian) probability distribution function (pdf) is an appropriate choice described by expected value (mean μ) and variance (standard deviation σ).

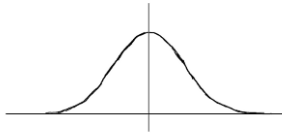


Fig. 7. Normal distribution

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2)$$

If the same observation could be repeated multiple times (which is strictly speaking not possible, because these observations would be at different times), we could compute an average and a standard deviation from the values observed.

4.1 Influence on Object Formation and Summary Values

Errors in the observation influence the determination of the object boundary. The statistical error of the boundary follows for simple cases from Gauss' law of error propagation; the standard deviation σ_f for function $f(u, v, w)$ in terms of σ_u , σ_v , and σ_w is:

$$\sigma_r^2 = \sigma_u^2 \left(\frac{df}{du} \right)^2 + \sigma_v^2 \left(\frac{df}{dv} \right)^2 + \sigma_w^2 \left(\frac{df}{dw} \right)^2 \quad (3)$$

If the observation information processes allow a probabilistic description of the imperfections of the values, then the imperfections in the object boundary and summary value are equally describable by a probability

distribution. Assuming a pdf for the determination of the boundary, one can describe the pdf for the boundary line (Fig. 8). It is an open question whether the transformations of probability density functions associated with boundary derivation and derivation of summary values preserve a normal distribution, i.e., if observation processes described by imperfections with a normal distribution produce imperfections in boundary location and summary values that are describable again by a normal distribution.

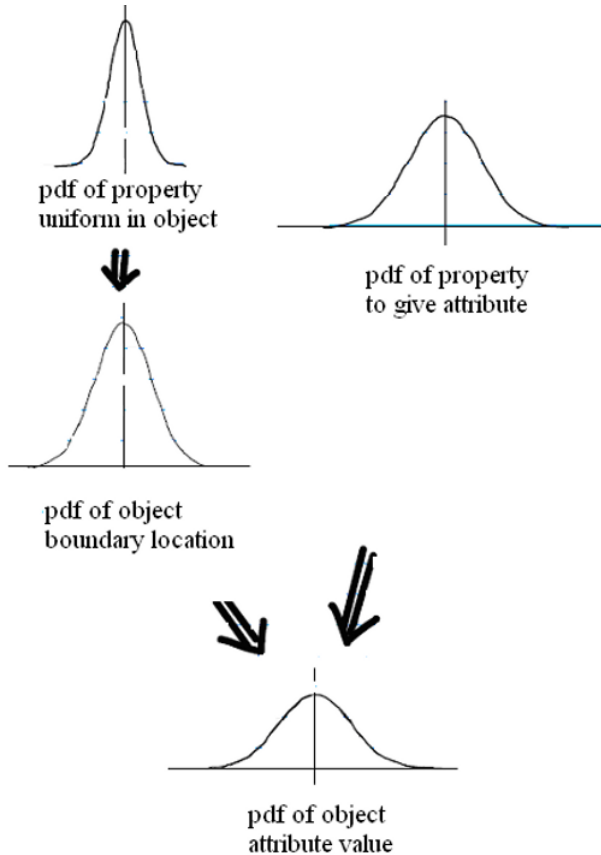


Fig. 8. The transformation of pdf from observations to object attribute values

4.2 Classifications

Distinctions describe limits in the attribute values of an object, whether the object can or cannot be used for a specific interaction and thus is mentally

classified as a particular category. The decision whether the values for an object are inside the limits or not is more or less sharp and the cutoff usually gradual (Fig. 9). The distinctions and classifications are therefore fuzzy values, i.e., membership functions as originally defined by Zadeh (1974). This is a fuzzy membership value for the category because neither the relevant attribute values of the object nor the limits for classification are known accurately. Here error propagation usually comes to an end, because the situation and the mental assessment include various correlations between the multiple relevant aspects; it is more complex than the prototypical engineering decision. Classification reduces precision to what is relevant in the context of an operation; it also reduces uncertainty and increases reliability.

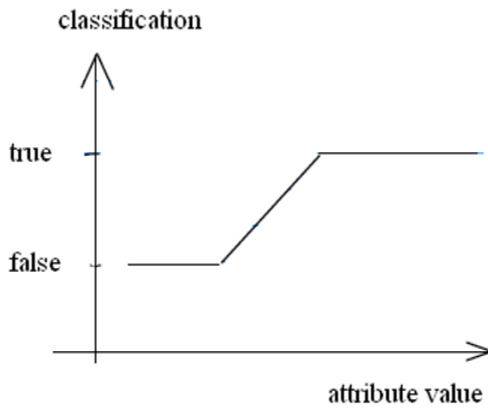


Fig. 9. Classification of objects results in fuzzy membership values

4.3 Qualitative Description

The assumption that the influence of the random effects follows a normal distribution is usually justified for physical observations. It is then reasonable to use the standard deviation σ to describe quantitatively the *precision* of observation data, even approximatively for data derived from point observations.

5 Finite Observation Devices

In this section the effects of the finiteness of physical observation instruments are discussed. The limits are threefold:

- the sensors are not infinitely small but measure over an extended area and time,
- only a finite number of observations are possible, and
- only a finite number of different readings to represent the observation value is possible.

5.1 Effects of Size of Sensor

A sensor cannot realize a perfect observation at a perfect point in space or time. Any physical observation must integrate the effects of a physical process over a region during a time. The time and region over which the integration is performed can be made very small. For example, a pixel sensor in a camera has a size of $5/1000$ mm and integrates (counts) the photons arriving in this region for as little as $1/5000$ sec but it is always of finite size and duration. The size of the area and the duration influences the result and give a “scale” to an observation.

The sensor can be modeled as a convolution of the perfect observation with a Gaussian. The effects of the size of the observation device is as unavoidable as the random perturbations of the observations, which is more widely recognized, discussed and given a formal model (summarized in the previous section). The intended point-observation $v = f(x, y, z)$ cannot be realized but effectively the device reports the average value for a small area and a small time interval $\underline{\varepsilon}$.

$$v(x, y, z, t) = \int_{-\underline{\varepsilon}}^{+\underline{\varepsilon}} f((x, y, z, t) - \underline{\varepsilon}) d\underline{\varepsilon} \quad (4)$$

where $\underline{\varepsilon}$ ranges over the size (x, y, z) and the time (t) interval used by the sensor. The region ε (in space and time) is called the support of the sensor. The imperfection in the observation process can be formalized as a special case of convolution. Convolutions are shift and linear invariant transformations, meaning shifting a signal in space gives the shifted result (5.1) and the addition of two inputs gives the addition of the two outputs (5.2) (for two signals g and h and a transformation f). This is:

$$\begin{aligned}
 g'(x) &= f(g(x)) \\
 h'(x) &= f(h(x)) \\
 g'(x-a) &= f(g(x-a)) \quad (5.1) \\
 \text{mit } \alpha g'(x) + \beta h'(x) &= f(\alpha g(x) + \beta h(x)) \quad (5.2)
 \end{aligned}$$

These seem to be fundamental rules for observation systems of a spatio-temporal reality.

The formal model is a convolution of $f(\underline{x}, t)$ with a kernel $k(\underline{\varepsilon})$

$$k_{(\underline{\varepsilon})} \text{ is } v(\underline{x}, t) = \int f((\underline{x}, t) - \underline{\varepsilon})k(\underline{\varepsilon})d\underline{\varepsilon}. \quad (6)$$

The observation $f(\underline{x}, t)$ is multiplied by the kernel value $k(\underline{\varepsilon})$, which is non-zero only for a small region around zero (the support) and for which

$$\int k(\underline{\varepsilon})d\underline{\varepsilon} = 1. \quad (7)$$

For sensors in cameras, the effect can be modeled with a convolution with the size of the sensor elements (Fig. 10a). Together with the inevitable blurring of the optical system, the imperfection of the observation can be approximated by a convolution with a Gaussian kernel (Fig. 10b).



Fig. 10. a) a pillbox function describing a camera sensor b) the Gaussian with a variance σ

5.2 Effects of Finite Number of Observations

The sampling theorem addresses another related limitation of real observations: It is impossible to observe infinitely many points; real observations are limited to sampling the phenomenon of interest at finite number of points.

Sampling introduces the danger that the observations may suggest a signal that is not present and an artifact of the sampling (Horn 1986). The sampling theorem (a.k.a. Nyquist law) states that sampling must be twice as frequent as the highest frequency in the signal to avoid artifacts (so-called aliasing). The signal must be filtered and all frequencies higher than half the sampling frequency cut off (low-pass filter). In the audio world the sampling theorem is well known, but it applies to multi-dimensional

signals as well: including sampling by remote sensing or sensor network in geographical space. It maybe appear strange to speak of spatial frequency, but it is effective to make the theory available to GIScience, where it must be applied to all dimensions observed (2 or 3 spatial dimensions and the temporal dimension).

Filtering out high frequencies, i.e., small objects in the image, must be performed before sampling. Cameras, our eyes and remote sensing devices by their construction from small sensor elements tightly packed attenuate high frequencies sufficiently to avoid aliasing modeled in a convolution with a pillbox function (Fig. 10a). The limitations of the optical system producing blur is similarly a low-pass filter, which can be modeled as a convolution with a Gaussian kernel (Fig. 10b). The two effects (finite highly packed sensors and optical blur) create observations, which are suitably filtered by a low-pass filter to avoid aliasing. Point sensors spread at a distance however do not filter high frequencies and artifacts due to aliasing may appear in the data, if frequencies higher than half the sampling rate are present in the terrain. The regular patterns of vineyards, a high frequency signal in space has been observed to produce artifacts in laser scanning data (personal communication Wolfgang Wagner January 2009).

The finiteness of observations introduces effectively a scale into the data. It limits the resolution to objects twice the size of the sampling rate. From observations at one scale more generalized data on a larger scale (i.e., cartographically speaking, a smaller scale) can be produced, approximately by convolution with a Gaussian, but data of higher resolution cannot be derived. The sampling rate effectively limits the zooming-in to obtain more detail. Proper observations avoiding aliasing can be formally modeled as a convolution with a Gaussian kernel acting as a low-pass filter followed by sampling. For this situation it appears reasonable to say that the observation has the “scale” corresponding to half the sampling frequency ν , which is the cutoff frequency of a proper low-pass filter. A numerical description of “scale” could then be $1/\nu$ with dimension time (second) and length (meter) respectively; giving the size of the smallest detail included.

It is debatable whether to call this scale, adding one more sense to the dozen already existing, or to use a term like resolution or granularity. I prefer resolution because speaking of a dataset and stating its resolution, for example as “30 m in space and 1 year in time”, extends the current usage reasonably.

5.3 Effects of Finite Representation of Observations

The representation of the observation is again finite. Only values from a range can be used. For example in photography and remote sensing, the intensity (amplitude) of the signal is given by an 8 bit value allowing values between 0 and 255 ($2^8 - 1$)—the so-called dynamic range. In a properly designed observation system the dynamic range is smaller than the precision of the sensor and has therefore a negligible effect.

5.4 Effects of Scales on Object Formation

5.4.1 Size of smallest objects detected

The scale of the observation limits the smallest object that can be detected; objects with extension in one dimension less than the scale are not observed and their extent is merged with the neighbors. This applies to small separating objects as well; roads or streams separating two fields are not picked up at large scale (low resolution) sampling and two separate fields appear as one.

5.4.2 Effects on attribute values

Attribute values are derived from two different observations: one signal is used to determine the object boundary the other is integrated over the region of the object to give the attribute value (subsection 3.4). If the two scales are comparable, the result will be meaningful at this scale. If, however, the scales are different, the interpretation of the result is difficult. The result has the larger scales of the two; it seems possible to treat this question formally in the framework designed here and I leave this as open question.

5.4.3 Effects on object classification

The scale of observation, influencing directly the object formation influences indirectly the classification in geographic data. This is relevant, where the size of an object influences the classification especially if the class is distinct by a size, e.g., small buildings vs. larger buildings. A recent study on reserves of land zoned 'residential' assumed that a building to qualify as a residential building had a minimal footprint of 60 sq.m. (Riedl 2009). In data of a scale m one does not expect objects smaller than $f \cdot m^2$, where f is the maximally expected indicating how different such objects are from a square (respective cube).

5.5 Qualitative Descriptors

The finiteness of the sensor is affecting the data in 3 ways:

- the sampling rate
- the support of the sensor
- the dynamic range of the sensor.

The support of the sensor produces in a well-designed observation system the necessary low-pass filter to satisfy the sampling theorem for the sampling rate used. In this case a description with the sampling rate in space and time alone is sufficient. Dynamic range in a well-designed sensor is such that the effects are less than the noise in the observations.

6 Scale as a Summary Description

In a well-designed observation system, the inevitable imperfections introduced by the observation system are by design balanced. There is no point in taking more samples than necessary from a band—with limited signal or observing with more precision a low resolution (band limited) signal. The appropriate relation between the low-pass filter and the appropriate precision of the observation depends on the amplitude (energy) of the signal for different frequencies. In general more precision in the observation than what is filtered away will be unnecessary.

Noting that quantitative descriptors for data quality are neither obtained nor required to be very precise, it is sufficient if the four characteristics (precision, sampling rate, highest frequency in signal, and dynamic range) of the observation system are approximately corresponding. Then they can be described with a single quantity, for which I suggest to use the term scale.

If the highest frequency in the signal is ν , corresponding to a wavelength λ , then the resolution is 2λ and the smallest objects discernable are at least of size 2λ in any dimension. The spatial-temporal precision should then be of the same order ($\sigma \approx \lambda$) and the attribute precision comparable to the amplitude in signal frequencies higher than ν .

The traditional map scale, describing the result of a cartographic rendering process, is organized around the accuracy of the human eye and limitation of the reproduction process. Assuming that one tenth of a millimeter is the graphical resolution, the “scale” describes the size of this minimal graphical element ($1/10$ to 1 millimeter) in reality. A map scale of 1:50.000 indicates that the smallest object expected is 5 m to 50 m, and precision of location is approximately the same. Spatial resolution expressed in milli-

meters (50m = 50.000 mm) gives the customary scale denominator. Different national mapping standards vary somewhat, but this describes the expectations of a map reader and his assessment what use the map is fit for realistically.

As a guideline, traditional map scale is therefore a reasonable comprehensive descriptor for the quality of a balanced data product. Combining datasets of similar “scale” produces most likely reasonable results. Combining datasets with very different “scales” requires care and the four different characteristics of observation quality must be considered separately. It is probably acceptable that for datasets for which only a summary description with scale is given, the individual characteristics are “reconstructed” assuming a balanced observation method.

7 Conclusions

Physical observation systems deviate in two inevitable and non-avoidable respects from the perfect point observation of the properties of reality:

- Random perturbation of results
- Finite spatial and temporal extent over which the observation system integrates.

Random effects are described by probability distribution function pdf and the propagation of these follow in simple cases Gauss’ Law of error propagation, in general transformations for the probability distribution must be computed.

The effects of finite support for the observation can be modeled as a convolution with a Gaussian Kernel and the non-zero extend of the kernel determines the “scale” of the observation. The signal must be filtered before the sampling with a low-pass filter to cut off all frequencies above half the sampling frequency. The typical sensors for remote sensing or photogrammetry achieve this and can be modeled as convolution with a Gaussian kernel.

In a balanced, well-designed observation system, the attenuation of frequencies above the half the sampling frequency is sufficient to avoid artifacts in the result (aliasing). Precision for the signal corresponds to the resolution.

For a dataset obtained with a well-balanced observation system, a characterization of the quality by a spatial and temporal scale is reasonable. It allows decision by users about the uses the data is fit for. A detailed analysis is necessary if multiple signals are observed by different observation systems, which is the regular case for GIS. The improvements of interoperability

allow the use of datasets from different sources. If they are combined, the analysis must detail the four characteristics for each signal and consider its combination. The limitations resulting from analog overlay of detailed and less detailed maps are known—they are not less severe in a digital system but less visible. The description given here shows how they can be dealt with analytically.

Using a tiered ontology where point observations are separated from object descriptions allows to follow how the imperfection introduced by random error and scale propagate to objects and their attributes. The analysis of the physical observation process and its formalization as a combination of random noise and a convolution with a Gaussian Kernel opens the door for a formal treatment of the effects in particular situation. Recommended is research to understand how data of different scale interacts and how from a dataset with small scale a dataset with a larger scale can be derived. Previous research by Openshaw et al. (1987) on the modifiable areal unit problem (MAUP) documents the importance of the question. The framework allows a formal treatment, but does not answer the question what the right scale to describe a phenomenon is. A recent paper by Gabora, Rosch and Aerts (2008) discusses the transformation of concepts (classes) between contexts. Changes in scale can be modeled as scale change and it appears promising to see if the approach of Gabora et al. applies.

Information is used to make a decision; this may be a simple, everyday decision in street navigation—“do I turn right or left here?”—decisions leading to important and complex actions—“is a new hospital building at location X necessary?”—or even indirectly connected to action as in the testing of scientific hypothesis that leads to scientific rules. A decision can always be reduced to a binary question and thus brought to a comparison of two values, from which a yes or no answer follows. Formally a decision is described as a test: $R - S > 0$. If imperfections affect R and S and formal models exist for these imperfections, the imperfection of the decision—i.e., the risk that the decision is wrong—can be tested. In particular, the scale for the observation of R and S must be comparable; this means for ordinary suplications that the scale of the observation should be comparable to the scale of the action we decide on.

Scale effects in geographic data are not yet well understood, despite many years of being listed as one of the most important research problems (Abler 1987; NCGIA 1989b; NCGIA 1989a; Goodchild et al. 1999). It is hoped that the conceptual clarification achieved here and the formalization using convolutions contributes to advancing research in scale effects in information processes.

Acknowledgements

These ideas were developed systematically for a talk I presented at the University of Münster. I am grateful to Werner Kuhn for this opportunity.

References

- Abler R (1987) Review of the Federal Research Agenda. In: International Geographic Information Systems (IGIS) Symposium (IGIS'87), The Research Agenda, Arlington, VA
- Cardelli L (1997) Type Systems. In: Tucker AB (ed) Handbook of Computer Science and Engineering, CRC Press, pp 2208–2236
- Chrisman N (1987) Fundamental Principles of Geographic Information Systems. In: Auto-Carto 8, Baltimore, MA, ASPRS & ACSM
- Couclelis H (1992) People Manipulate Objects (but Cultivate Fields): Beyond the Raster-Vector Debate in GIS. In: Frank AU, Campari I, Formentini U (eds) Theories and Methods of Spatio-Temporal Reasoning in Geographic Space, Springer, Berlin Heidelberg New York, LNCS 639, pp 65–77
- Egenhofer MJ, Frank AU (1986) Connection between Local and Regional: Additional “Intelligence” Needed. In: FIG XVIII International Congress of Surveyors, Toronto, Canada (June 1-11, 1986)
- Frank AU (2006) Distinctions Produce a Taxonomic Lattice: Are These the Units of Mentalese? In: International Conference on Formal Ontology in Information Systems (FOIS), Baltimore, Maryland, IOS Press
- Frank AU (2001) Tiers of Ontology and Consistency Constraints in Geographic Information Systems. *International Journal of Geographical Information Science (IJGIS)* 75(5 (Special Issue on Ontology of Geographic Information)): 667–678
- Frank AU (2003) Ontology for Spatio-Temporal Databases. In: Koubarakis M, Sellis T, Frank AU, Grumbach S, Güting RH, Jensen CS, Lorentzos N, Manolopoulos Y, Nardelli E, Pernici B, Schek H-J, Scholl M, Theodoulidis B, Tryfona N (eds) *Spatiotemporal Databases: The Chorochronos Approach*, Springer, Berlin Heidelberg New York, pp 9–78
- Frank AU (2007) Data Quality Ontology: An Ontology for Imperfect Knowledge. In: Winter S, Duckham D, Kulik L, Kuipers B (eds) *Spatial Information Theory, 8th International Conference, COSIT 2007*, Melbourne, Australia, September 19-23, 2007, Proceedings, Lecture Notes in Computer Science 4736, Springer, Berlin Heidelberg New York, pp 406–420
- Frank AU (2008a) Analysis of Dependence of Decision Quality on Data Quality. *Journal of Geographical Systems* 10(1): 71–88
- Frank AU (2008b) Data Quality - What Can an Ontological Analysis Contribute? In: *Spatial Accuracy Assessment in Natural Resources and Environmental Sciences 2008*, Shanghai, China, WorldAcademicPress
- Frank AU (draft 2005) *Ontology for GIS*. Vienna, Technical University Vienna, Institute for Geoinformation and Cartography

- Gabora L, Rosch E, Aerts E (2008) Toward an Ecological Theory of Concepts. *Ecological Psychology* 20(1): 84–116
- Gibson JJ (1986) *The Ecological Approach to Visual Perception*, Hillsdale, NJ, Lawrence Erlbaum
- Goodchild MF, Egenhofer MJ, Kemp KK, Mark DM, Sheppard E (1999) Introduction to the Varenius Project. *International Journal of Geographical Information Science (IJGIS)* 13(8): 731–745
- Goodchild MF, Proctor J (1997) Scale in a Digital Geographic World. *Geographical & Environmental Modelling* 1(1): 5–23
- Grenon P, Smith B, Goldberg L (2004) Biodynamic Ontology: Applying BFO in the Biomedical Domain. In: Pisanelli DM (ed) *Ontologies in Medicine*, IOS Press, Amsterdam, pp 20–38.
- Gruber, T. (2005). "TagOntology - a way to agree on the semantics of tagging data." Retrieved October 29, 2005., from <http://tomgruber.org/writing/tagontology-tagcapm-talk.pdf>.
- Guarino, N. (1995). "Formal Ontology, Conceptual Analysis and Knowledge Representation." *International Journal of Human and Computer Studies*. Special Issue on Formal Ontology, Conceptual Analysis and Knowledge Representation, edited by N. Guarino and R. Poli 43(5/6).
- Heidegger, M. (1927; reprint 1993). *Sein und Zeit*. Tübingen, Niemeyer.
- Horn, B. K. P. (1986). *Robot Vision*. Cambridge, Mass, MIT Press.
- Husserl (1900/01). *Logische Untersuchungen*. Halle, M. Niemeyer.
- Krantz DH, Luce RD, Suppes P, Tversky A (1971) *Foundations of Measurement*. New York, Academic Press
- Kuhn W (2007) An Image-Schematic Account of Spatial Categories. In: Winter S, Duckham D, Kulik L, Kuipers B (eds) *Spatial Information Theory, 8th International Conference, COSIT 2007, Melbourne, Australia, September 19-23, 2007, Proceedings, Lecture Notes in Computer Science 4736*, Springer, Berlin Heidelberg New York
- Lam N, Quattrochi DA (1992) On the issues of scale, resolution, and fractal analysis in the mapping sciences. *The Professional Geographer* (44): 88–98
- Marr D (1982) *Vision*. New York, N.Y., W.H. Freeman
- McCarthy J, Hayes PJ (1969) Some Philosophical Problems from the Standpoint of Artificial Intelligence. In: Meltzer B, Michie D (eds) *Machine Intelligence 4*. Edinburgh, Edinburgh University Press, pp 463–502
- NCGIA (1989a) The U.S. National Center for Geographic Information and Analysis: An Overview of the Agenda for Research and Education. *International Journal of Geographical Information Science (IJGIS)* 2(3): 117–136
- NCGIA (1989b) Use and Value of Geographic Information Initiative Four Specialist Meeting, Report and Proceedings, National Center for Geographic Information and Analysis; Department of Surveying Engineering, University of Maine; Department of Geography, SUNY at Buffalo
- Openshaw S, Charlton M, Wymer C, Craft A (1987) A Mark 1 Geographical Analysis Machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems* 1(4): 335–358
- Orth B (1974) *Einführung in die Theorie des Messens*. Verlag W. Kohlhammer, Stuttgart, Berlin, Köln, Mainz

- Raubal M (2002). *Wayfinding in Built Environments: The Case of Airports*. Münster, Solingen, Institut für Geoinformatik, Institut für Geoinformation.
- Reitsma F, Bittner T (2003) Process, Hierarchy, and Scale. In: *Spatial Information Theory, Cognitive and Computational Foundations of Geographic Information Science, International Conference COSIT'03*
- Riedl M (2009) Erstellung von Baulandbilanzen in Tirol. In: 15. Internationale Geodätische Woche Obergurgl, Ötztal Tirol, Wichmann
- Robinson V, Frank AU (1987) Expert Systems Applied to Problems in Geographic Information Systems: Introduction, Review and Prospects. In: *Auto-Carto 8*, Baltimore, MA, ASPRS & ACSM
- Schneider M (1995) *Spatial Data Types for Database Systems*. Hagen, FernUniversität
- Searle JR (1995) *The Construction of Social Reality*. New York, The Free Press
- Stefanidis A, Nittel S (2005) *Geosensor Networks*. Boca Raton, Florida: CRC Press
- Timpf S, Raubal M, Kuhn W (1996) Experiences with Metadata. In: 7th Int. Symposium on Spatial Data Handling, SDH'96, Delft, The Netherlands (August 12-16, 1996), Faculty of Geodectic Engineering, Delft University of Technology
- Tomlin CD (1983) *A Map Algebra*. Harvard Computer Graphics Conference, Cambridge, Mass.
- Zadeh LA (1974) Fuzzy Logic and Its Application to Approximate Reasoning. In: *Information Processing*, North-Holland Publishing Company
- Zadeh LA (2002) Some Reflections on Information Granulation and Its Centrality in Granular Computing, Computing with Words, the Computational Theory of Perceptions and Precisiated Natural Language. In: *Data Mining, Rough Sets and Granular Computing*, Heidelberg, Germany, Physica-Verlag GmbH
- Zaibert L, Smith B (2004) Real Estate - Foundations of the Ontology of Property. In: Stuckenschmidt H, Stubkjaer E, Schlieder C (eds) *The Ontology and Modelling of Real Estate Transactions: European Jurisdictions*, Ashgate Pub Ltd, pp 35–51

Semantic Engineering

Werner Kuhn

Institute for Geoinformatics (ifgi), University of Muenster
Robert-Koch-Str. 26-28, D-48149 Muenster (Germany)
kuhn@uni-muenster.de

Abstract

The chapter shows how minimal assumptions on difficult philosophical questions suffice for an engineering approach to the semantics of geospatial information. The key idea is to adopt a conceptual view of information system ontologies with a minimal but firm grounding in reality. The resulting constraint view of ontologies suggests mechanisms for grounding, for dealing with uncertainty, and for integrating folksonomies. Some implications and research needs beyond engineering practice are discussed.

1 Introduction

Many computer scientists, geographers, geoscientists, cognitive scientists, philosophers, and knowledge engineers are concerned today with solving semantic problems posed by data about the environment. Their work has diverging goals and heated debates often ensue on foundations. For example, in *Beyond Concepts: Ontology as Reality Representation*, Barry Smith (2004) exposes some confused uses of the term *concept* in ontology. He proposes to replace concepts as the subject matter of ontologies by “the universals and particulars which exist in reality” and goes on to show that this choice yields a more precise understanding of foundational ontological relations, such as *is-a* or *part-of*. While he demonstrates the value of **distinguishing** universals and particulars, his arguments do not support

abandoning the notion of concept, as elusive or abused as it may be. Debates on *universal* and *particular* are older and not easier to settle than those on *concept*. For example, the question what it means for particulars (such as Lake Constance) or universals (such as lake) to “exist in reality” remains unsettled. Thus, Smith’s critique of concepts is mainly a (justified) exposure of some sloppy language use and modeling.

In defense of concepts, and in an information system context, this chapter advocates a pragmatic stance and an engineering view of semantics. A vast body of literature on ontology engineering for conceptual modeling (see, e.g., Guarino and Weltri 2002; Guizzardi and Halpin 2008) shows how productive it can be to avoid throwing out the baby of concepts with the bathwater of its abuses. I will argue that this is so because

1. information system ontologies are only meant to *constrain the use and interpretation* of terms; they do not specify “the meaning” of these terms, much less “the existence” of universals and particulars in reality;
2. ontological constraint networks are *groundable* in physical properties of the environment; for semantics, no other assumptions are needed about reality.

These two assumptions support a linguistic *and* an engineering reading of concepts, make these two views compatible with each other, and anchor ontologies in reality. They commit to a mind-independent reality, but one in which no objects, universals or particulars need to be posited, only stimuli, which humans can detect and build concepts from.

The first assumption recalls Guarino’s characterization of an ontology as “a set of logical axioms designed to account for the intended meaning of a vocabulary” (Guarino 1998). However, following the saying that “words don’t mean, people do” and Putnam’s arguments that meaning is not an object (Putnam 1975), I consider meaning to be a process. Furthermore, I treat this process as an engineering artifact. Similar to the processes running in a chemical plant, meaning processes can then be constrained in how they run: what people mean when they use a term, and how others interpret the term, can be described and influenced. Dictionaries or feature attribute catalogues, for example, constrain the uses and interpretations of words or geodata, respectively.

The second assumption ties ontological constraints to reality. Instead of a simplistic correspondence between terms and objects in reality, which is clearly untenable, it suggests a minimal and sufficient grounding of ontological constraint networks in elementary physical properties of the world. A related paper (Schneider et al. 2009) demonstrates and formalizes this grounding process, drawing heavily on Gibson’s meaningful environment

(Gibson 1986). Here, we will just posit the grounding capability as such and relate it by analogy to the grounding of geodetic networks.

Based on these two core assumptions, the chapter lays out an engineering view of semantics. The view has its roots in ontology engineering, but has a purely semantic purpose. It puts concepts (which are considered to be always associated with terms) at the center of attention and acknowledges that their descriptions are necessarily incomplete. Its goal is to enable information users and providers to constrain the uses and interpretations of their terms. *A semantic engineer designs processes of language use and interpretation.*

The chapter first shows that concepts can be treated as symbolic and social entities subject to constraints (section 2). Then, it explains the resulting view of ontologies as constraint networks (section 3), an understanding of grounding resulting from it (section 4), an integration strategy for folksonomies (section 5), and a mechanism for dealing with uncertainty (section 6). It concludes with a discussion of research challenges (section 7).

2 An Engineering View of Concepts

As Smith (2004) states, the lack of convincing definitions of *concept* and related terms like conceptualization is partly due to “the fact that these terms deal with matters so fundamental to our cognitive architecture (comparable in this respect to terms like ‘identity’ or ‘object’) that attempts to define them are characteristically marked by the feature of circularity.” Replacing *concept* by *universal* and *particular*, however, does not solve this problem, as the age-old debates on realism, nominalism, and conceptualism show. For solving semantic problems, it may be more productive to agree on minimal requirements imposed on the notion of concept. This is what I attempt to do here, limiting foundational claims to relatively uncontroversial ones, and not attempting a formal definition of *concept*. This section states these claims and proposes an interpretation of the popular semantic triangle capturing them.

2.1 A Triadic Notion of Concepts

For the purposes of semantic engineering, it is necessary and sufficient to posit a threefold nature of concepts, involving

- *terms* (symbols, words, expressions),
- which evoke and *express ideas* (thoughts) and
- are used to *refer to reality*.

For example, in speakers of English, the word *lake* evokes an idea of a water body connected to other water bodies and having properties like a relatively flat surface and a water depth. The word may be used to refer, for instance, to that large amount of blue, wet substance near Constance, as an instance of the kind (*lake*) or as a named individual (*Lake Constance*). For German speakers, the words *See* and *Bodensee* play the same roles, respectively.

Concepts are considered here to be associated with terms in a language, not detached from them, so that the English word *lake* and the German word *See* belong to two different concepts, regardless of whether they are used to refer to the same parts of reality or not (see also Mark 1993). The German term *Begriff* may make this close association between words and thoughts more explicit than the more abstract English term *concept*; its root *begreifen* (touch) furthermore points to the embodied nature of concepts.

2.2 The Semantic Triangle Revisited

The triadic notion of concepts goes back as far as Aristotle (Sowa 2000) and is often represented by a semantic (or semiotic, or meaning) triangle (Ogden and Richards 1923), with one corner for each of the three aspects (Fig. 1).

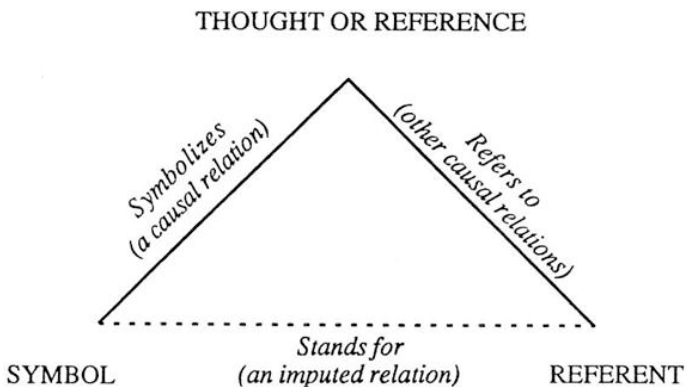


Fig. 1. A form of the meaning triangle, adapted from (Ogden and Richards 1923)

Note that Ogden and Richards (1923) avoid the term “concept” in their triangle altogether and point to many imprecise uses of the term. For the purpose of semantic engineering, it is sufficient to posit that the corners of the meaning triangle represent the three aspects of concepts identified above.

Many discussions of the semantic triangle make assumptions about each corner that are difficult to justify. For example, they label the top corner

“Concept”, claim that the REFERENT corner represents “Objects”, and define SYMBOLS such as to carry some fixed meaning in and of themselves. These and other assumptions about meaning are unnecessarily strong and probably wrong. The following discussion relaxes some assumptions, makes the remaining ones more precise, and emphasizes the social embedding of the triangle, which is normally neglected in its discussion.

First, the top corner of the triangle shall, for our purposes, remain a black box to which semantic engineers have no access, except by assuming that the use of symbols expresses and evokes some thoughts in some people, and that these thoughts are shaped by observations of reality.

Second, the right corner of the triangle is understood here as anything external to minds that is shared in an information (or language) community. It provides physical stimuli, which people can observe and agree on. Reality, in this context, is what we observe and what we talk about using symbols. Objects, particulars, and universals do not need to be assumed to exist in a mind-independent reality; they can be cognitively or socially constructed. This position does not exclude stronger claims about reality, but avoids the two-fold circularity of defining universals “as that in reality to which the general terms used in making scientific assertions correspond” and particulars as “the instances of such universals” in (Smith 2004).

Third, the left corner contains the symbols of the language whose semantics is in question, including symbols denoting relationships (such as flowing into). For artificial languages, like those of information systems, one can *design* these symbols as well as the conditions for their use. This language design capacity reinforces the idea of semantic engineering.

The *edges* of the triangle represent relations between the corners, generated by human activities:

1. People *observe* reality and form thoughts; for example, the repeated occurrence of extended horizontal water surfaces may suggest a category of lakes.
2. People *express* thoughts through symbols; for example, one may notice a water surface, while flying over it, and say “we are flying over a lake”.
3. Communication succeeds when others interpret how the symbols *refer* to reality; for example, a seat neighbour on the plane might respond “it must be Lake Constance”.

With the edges of the triangle representing *many-to-many* relations, perspectivalism and polysemy are fully admitted, as they should be. Reality induces all kinds of thoughts, depending on perspectives taken, which are in

turn expressed in multiple ways as symbols, and the symbols are used to refer to reality in many ways, even within a single language community.

The triadic notion of concepts avoids their reduction to purely linguistic or purely mental entities. The symbolic and mental sides are necessary and inseparable components of concepts and the reference to reality grounds them. The concepts constructed by a language community are not arbitrary, and they are not just entities created by modelers. Rather, they are what Smith calls “tools (analogous to telescopes or microscopes) which we can use in order to gain cognitive access to corresponding entities in reality” – except that *some of* the corresponding entities are mentally and socially constructed, while others are directly observable.

Finally, the proposed view of concepts also acknowledges their social aspects. In information systems, as well as in communication in general, terms can only be used to refer to something if there is a language community establishing and sustaining this use. The semantic triangle does not make this social aspect explicit, which is one of its weaknesses. Implicitly, however, all its corners and edges require concepts to be situated in a community sharing a language (or parts of it).

3 Ontologies as Networks of Constraints

Ontologies, in our semantic engineering view, constrain the use and interpretation of terms in an information community. For example, a hydrology ontology constrains how terms like *lake* or *waterbody* should be used and interpreted. The non-logical symbols of an ontology stand for concepts and relations and its logical sentences constrain these. For instance, the constants *lake* and *waterbody* in the sentence

lake is-a waterbody

stand for the concepts *lake* and *waterbody*, respectively. By committing to the ontology, a hydrological information community constrains these two concepts through the *is-a* relation. The consequence is that anything stated about water bodies applies also to lakes. For example, the ontology could state that

waterbody has-a waterdepth

and thereby also constrain all lakes to have a water depth quality. By adding more and more sentences, such as

river is-a waterbody

a network of constraints is incrementally being built up, narrowing the possible interpretations and uses of terms.

Some symbols may be introduced in the ontology for completeness or convenience, without necessarily expressing a domain notion. For example, in a sentence

every river flows_into a waterbody

the relation *flows_into* may express a notion of flowing into used in the hydrology community, but it may also be an auxiliary concept used in the ontology only.

While the notion of *concept* is typically reserved for universals in the literature, ontologies can also constrain terms for particulars, in sentences like

LakeConstance instance-of lake

or

Rhine flows_into LakeConstance.

This generalization allows for reasoning on individuals in the ontology, not just in a database or GIS, where this kind of reasoning is typically (and often more efficiently) performed. Gazetteers are a good example for the need of a combined reasoning on universals and particulars (Janowicz and Keßler 2008). Also, ontological specifications of geographic kinds, like lake or mountain, may refer to the (individual) surface of the Earth, of which all their instances are parts.

The semiotic function of ontologies themselves (representing concepts in logical languages) does not require the second meaning triangle that is sometimes proposed (Sowa 2000). The symbols of the ontology can be taken to *be* the symbols of the object language (for example, of hydrology terms), or syntactic variants of them, expressing the same thoughts and referring to the same reality.

4 Grounding Constraint Networks

Treating ontologies as networks of constraints can give us an understanding of what it means to ground them. The nodes and edges of a network of concept specifications can be further constrained by observations. For example, the node *lake* in the ontology can be tied to polygons representing lakes in a GIS database, as proposed in (Bennett et al. 2008), or the node *waterdepth* can be tied to an observation procedure, as in (Schneider et al. 2009). Such observational information is then propagated through the network and further restricts possible interpretations of all connected terms. If

it is supplied in symbols that are grounded in physical reality, this grounding propagates through the network.

Grounding is a process of adding information on the variables of a conceptual network through observations anchored in physical stimuli. This idea concurs with Quine's notion of observation sentences (Quine 1960). Anchoring typically occurs in measurement units, other reproducible conventions about measurement (such as agreements on zero values), and fundamental observable properties of the environment (like the fact that two different media are separated by a surface). While a complete theory of ontology grounding remains to be worked out, I will explain the main idea here using a geodetic analogy. A worked out example and the relation to environmental psychology are presented in (Schneider et al. 2009).

Geodesists are familiar with the idea of grounding a constraint network: using triangulation networks, they compute coordinates from observations of distances and directions. The distance and directions are expressed as constraints, which are parameterized in the coordinates and thereby map the coordinate space to an observation space (Vaníček et al. 1982). The networks are grounded through measurement units and externally supplied coordinate values, which are both anchored in the physics of the earth.

The grounding of triangulation networks is called a *geodetic datum*. It ties the social constructions of coordinate systems (in particular, their equator and zero meridian) to the body of the earth. Broadly speaking, the earth's shape determines the ellipsoid on which the coordinates are defined, the mass center anchors it in space, and the rotation axis orients it. The essence of this grounding scheme is to achieve *reproducible* interpretations of coordinates: one can take any coordinates and reconstruct the corresponding real-world location, at least in principle.

A geodetic datum determines the interpretations of the coordinate *concepts* used to describe location, i.e. of coordinates as such, not just of particular coordinate values. It explains the notions of latitude and longitude operationally, by giving a recipe of how they are measured. Seen as networks constraining concepts, triangulation networks constrain the interpretation of coordinates, distances and directions. Practically, these pose no semantic problems, since methods exist to compute the necessary interpretations. For coordinates, these methods are the coordinate reference systems (ISO 2002) commonly used in GIS and other geospatial information technology. For distance and direction measurements, they are the SI system of measurement units (SI). The former, of course, are themselves anchored in the latter.

Conceptually, this interpretation procedure for coordinates supplies an analogy for interpreting terms like *lake* or *waterdepth*. Both kinds of interpretation processes, geodetic and general, first map the symbols to observations and then ground this mapping in physical reality. Such an analogy

is at the heart of the notions of a semantic reference system and semantic datum introduced in (Kuhn 2003; Kuhn and Raubal 2003). Here, we have extended it to a constraint view of ontologies, to clarify the notion of ontology grounding.

It should be noted that grounding can never be absolute. A geodetic datum rests on geophysical models (e.g., for the mass distribution of the earth) and on astronomical frames of reference (star positions). Strictly speaking, these assumptions harm the reproducibility of coordinate positions. More generally, a semantic datum can only shift the need for interpretation to a reference frame at the next level. Practically, this shift should (by design) solve most semantic problems. Philosophically, however, the caveat may be useful to consider by ontologists making stronger assumptions about reality.

5 Integrating Folksonomies with Ontologies

A further gain of a constraint view of ontologies is that it connects ontologies to folksonomies. Folksonomies are non-hierarchical lists of keywords (tags) linked to information resources. For example, a web site describing a bicycle tour around Lake Constance¹ has been tagged with the terms *cycle*, *tour*, *austria*, *bregenz*, *lake*, *constance* by a user of the social bookmarking site *delicious.com*².

Folksonomies are not related to taxonomies, despite their name, but provide data about the terms people associate with contents. They do not contain logical axioms, but tuples linking terms to resource identifiers (and to the tagging users). These tuples constrain interpretations of the terms used as tags, by showing their use and its evolution over time. They constrain the interpretations bottom-up, complementing the prescriptive top-down constraints of ontologies. For example, *delicious.com* reveals what contents are tagged by terms like *lake* and/or *river*.

Folksonomies are easy to generate and use, do not burden users or producers with difficult modeling tasks, and clearly have something to tell us about the semantics of their terms. They are, in fact, an increasingly popular form of empirical semantic data. Other such forms come from data mining and similar knowledge extraction methods. All these inductive approaches to semantics play a key role in the automated learning and maintenance of terminological constraint systems.

¹ http://www.bicyclegermany.com/lake_constance.htm

² <http://delicious.com/url/a8dccabc65ed02711e150a743f226fff?show=all>

6 Accommodating Uncertainty

Constraint networks provide great flexibility in information handling, by admitting any number of constraints on any of their variables. As a consequence, they have to provide methods to accommodate uncertainty, since the stated constraints may over- or underdetermine an exact solution. In the case of triangulation networks, as well as for many other cases, one considers all variables and observables to be stochastic, i.e., having a probability density distribution (Vaníček 1982). Relative weights on the constraints can then be derived from knowledge or assumptions about this distribution, i.e., about the precision of the constraints.

Zadeh's Generalized Constraint Language (Zadeh 2008) provides the formal framework to extend this idea to general cases of "computing with words". Perception-based statements, i.e., observations, can be precisiated by any suitable means (for example, by probability distributions or fuzzy set membership curves), and their impact on a solution for the network variables can be computed.

This methodology of Zadeh bridges the traditional two-valued logic ontologies to the constraint-based view suggested here, where semantic information is *by default* considered to be uncertain. The main difference to geodetic or other geometrically well-defined cases is that conceptual networks have no clear-cut degrees of freedom. Grounding can therefore not easily be determined to be sufficient, but the geodetic ideas that

- grounding spreads through the network;
- arbitrary observational information can be added;
- assumptions on the relative precision of this information serve to weigh its impact;

remain valid in the "computing with words" scenario of semantic engineering.

7 Conclusions

Semantic engineering constrains interpretations of terminologies. It improves mechanisms for information sharing by using semantic web and social web technology to formulate and evaluate constraints on interpretations. It makes only minimal assumptions about difficult philosophical issues (reference, realism vs. nominalism, cognitive processes), in order to allow for pragmatic solutions to semantic problems.

The proposed engineering view of semantics avoids the pitfalls of treating ideas decoupled from language (and thereby treading on thin ice re-

garding testability of its hypotheses) or treating terminology decoupled from its use (and thereby limiting semantics to linguistic relations). Instead, it rests on a notion of concepts that necessarily involves expressions in a language and ties them to reality. It specifies concepts in constraint networks and grounds ontological constraints in observations of reality. Thereby, it admits conceptual theories, but avoids engaging in psychological speculation about what ideas people may have about the world. The matter of study (and of engineering design) is how people apply terms to refer to something in the world that is either commonly observable in an information community or traceable to something that is.

What these observable aspects exactly are is a question discussed elsewhere (Schneider et al. 2009). It constitutes one of the core research questions raised by semantic engineering. Frank has proposed ontological tiers to capture references to reality at multiple levels of abstraction (Frank 2001). Here, I refrain from assuming anything about such levels (e.g. about their ordering or about the role of objects) and only posit that observation sentences (in the sense of Quine) exist, so that primitive symbols can be interpreted through ostension. While there may be philosophical quibbles against this position as a general requirement, it appears to rest on solid ground in the context of geospatial information, which is per definition rooted in observations of the environment.

Ontology research has made limited use of the idea that ontologies are networks of constraints on concepts. Yet, concept networks have been a central idea in dealing with semantics for a long time, both in linguistics (Langacker 1987) and in computing (Woods 1985). Networks of constraints are a standard device in many areas of engineering and computing. The chapter has shown that ontologies seen as constraint networks supply mechanisms for grounding, for accommodating uncertainty, and for integrating folksonomies.

Mechanisms for the formal treatment of ontological constraint networks, including their grounding and uncertainty, remain to be refined and implemented. It appears that model theory is a sufficient formal basis, if observable aspects of reality are admitted as models, as proposed, for example, in (Hayes 1985). These models are then algebraic, consisting of observable qualities and their changes, and transcend the naïve set-based model theory of formal semantics.

Picking out *one* symbol and considering only *one* sense of it turns the relations at the edges of the meaning triangle into functions. This allows for a categorical formalization of the triangle, where the *refer* function is treated as a composition of *observe* and *express* functions. Thereby, semantic theories may get connected to theories of change and action (Kuhn 2005), explaining semantics through observable effects of processes. For example, this would make it possible to explain why the seat neighbor in

the flight over Lake Constance might leave his seat after the above dialogue to stretch his legs before an expected landing in Zurich.

The combination of linguistic, mental, empirical, and social aspects of concepts advocated here allows for constraining how information producers and consumers interpret terms. It permits agreements on such interpretations in the form of ontologies and it can deal with their evolution over time. This pragmatic position has nothing to do with “cultural relativism”. It rests on the basic scientific paradigm of knowledge derived from observations. It is compatible with, but does not require, a stronger form of realist semantics, but avoids some pitfalls of both, realist and cognitive semantics. For example, it has no need to invoke truth independently of meaning, or to decide which entities have correspondents in reality and which not, nor does it have to assume unverifiable cognitive mechanisms. The only cognitive claim is that humans interpret the terms they use, and that this interpretation is ultimately based on ostension. An appropriate philosophical basis for such a view is radical constructivism (Glaserfeld 2002), which treats human conceptualizations and knowledge as constructions, constrained by observations and interactions with other individuals and with the environment.

Smith’s program of *ontology as science* (Smith 2008) is compatible with, but cannot replace an engineering approach; at least not in the context of geospatial information, which exhibits multiple and often conflicting conceptualizations of reality. A single reference ontology (which Smith pursues for biomedical information) is unlikely to emerge any time soon for geospatial domains.

Meanwhile, putting application terminology on a solid basis in the form of foundational ontologies (such as DOLCE, BFO, CIDOC-CRM, SUMO) helps making sensible general distinctions. For example, universals and particulars, endurants and perdurants, or different types of qualities are usefully distinguished, though the distinction ultimately rests on human conceptualizations. DOLCE’s distinctions suggest a notion of primitive qualities, which support grounding in observations. This kind of anchoring of ontologies is likely to support ontology mappings at least as effectively as the abstract scientific notions from a reference ontology, which are almost guaranteed to be interpreted differently in multiple applications.

Acknowledgments

Support for this work was provided in part by the European Commission (IST project SWING, No. FP6-26514). Many discussions in MUSIL (<http://musil.uni-muenster.de>) and with other colleagues (Andrew Frank,

David Mark, Boyan Brodaric and others) as well as comments from two anonymous reviewers have shaped the ideas expressed and helped me articulate them. I am grateful to Lotfi Zadeh for asking how my approach handles uncertainty, to Andrew Frank for asking where objects come from, and to Mike Worboys for asking why we need to talk about anything beyond terms. The chapter gives some preliminary answers.

References

- Bennett B, Mullenby D, Third A (2008) An Ontology for Grounding Vague Geographic Terms. In: Eschenbach C, Grüninger M (eds) *Formal Ontology in Information Systems (FOIS 2008)*, IOS Press, pp 280–293
- Frank AU (2001) Tiers of ontology and consistency constraints in geographical information systems. *International Journal of Geographical Information Science (IJGIS)* 15(7): 667–678
- Gibson JJ (1986) *The Ecological Approach to Visual Perception*, LEA Publishers, Hillsdale, NY
- Glaserfeld Ev (2002) *Radikaler Konstruktivismus: Ideen, Ergebnisse, Probleme*
- Guarino N (1998) Formal Ontology and Information Systems. In: Guarino N (ed) *Formal Ontology in Information Systems (FOIS'98)*. IOS Press, Amsterdam, Trento, Italy, pp 3–15
- Guarino N, Welty C (2002) Evaluating Ontological Decisions with ONTOCLEAN. *Communications of the ACM* 45: 61–65
- Guizzardi G, Halpin T (eds) (2008) Special Issue: Ontological Foundations for Conceptual Modeling. *Applied Ontology* 3(1-2)
- Hayes PJ (1985) The Second Naive Physics Manifesto. In: Moore Ha (ed) *Formal Theories of the Common-sense World*. Ablex, Norwood, NJ, pp 1–36
- ISO (2002) ISO 19111 - Spatial referencing by geographic coordinates, ISO TC 211
- Janowicz K, Keßler C (2008) The Role of Ontology in Improving Gazetteer Interaction. *International Journal of Geographical Information Science (IJGIS)* 22(10): 1129–1157
- Kuhn W (2003) Semantic Reference Systems. *International Journal of Geographic Information Science (IJGIS, Guest Editorial)* 17: 405–409
- Kuhn W (2005) Geospatial Semantics: Why, of What, and How? *Journal on Data Semantics*: 1–24
- Kuhn W, Raubal M (2003) Implementing Semantic Reference Systems. In: Gould MF, Laurini, R, Coulondre S (eds) *6th AGILE Conference on Geographic Information Science*, Presses Polytechniques et Universitaires Romandes, April 24-26, 2003, Lyon, France, pp 63–72
- Langacker RW (1987) *Foundations of Cognitive Grammar*, vol. 1: Theoretical Prerequisites, Stanford University Press, Stanford
- Mark DM (1993) Toward a Theoretical Framework for Geographic Entity Types. In: Frank AU, Campari I (eds) *Spatial Information Theory: Theoretical Basis*

- for GIS, Lecture Notes in Computer Science 716, Springer Heidelberg Berlin New York, pp 270–283
- Ogden CK, Richards IA (1923) *The Meaning Of Meaning*, Harcourt Brace Jovanovich
- Putnam H (1975) *Mind, Language and Reality*, Cambridge University Press, Cambridge, MA
- Quine WVO (1960) *Word and Object*, The MIT Press, Cambridge, MA
- Scheider S, Janowicz K, Kuhn W (2009) *Grounding Geographic Categories in the Meaningful Environment*, MUSIL working papers, Institute for Geoinformatics (ifgi), University of Muenster, Muenster (Germany)
- Smith B (2004) *Beyond Concepts: Ontology as Reality Representation*. In: Varzi A, Vieu L (eds) *Proceedings of FOIS*
- Smith B (2008) *Ontology (Science)*. In: Eschenbach C, Grüninger, M (eds) *Formal Ontology in Information Systems (FOIS 2008)*, IOS Press
- Sowa JF (2000) *Knowledge Representation. Logical, Philosophical, and Computational Foundations*. Brooks Cole, Pacific Grove, CA
- Vaniček P, Krakiwsky EJ (1982) *Geodesy: The Concepts*, North-Holland, Amsterdam
- Woods WA (1985) *What's in a link: Foundations for Semantic Networks*. In: Levesque RJBaHJ (ed) *Readings in Knowledge Representation*. Morgan Kaufman, pp 218–241
- Zadeh LA (2008) *Is there a need for fuzzy logic?* *Information Sciences* 178: 2751–2779

A Common Spatial Model for GIS

Christopher Gold

Computing and Mathematics, University of Glamorgan, Pontypridd, Wales

Abstract

This chapter attempts to describe the role of tessellated models of space within the discipline of Geographic Information Systems (GIS). We look at some of the basic operations in GIS, including dynamic and kinetic applications. We examine issues of topology and data structures, and produced a tessellation model that may be widely applied both to traditional “object” and “field” data types. Based on this framework it can be argued that tessellation models are fundamental to our understanding and processing of geographical space, and provide a coherent framework for understanding the “space” in which we exist.

1 Introduction

This chapter attempts to describe the role of tessellated models of space within the discipline of Geographic Information Systems (GIS) – a specialty coming largely out of Geography and Land Surveying, where there was a strong need to represent information about the land’s surface within a computer system rather than on the original paper maps. We start with a quick look at “Geographic Data”.

The more one looks at these issues, and how they can be represented, the more one realizes that there are many different “Models of Space”, often mental models that may differ between one user and another. These may not be easy to classify, but they form our understanding of what manipulations are feasible for the space we are working in. A simple example

is a terrain model: we may think of it as a space-covering grid or triangulation; a regular lattice of points; a cellular structure around each data point; and several more. We will focus on what appears to be the key issue: the distinction between space-covering “fields” and discrete “objects”.

This leads us directly to a key question – if fields allow us to express spatial relationships, how can we handle “Discrete Objects”? We propose the concept of “Spatial Extension” to handle this issue, and thus introduce the Voronoi diagram as a fundamental idea in GIS.

Our field model thus covers the relevant space, but does not necessarily represent the value of a continuous function. Indeed, it is formed of discrete tiles having their own particular attributes and, as they are connected, some representation of adjacency. This leads to a discussion of “Data Structures, Algorithms and Applications”: the idea of the dual; the main categories of algorithms; and examples of useful applications of these methods. We emphasize the importance of the dual of a planar graph, as well as of tessellations of spatially extended objects – the Voronoi diagram (VD) and its dual, the Delaunay triangulation (DT).

Based on this framework it can be argued that tessellation models are fundamental to our understanding and processing of geographical space, and provide a coherent framework for understanding the “space” in which we exist.

2 Geographic Data

In GIS data is often classified as of either “field” or “object” types. Field data implies spatial continuity whereas objects are unconnected entities, such as houses. Field data again may be classified as having discrete or continuous attributes – for example land use classification vs. temperature. Traditionally tessellations are used for discrete attribute fields, and not for objects. Point observations of fields may be interpolated over the whole space and either be classified and treated as discrete, or else modelled as surfaces – as in triangulated terrain models.

However, recent re-examination of tessellations suggests that they are valuable for object-type data as well – if they are based on an adaptive tessellation such as the VD rather than a spatial partitioning like the quad-tree. Following from this is the idea that cell adjacency relationships describe the relative positions of the generator objects (a local coordinate system). It should be remembered that one of the original, and ongoing, concerns of GIS is the valid construction of connected graphs (polygons, rivers etc.) from approximate co-ordinate data.

Finally, while the Voronoi and Delaunay duality is well known (Fig. 1), there are many applications where both of these structures are needed si-

multaneously – for example in crust and skeleton construction and catchment area modelling (Gold and Dakowicz 2005). Data structures in 2D or 3D that preserve these relationships, and associated attributes, are becoming more interesting.

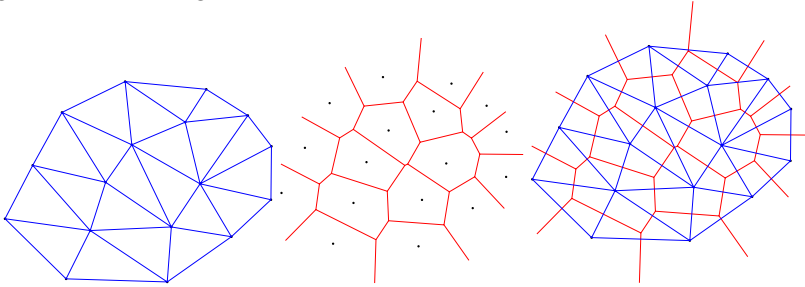


Fig. 1. Delaunay triangulation; Voronoi diagram; their combination

It is traditional in GIS to talk about geometry (coordinate information) and topology (connectivity information) separately when talking about vector data consisting of points, lines and polygons. More complex objects (e.g., building outlines) are composed of these elements, with attributes attached. The same approach may be used for space-covering polygon tessellations, where boundaries are shared by adjacent polygons, etc.

However, these data points (vertices) are useless without some connectivity information – at a minimum an ordering of vertices around the building to “connect the dots”. If we have a connected set of polygons, such as administrative districts forming a tessellation, we must not only connect the dots along the boundaries, but also connect the boundaries around each polygon. There is a variety of ways to construct this “topology”, and this is still a subject of research interest. Thus we have geometry “objects” or “entities” as well as topological ones.

While topological entities do not usually need additional information beyond connectivity, the remaining entities are attempts to represent aspects of the real world. Each entity type may have specific properties, such as semantics, function, time and scale, in addition to various attributes such as colour, population, area etc.

3 Models of Space

Perhaps the simplest model of space “in the head” is a set of points on the plane, with no specified relationships between them. Nevertheless, some sorts of relationships are often imagined by the viewer, based on relative proximity. Many relationships that are “obvious” to the viewer, such as

clusters, are not at all obvious to a computer, and this causes a good deal of confusion. As noted by Andreas Thomsen: “*Topology is intrinsically related to space – it exists whether you like it or not. It can be either implicit or explicit in the data model.*” (Gold and Anton 2007) The same holds true of other disconnected objects in the map – for example, building polygons or linear features.

Alternatively we may imagine some continuous model of space – a field. One example is a terrain model: we may think of it as a grid made of square cells; a grid formed from a regular lattice of (either unconnected or connected) points; a triangulation connecting arbitrary data locations to give an adjacency network; a space-filling triangulation formed of flat plates; or a cellular structure representing the proximal region for each data point. In addition, we may want to imagine movement between cells: possibly water runoff. Here we need to have both a container for the current water in some cell and some adjacency statement to permit movement. If we think of a grid of square cells, these form the “buckets”, and their four neighbours give the adjacency. However, if we need to estimate slopes, in order to simulate flow rates, we immediately switch to the dual, with a node at the centre of each cell, connected to its neighbours, each with a slope based on the elevation difference of the nodes. We are already working with two spatial models at once. The same is true for irregular distributions: proximal (Voronoi) cells form the buckets in our “Block World” and the dual triangulation of their generating nodes gives the slopes (and perhaps the planar TIN model) in our “Slope World”. Similarly, any polygon map has a dual based on the adjacency given by shared boundaries. Duality is hidden everywhere – even in the apparently distinct problem of identifying road centrelines or river watersheds: a subset of the dual of the boundary or river simplifies the job greatly.

The oldest argument in GIS was therefore between a “raster” (grid) view of the world - a field model, partitioning space on the basis of coordinates, and a “vector” view – consisting initially of disconnected 1D line segments, each associated with some cartographic object. Transfer between one and the other inevitably produced loss of precision. We will focus next on what appears to be the biggest issue: the distinction between “fields” and “objects”.

3.1 Objects and Fields

GIS has traditionally separated the mapping of objects, and of fields, and has used separate data structures for each. Perhaps the simplest distinction is based on what may be the most fundamental GIS query: “What is here?” when pointing at some map location. For fields there is always an answer

(perhaps temperature or population density) within the boundary of the map. For objects the answer may be “nothing” – e.g., for a map of houses the answer will be “House” or “Not House.” Thus for fields, with complete map coverage, usually by pixels, polygons or triangles, there is some implicit or explicit definition of adjacency where two cells meet at a common boundary.

This is not present for discrete unconnected objects, creating difficulties when the mapping of moving objects is desired. While no interaction between a moving object and its neighbours may be needed for simple applications, any attempt at real-world simulation needs to maintain proximity relationships between them in order to prevent unwanted collisions, or perhaps to permit managed interaction. It should be noted that this applies for dynamic (insertion and deletion) situations as well as kinetic ones involving point movement. Another fundamental GIS operation – interpolation – is of this type: a query point is considered to have been inserted at the desired location; its set of neighbours obtained; and an attribute (e.g., elevation) estimate calculated from this set. In all these dynamic or kinetic cases a set of neighbouring objects is required, based on spatial proximity. For simulation of the real world some time scale is required, but this is not necessary where movement is merely part of the construction process – for example when simulating the drawing of a line.

Many techniques have been developed to obtain this set, often based on metric distance (Gold 1992). However, many inconsistencies have been noted where pure distance has been used (Jones et al. 1995). A more consistent approach is based on a field spatial model – an object set is converted to a field model by calculating a proximal map: each object has an associated polygon representing the portion of the map closer to that object than to any other. This partitions the map into polygons and thus adjacency relationships can be defined based on the common boundaries. Thus the “What is here?” query is transformed to “What is closest to here?” As a query of a map-covering field, an answer is always obtainable. These proximal maps – called Thiessen polygons, Dirichlet cells or Voronoi diagrams – give a spatial model providing a unifying concept between objects and fields as well as adjacency information.

Other work on this topic includes Winter and Frank (2000), who discuss the integration of vector and raster models by constructing the “skeleton” of the raster – by which they mean the edges and nodes forming the boundaries of each pixel. Goodchild et al. (2007) describe geo-fields and geo-objects, and how to convert between them: obtaining geo-fields from geo-objects is essentially an interpolation problem. They describe three “complete” discretizations of geo-fields of points, where the interpolation method is evident from the data type, and three “incomplete” cases where it is not. They neither describe the “proximal” case, nor the case of non-point

objects. By contrast, we use the same proximal model, or its dual, for all point, line segment and compound cases.

3.2 Objects: Spatial Extension

Space is continuous – so are fields, which, in GIS, attempt to represent space by tessellations as we need to break down spatial continuity into discrete elements for computer storage and processing. Then, of course, in order to simulate our original space, we need to sew the tiles back together with an appropriate topological structure. In effect, we first classify our space into tiles, and then classify the attributes in each tile. As any cartographer would admit, classification inevitably distorts the truth – but we have little choice.

The fundamental GIS problem (which is the same as the CAD problem in 3D) is how to produce a connected structure (a field of some type, with some kind of topology) from discrete objects with imprecise coordinate data. It is known that detecting adjacency/connectivity from metric proximity alone is invalid. A valid approach is to generate the VD of our point objects in the embedding space: this gives sufficient information to determine the connectivity of our desired structure. This is perhaps consistent with Einstein's (1961) comment:

“I wished to show that space-time is not necessarily something to which one can ascribe a separate existence, independently of the actual objects of physical reality. Physical objects are not in space, but these objects are spatially extended. In this way the concept of ‘empty space’ loses its meaning.”

Thus if we expand our individual cartographic objects (points, houses etc.) on a regular basis until their “bubbles” meet we have produced a “field” model from our independent objects, with its associated topological structure. In Euclidean space, with point generators and a fixed expansion rate, this produces the ordinary VD. If we connect each pair of generators whose cells touch we get the DT. This forms the “dual” graph of the Voronoi representation, and vice versa. Similarly, primal and dual structures may be preserved for traditional polygon (choropleth) maps.

4 Data Structures, Algorithms and Applications

The connectivity model or “topology” for tessellations consists of nodes and edges – it is a graph. There are many computer algorithms that perform various types of searches on graphs, and they are basic to very many applications. Graphs may be un-weighted or weighted (with values on

edges or nodes), they may be undirected or directed (allowing traversal in only one direction) and they may be unconnected or connected. They may also be non-planar or planar, and in planar graphs any minimally closed paths form regions that may be entities in their own right, for example as map polygons.

For a planar map such as this a dual graph exists, where the region entity is treated as a node, and nodes in adjacent regions are connected. The node in the original graph is now surrounded by a loop of edges, becoming a region in the dual graph. Except where more than three regions meet at a node the result is a triangulation. If we spatially extend a set of points, as described previously, we generate a set of regions, the simple VD, which has a dual, the DT. Because the VD is unique the DT is unique – which is not the case for other triangulations. Data structures exist to represent these planar graphs – a notable example is the Quad-Edge (Guibas and Stolfi 1985), as it represents both the primal and the dual at the same time.

The difficulty of finding the order of edges around a node in traditional GIS is because the 1D edges have relationships that are not fully described in 2D space. The ideal solution is to spatially extend these edges so they generate a space-filling tiling where all topological relations are fully described.

4.1 Classes of Algorithms

We may examine our mapping (and tessellations) from a temporal point of view. The simplest case is a static map - either where nothing changes, where only the viewpoint changes (pan and zoom in 2D, or fly-throughs in terrain or city models). Alternatively, the map structure remains unchanged, but attributes vary – as in the previous surface runoff simulation, or slide-shows illustrating population change. Algorithms may be batch processes, where intermediate results are not available until construction is finished, or incremental ones where objects are added one at a time. (Batch processes are often faster, but incremental ones are often easier and more robust.) A useful example is DT construction: Fortune's (1987) algorithm is faster, but the incremental algorithm (Guibas and Stolfi 1985) with one point inserted at a time is frequently used, as incremental processes consist of smaller, simpler steps that are easier to explain and implement. This is particularly useful for large geographic data sets, as any system failures would require a batch process to start again, where an incremental method may be restarted following the last successful modification. It should be noted that most robustness problems are caused by topological errors due to finite-precision computer arithmetic.

Algorithms are called dynamic not, as is frequently supposed, because they model dynamic actions in the real world, but because the related data structures may be updated locally, without the need to rebuild everything for each change. In many cases this simply requires object or point deletion methods, as well as the previous incremental insertion ones. This greatly simplifies system design. Kinetic algorithms allow the related data structures to be preserved during object movement, but are more difficult to implement. We will summarize algorithms and applications based on whether they are static (incremental), dynamic or kinetic.

4.2 Incremental Algorithms and Applications

Constructing a continuous terrain model from scattered elevation data used to be done by estimating values on a grid by averaging the values of nearby points: a form of interpolation. Today a TIN model (Triangulated Irregular Network) is more common: for simple cases assuming flat triangular plates is sufficient. The DT is used as local data changes are guaranteed to make only local changes to the network.

Where total volume estimates of rock (e.g., coal) are needed, the most stable approach is not to attempt to construct upper and lower bounding surfaces, but to assign the estimated thickness at any location to that of the nearest data point. This produces Voronoi prisms.

Integrated Finite Difference (IFD) flow modelling simulates flow between adjacent irregular cells that have boundaries perpendicular to the dual edges connecting adjacent elevation values: this is a property of the VD. (Dakowicz and Gold 2007).

Where it is necessary to construct polygon maps rapidly, e.g., for annual forest mapping, an effective method is to “roll” the digitizing cursor round the interior of each polygon, generating multiple points, each with the polygon label. Generating the VD and extracting the boundary between cells with differing labels provides a simple topologically-connected polygon map, a “labelled skeleton” (Gold et al. 1996).

4.2.1 Crust and Skeleton

The labelled skeleton is only effective for closed polygons, and requires digitizing both sides of each boundary. For open networks or scanned polygon boundaries with sufficient point density the “geometric skeleton”, as well as the “crust” of connected boundary points may be extracted from the simple DT/VD (Gold 1999; Gold and Snoeyink 2001). Fig. 2a shows the DT/VD and Fig. 2b shows the crust and skeleton obtained by accepting only one part of each Voronoi/Delaunay edge pair. The skeleton is “hairy”

due to irregularities in the scanned boundary. Smoothed results may be obtained by iterative adjustment of boundary points so that they fall on adjacent circumcircles (Thibault and Gold 2000). If polygons were formed by the boundaries of text characters the generated skeletons form a simplified representation of the character, which can assist in character recognition.

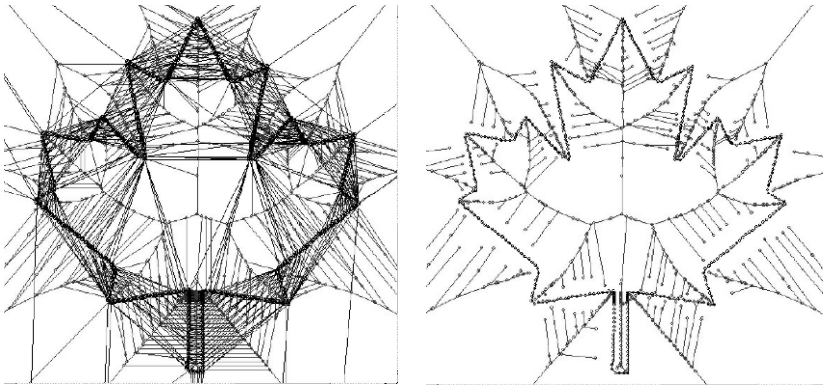


Fig. 2. (a) Delaunay plus Voronoi (b) Crust plus skeleton

Fig. 3 shows a portion of a scanned cadastral map, where image filtering has identified the boundaries between light (background) and dark (text). Generating the skeleton gives a connected topological set of property boundaries, the position of buildings within properties, characters forming labels, the grouping of characters to form labels, and the identification of labels with properties (Gold 1997).

Scanned contour maps may be used to generate TIN type terrain models, but along ridges and valleys some triangles will be “flat”, with the same elevation at all vertices. Using the ridge and valley portions of the skeleton generates intermediate secondary data points that break up the flat triangles, and their elevations may be estimated by using certain assumptions of constant slope (Thibault and Gold 2000; Dakowicz and Gold 2002).

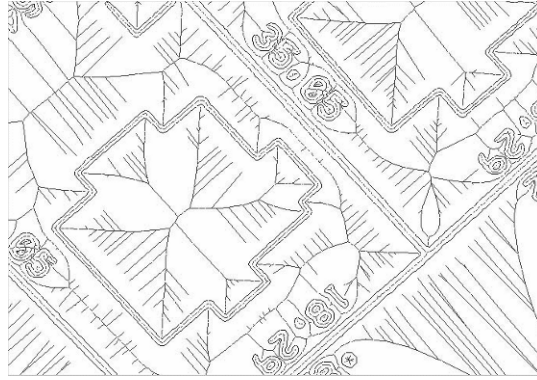


Fig. 3. Scanned cadastral map: crust and skeleton

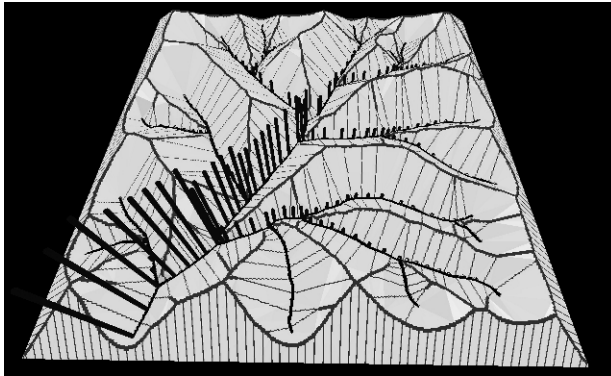


Fig. 4. Terrain from crust and skeleton of the drainage network, together with cumulative catchment areas from the Voronoi cells.

Fig. 4 shows the skeleton of a digitized river system, together with the VD of the river points. Blum (1967) originally defined the skeleton, or “Medial Axis Transform”, and also suggested that the height of a skeleton point could be estimated as equivalent to the radius of the Voronoi vertex forming the skeleton point. Fig. 4 shows the resulting terrain model. Assuming a fixed rainfall throughout the map, the volume of water in each Voronoi cell is proportional to the cell size, and the cumulative sum of these volumes downstream gives an estimate of the total water flow at any river node. This cumulative catchment model provides a first-order approximation of flow, based on the drainage network alone.

4.3 Dynamic Algorithms and Applications

Algorithms for maintaining GIS data structures are usually based on a line-intersection spatial model, rather than a tessellation. The tessellation model that is most commonly used in GIS is the constrained DT (Jones et al. 1995), which is usually static, in that all vertices are added first to give a simple DT (Chew 1989), and then constrained edges are added to give enforced boundaries. Deletion of these edges is not addressed: even point deletion in the simple DT is only recently described (Devillers 1999; Mostafavi et al. 2003).

Second Order Voronoi Diagrams

As the VD of a set of points gives the regions closest to each point, this may be used to find the closest service (e.g., hospital) to any map location. However, if this service is out of order, perhaps in a disaster, the closest service needs to be recalculated. This can be done by deleting the appropriate point in the VD. (More formally, the ordered order-2 VD gives the regions with the closest and second-closest services – see Okabe et al. 2000.)

The same approach may be used as an interpolation method. Traditionally local interpolation (based on a small set of neighbouring data points) used some form of inverse distance to weight the contribution of the elevation of each data point. The difficulties with this approach are described in Gold (1989). Sibson interpolation (Sibson 1981) removed many of these problems by inserting a dummy point at the location of interest, then removing it, and calculating the areas of the adjacent Voronoi cells “stolen” by the point insertion. See (Gold 1989). This has proved to be a very convenient approach for many applications: for example where data points are anisotropically distributed around the query point.

4.4 Kinetic Algorithms and Applications

Many dynamic applications may be performed with these insertion and deletion operations. Examples are: Sibson interpolation (Sibson 1981), map editing; simulation of robot movement by removing the point from one location and placing it in the next, etc. The data structures themselves are also dynamic, being capable of local update. However, for further applications true kinetic properties are required, where it is necessary to know when the topology will change (given the current trajectory of the point), to move there, to update the topology, and to continue. This has been achieved in 2D (Gold 1990; Roos 1990; Guibas et al. 1991; Mostafavi and

Gold 2004): when change (movement etc.) destroys these kinetic properties then local updating is required.

Applications of the kinetic VD include ship navigation (Gold et al., 2004), free-Lagrange fluid flow simulation (Mostafavi and Gold 2004) and interactive constrained DT and line-segment VDs (Gold 1990, 1994; Gold and Dakowicz, 2006).

With the kinetic VDs a new type of GIS system for maritime navigation safety has been developed. The system takes advantage of the properties of the kinetic VD (which implies that the first possible collision must be with one of the moving point's Voronoi neighbours) and uses the Quad-Edge data structure for the real-time maintenance of the spatial relationships of ships and other navigational objects, which are used for collision detection and avoidance (Goralski and Gold 2007).

Mostafavi and Gold (2004) used the kinetic VD to simulate global tides. Shorelines were marked by double lines of fixed points, and a regular pattern of Voronoi cells were placed throughout the oceans, each representing a fixed volume of water. The free-Lagrange method (Fritts et al. 1985) was used to handle the forces on each cell – both from the neighbouring cells and from the moon.

Managing the Voronoi structure during point movement has obvious applications in the simulation of real-world processes. However, the same approach may be used for the simulation of the map drawing process itself: the moving point simulates the pen, which leaves a trace of its path behind.

The most obvious example of this is the construction of the constrained DT, where certain triangle edges are forced to conform to some cartographic feature, such as a building boundary. The trace of the pen's path is formed by forbidding the triangle edge connecting the starting point and the pen to be switched, no matter how far the pen moves (see Gold and Dakowicz 2006). Constrained edges are deleted by reversing the movement of the pen.

Line-Segment VD

Since line segments are required to define cartographic features, and not just points, the kinetic model allows the VD/DT to be maintained as the pen point moves through the tessellation, trailing a line segment behind that is itself a Voronoi generator, giving the line-segment VD. While the dynamic model could be developed with only the CCW and InCircle geometric predicates used in Computational Geometry (Guibas and Stolfi 1985; Shewchuk 1997), giving a guaranteed robustness, the kinetic model requires additional tests to handle near-degenerate cases, especially point collision.

In either the constrained DT or the line-segment VD, all the update operations used have their inverses, as point movement may expand or contract the trailing line (Mioc et al. 1999). Preserving the topological relationships during construction means that potential collisions may be detected in advance, and the appropriate join operations implemented. This is simplified as the lines and their proximal regions are embedded in two-dimensional space.

The difficulties with this approach have been that algorithms for the construction of the line-segment VD are extremely complex (Sugihara et al., 2000) and sensitive to the limitations of computer floating-point arithmetic (Shewchuk 1997). Held (2001) stated that it took him ten years to achieve this for a static algorithm. Karavelas (2004) described a robust incremental algorithm (without deletion). Gold and Dakowicz (2006) built a kinetic line-segment VD algorithm, where points and line segments may be inserted, intersected and deleted – which is particularly useful for a dynamic GIS. It appears to be robust in practice. Fig. 5 illustrates the spatial relationships that may be obtained from this spatial model. The connectivity of building outline segments and the interior portions of their Voronoi cells is clear, as is the associated set of exterior cells, giving a good definition of proximity to any building, as well as defining building adjacency for navigation, path planning, etc. As a field model of discrete objects it may be combined with other field models as desired.

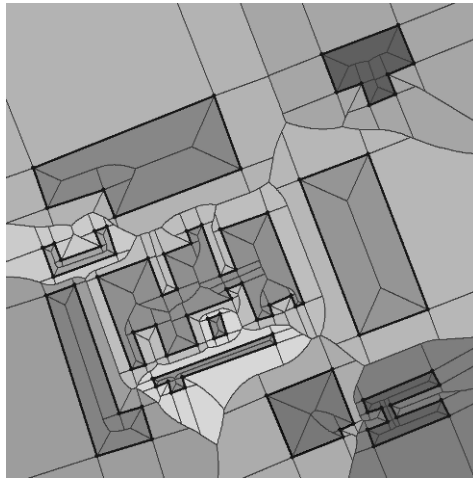


Fig. 5. Neighbour relationships from Line-Segment VD

5 Conclusions

We have looked at some of the basic operations in GIS, including dynamic and kinetic applications. We have examined issues of topology and produced a tessellation model that may be widely applied both to traditional “object” and “field” data types. Based on this framework it can be argued that tessellation models are fundamental to our understanding and processing of geographical space, and provide a coherent framework for understanding the “space” in which we exist.

This paper has restricted itself to 2D applications of tessellations, with a brief look at time in the form of simulation – either of real-world processes or of the map construction process. The same principles apply in 3D – see Ledoux and Gold (2007) for a discussion. Dynamic and kinetic algorithms for points have been developed, and have practical applications in geology and flow modelling, for example, but the inclusion of line segments and faces in the 3D model is still a research topic.

References

- Blum H (1967) A transformation for extracting new descriptors of shape. In: Whaten-Dunn W (ed) *Proceedings of the Symposium on Models for the Perception of Speech and Visual Form*, MIT Press, Cambridge, Mass., pp 362–380
- Chew P (1989) Constrained Delaunay Triangulations, *Algorithmica* 4, pp 97–108
- Dakowicz M, Gold CM (2002) Extracting Meaningful Slopes from Terrain Contours. In: *Proceedings of the Computational Science - ICCS 2002*, Amsterdam, The Netherlands, Lecture Notes in Computer Science, Vol. 2331, Springer, Berlin Heidelberg New York, pp 144–153
- Dakowicz M, Gold CM (2007) Finite Difference Method Runoff Modelling Using Voronoi Cells. In: *Proceedings of the 5th ISPRS Workshop on Dynamic and Multi-dimensional GIS*, Urumchi, China, pp 55–60
- Devillers O (1999) On deletion in Delaunay triangulations. In: *Proceedings of the 15th Annual ACM Symposium on Computational Geometry*, pp 181–188
- Einstein A (1961). *Relativity: The Special and general Theory*, 15th Edition (Lawson RW, translator), New York, Bonanza Crown
- Fortune S (1987) A sweepline algorithm for Voronoi diagrams. *Algorithmica* 2: 153–174
- Fritts MJ, Crowley WP, Trease H (1985) *The Free-Lagrange Method*. Lecture Notes in Physics Vol. 238, Springer, Berlin Heidelberg New York
- Gold CM (1989) Chapter 3 - surface interpolation, spatial adjacency and G.I.S. In: Raper J (ed), *Three Dimensional Applications in Geographic Information Systems*, Taylor and Francis, London, England, pp 21–35

- Gold CM (1990) Spatial data structures - the extension from one to two dimensions. In: Pau LF (ed) *Mapping and Spatial Modelling for Navigation*, Springer, Berlin Heidelberg New York, pp 11–39
- Gold CM (1992) The Meaning of “Neighbour”. In: *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*, Lecture Notes in Computing Science No. 639, Springer, Berlin Heidelberg New York, pp 220–235
- Gold CM (1994) Dynamic data structures: the interactive map. In: *Advanced Geographic Data Modelling - Spatial Data Modelling and Query Languages for 2D and 3D Applications*, Netherlands Geodetic Commission Publications on Geodesy (New Series) 40, pp121–128
- Gold CM (1997) Simple topology generation from scanned maps. In: *Proceedings of Auto-Carto 13*, ACM/ASPRS, 5, pp 337–346
- Gold CM (1999) Crust and anti-crust: A one-step boundary and skeleton extraction algorithm. In: *Proceedings of the ACM Conference on Computational Geometry*, pp 189–196
- Gold CM, Anton F (2007) Minutes, 3D Geo-Information Working Group on Modelling, 3D-Geoinfo-07 Workshop, Delft, <http://www.3d-geoinfo-07.nl>
- Gold CM, Dakowicz M (2005) The Crust and Skeleton – Applications in GIS. In: *Proceedings of the 2nd International Symposium on Voronoi Diagrams in Science and Engineering*, Seoul, Korea, pp 33–42
- Gold CM, Dakowicz M (2006) Kinetic Voronoi - Delaunay drawing tools. In: *Proceedings of the 3rd International Symposium on Voronoi Diagrams in Science and Engineering*, Banff, Canada, pp 76–84
- Gold CM, Snoeyink J (2001) A one-step crust and skeleton extraction algorithm. *Algorithmica* 30: 144–163
- Gold CM, Chau M, Dzieszko M, Goralski R (2004) 3D Geographic Visualization: The Marine GIS. In: Fisher PF (ed) *Developments in Spatial Data Handling - 11th International Symposium on Spatial Data Handling*, Springer, Berlin Heidelberg New York, pp 17–28
- Gold CM, Nantel J, Yang W (1996) Outside-in: an alternative approach to forest map digitizing. *International Journal of Geographical Information Systems (IJGIS)* 10: 291–310
- Goodchild MF, Yuan M, Cova TJ (2007) Towards a general theory of geographic representation in GIS. *International Journal of Geographical Information Science (IJGIS)* 21: 239–260
- Goralski IR, Gold CM (2007) Maintaining the Spatial Relationships of Marine Vessels Using the Kinetic Voronoi Diagram. In: *Proceedings of the ISVD 2007*, Glamorgan, UK, pp 84–90
- Guibas L, Stolfi J (1985) Primitives for the manipulation of general subdivisions and the computation of Voronoi diagrams. *Transactions on Graphics* 4: 74–123
- Guibas L, Mitchell JSB, Roos T (1991) Voronoi Diagrams of moving points in the plane. In: *Proceedings of the 17th International Workshop on Graph Theoretic Concepts in Computer Science*, Fischbachau, Germany, Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 70, pp 113–125

- Held M (2001) VRONI: an engineering approach to the reliable and efficient computation of Voronoi Diagrams of points and line segments. *Computational Geometry, Theory and Application* 18: 95–123
- Jones CB, Bundy GL, Ware JM (1995) Map generalization with a triangulated data structure. *Cartography and Geographic Information Systems* 22: 317–331
- Karavelas MI (2004) A robust and efficient implementation for the segment Voronoi Diagram. In: *International Symposium on Voronoi Diagrams in Science and Engineering 2004*, pp 51–62
- Ledoux H, Gold CM (2007) Simultaneous storage of primal and dual three dimensional subdivisions. *Computers, Environment and Urban Systems* 31: 393–408
- Mioc D, Anton F, Gold CM, Moulin B (1999) “Time travel” Visualization in a Dynamic Voronoi Data Structure. *Cartography and Geographic Information Science* 26: 99–108
- Mostafavi M, Gold CM (2004) A Global Spatial Data Structure for Marine Simulation. *International Journal of Geographical Information Science (IJGIS)* 18: 211–227
- Mostafavi M, Gold CM, Dakowicz M (2003) Dynamic Voronoi/Delaunay Methods and Applications. *Computers and Geosciences* 29: 523–530
- Okabe A, Boots B, Sugihara K, Chiu SN (2000) *Spatial Tessellations - Concepts and Applications of Voronoi Diagrams* (2nd edition), John Wiley and Sons, Chichester
- Roos T (1990) Voronoi diagrams over dynamic scenes. In: *Proceedings of the 2nd Canadian Conference on Computational Geometry*, pp 209–213
- Shewchuk JR (1997) Adaptive Precision Floating-Point Arithmetic and Fast Robust Geometric Predicates. *Discrete and Computational Geometry* 18: 305–363
- Sibson R (1981) A brief description of natural neighbour interpolation. In: Barnett V (ed) *Interpreting Multivariate Data*, Wiley, New York, USA, pp 21–36
- Sugihara K, Iri M, Inagaki H, Imai T (2000) Topology-oriented implementation – an approach to robust geometric algorithms. *Algorithmica* 27: 5–20
- Thibault D, Gold CM (2000) Terrain reconstruction from contours by skeleton construction. *GeoInformatica* 4: 349–373
- Winter S, Frank AU (2000) Topology in Raster and Vector Representation. *GeoInformatica* 4: 35–65

Computation with Imprecise Probabilities

Lotfi A. Zadeh^{1*}

Department of EECS, University of California, Berkeley, CA 94720-1776

Plenary lecture, International Conference on Information Processing and Management of Uncertainty for Knowledge-Based Systems (IPMU), Malaga, Spain, June 22-27, 2008.

Extended Abstract

An imprecise probability distribution is an instance of second-order uncertainty, that is, uncertainty about uncertainty, or uncertainty² for short. Another instance is an imprecise possibility distribution. Computation with imprecise probabilities is not an academic exercise – it is a bridge to reality. In the real world, imprecise probabilities are the norm rather than exception. In large measure, real-world probabilities are perceptions of likelihood. Perceptions are intrinsically imprecise, reflecting the bounded ability of human sensory organs, and ultimately the brain, to resolve detail and store information. Imprecision of perceptions is passed on to perceived probabilities. This is why real-world probabilities are, for the most part, imprecise.

¹ Dedicated to Peter Walley.

* Research supported in part by ONR N00014-02-1-0294, BT Grant CT1080028046, Omron Grant, Tekes Grant, Chevron Texaco Grant and the BISC Program of UC Berkeley.

What is important to note is that in applications of probability theory in such fields as risk assessment, forecasting, planning, assessment of causality and fault diagnosis, it is a common practice to ignore imprecision of probabilities. The problem with this practice is that it leads to results whose validity is in doubt. This underscores the need for approaches in which imprecise probabilities are treated as imprecise probabilities rather than as precise probabilities.

Peter Walley's seminal work "Statistical Reasoning with Imprecise Probabilities," published in 1991, sparked a rapid growth of interest in imprecise probabilities. Today, we see a substantive literature, conferences, workshops and summer schools. An exposition of mainstream approaches to imprecise probabilities may be found in the 2002 special issue of the Journal of Statistical Planning and Inference (JSPI), edited by Jean-Marc Bernard. My paper "A perception-based theory of probabilistic reasoning with imprecise probabilities" (Zadeh 2002), is contained in this issue but is not a part of the mainstream. A mathematically rigorous treatment of elicitation of imprecise probabilities may be found in "A behavioural model for vague probability assessments," by Gert de Cooman (2005).

The approach which is outlined in the following is rooted in my 1975 paper "The concept of a linguistic variable and its application to approximate reasoning" (Zadeh 1975), but in spirit it is close to my 2002 JSPI paper (Zadeh 2002). The approach is a radical departure from the mainstream. Its principal distinguishing features are: (a) imprecise probabilities are dealt with not in isolation, as in the mainstream approaches, but in an environment of imprecision of events, relations and constraints; (b) imprecise probabilities are assumed to be described in a natural language. This assumption is consistent with the fact that a natural language is basically a system for describing perceptions.

The capability to compute with information described in a natural language opens the door to consideration of problems which are not well-posed mathematically. Following are very simple examples of such problems.

1. X is a real-valued random variable. What is known about X is: (a) usually X is much larger than approximately a ; and (b) usually X is much smaller than approximately b , with $a < b$. What is the expected value of X ?
2. X is a real-valued random variable. What is known is that $\text{Prob}(X \text{ is small})$ is low; $\text{Prob}(X \text{ is medium})$ is high; and $\text{Prob}(X \text{ is large})$ is low. What is the expected value of X ?

3. A box contains approximately twenty balls of various sizes. Most are small. There are many more small balls than large balls. What is the probability that a ball drawn at random is neither large nor small?
4. I am checking-in for my flight. I ask the ticket agent: What is the probability that my flight will be delayed. He tells me: Usually most flights leave on time. Rarely most flights are delayed. How should I use this information to assess the probability that my flight may be delayed?

To compute with information described in natural language we employ the formalism of Computing with Words (CW) (Zadeh 1999) or, more generally, NL-Computation (Zadeh 2006). The formalism of Computing with Words, in application to computation with information described in a natural language, involves two basic steps: (a) precisiation of meaning of propositions expressed in natural language; and (b) computation with precisiated propositions. Precisiation of meaning is achieved through the use of generalized-constraint-based semantics, or GCS for short. The concept of a generalized constraint is the centerpiece of GCS. Importantly, generalized constraints, in contrast to standard constraints, have elasticity. What this implies is that in GCS everything is or is allowed to be graduated, that is, be a matter of degree. Furthermore, in GCS everything is or is allowed to be granulated. Granulation involves partitioning of an object into granules, with a granule being a clump of elements drawn together by indistinguishability, equivalence, similarity, proximity or functionality.

A generalized constraint is an expression of the form $X \text{ isr } R$, where X is the constrained variable, R is the constraining relation and r is an indexical variable which defines the modality of the constraint, that is, its semantics. The principal modalities are: possibilistic ($r = \text{blank}$), probabilistic ($r = p$), veristic ($r = v$), usuality ($r = u$) and group ($r = g$). The primary constraints are possibilistic, probabilistic and veristic. The standard constraints are bivalent possibilistic, probabilistic and bivalent veristic. In large measure, scientific theories are based on standard constraints.

Generalized constraints may be combined, projected, qualified, propagated and counterpropagated. The set of all generalized constraints, together with the rules which govern generation of generalized constraints from other generalized constraints, constitute the Generalized Constraint Language (GCL). Actually, GCL is more than a language—it is a language system. A language has descriptive capability. A language system has descriptive capability as well as deductive capability. GCL has both capabilities.

The concept of a generalized constraint plays a key role in GCS. Specifically, it serves two major functions. First, as a means of representing the meaning of a proposition, p , as a generalized constraint; and second,

through representation of p as a generalized constraint it serves as a means of dealing with p as an object of computation. It should be noted that representing the meaning of p as a generalized constraint is equivalent to precisiation of p through translation into GCL. In this sense, GCL plays the role of a meaning precisiation language. More importantly, GCL provides a basis for computation with information described in a natural language. This is the province of CW or, more generally, NL-Computation.

A concept which plays an important role in computation with information described in a natural language is that of a granular value. Specifically, let X be a variable taking values in a space U . A granular value of X , $*u$, is defined by a proposition, p , or more generally by a system of propositions drawn from a natural language. Assume that the meaning of p is precisiated by representing it as a generalized constraint, $GC(p)$. $GC(p)$ may be viewed as a definition of the granular value, $*u$. For example, granular values of probability may be defined as approximately 0.1, ..., approximately 0.9, approximately 1. A granular variable is a variable which takes granular values. For example, young, middle-aged and old are granular values of the granular variable Age. The probability distribution in Example 2 is an instance of a granular probability distribution. In effect, computation with imprecise probability distributions may be viewed as an instance of computation with granular probability distributions.

In the CW-based approach to computation with imprecise probabilities, computation with imprecise probabilities reduces to computation with generalized constraints. What is used for this purpose is the machinery of GCL. More specifically, computation is carried out through the use of rules which govern propagation and counterpropagation of generalized constraints. The principal rule is the extension principle (Zadeh 1965, 1975). In its general form, the extension principle is a computational schema which relates to the following problem. Assume that Y is a given function of X , $Y = g(X)$. Let $*g$ and $*X$ be granular values of g and X , respectively. Compute $*g(*X)$.

In most computations involving imprecise probabilities what is sufficient is a special form of the extension principle which relates to possibilistic constraints. More specifically, assume that f is a given function and $f(X)$ is constrained by a possibility distribution, A . Assume that g is a given function, $g(X)$. The problem is to compute the possibility distribution of $g(X)$ given the possibility distribution of $f(X)$. In this case, the extension principle reduces the solution of the problem in question to the solution of a variational problem (Zadeh 2006).

In summary, the CW-based approach to computation with imprecise probabilities opens the door to computation with probabilities, events, relations and constraints which are described in a natural language. Progression from computation with precise probabilities, precise events, precise

relations and precise constraints to computation with imprecise probabilities, imprecise events, imprecise relations and imprecise constraints is an important step forward – a step which has the potential for a significant enhancement of the role of natural languages in human-centric fields such as economics, decision analysis, operations research, law and medicine, among others.

References

- de Cooman G (2005) A behavioural model for vague probability assessments. *Fuzzy Sets and Systems* 154(3): 305–358
- Zadeh LA (1965) Fuzzy Sets. *Information and Control* 8: 338–353
- Zadeh LA (1975) The Concept of a Linguistic Variable and its Applications to Approximate Reasoning—I. *Information Sciences* 8: 199–249
- Zadeh LA (1999) From Computing with Numbers to Computing with Words—From Manipulation of Measurements to Manipulation of Perceptions. *IEEE Transactions on Circuits and Systems—I: Fundamental Theory and Applications* 45(1): 105–119
- Zadeh LA (2002) Toward a perception-based theory of probabilistic reasoning with imprecise probabilities. *Journal of Statistical Planning and Inference* 105: 233–264
- Zadeh LA (2006) Generalized theory of uncertainty (GTU) —principal concepts and ideas. *Computational Statistics & Data Analysis* 51(1): 15–46

Spatial Data Quality: Problems and Prospects

Gary J. Hunter¹, Arnold K. Bregt², Gerard B.M. Heuvelink², Sytze De Bruin² and Kirsi Virrantaus³

¹ Department of Geomatics, University of Melbourne, Parkville VIC 3010, Australia

² Center for Geo-Information, Wageningen UR, PO Box 47, 6700 AA, Wageningen, Netherlands

³ Department of Surveying, Helsinki University of Technology, Otakaari 1, Espoo, Finland

garyhnr@gmail.com, arnold.bregt@wur.nl, gerard.heuvelink@wur.nl, sytze.debruin@wur.nl, kirsi.virrantaus@tkk.fi

Abstract

This paper reflects upon the topic of spatial data quality and the progress made in this field over the past 20-30 years. While international standards have been established, theoretical models of error developed, new visualization techniques introduced, and metadata now routinely documented for spatial datasets, difficulties nevertheless exist with the way data quality information is being described, communicated and applied in practice by users. These problems are identified and the paper suggests how the spatial information community might move forward to overcome these obstacles.

1 Introduction

With the growth of Geographical Information Systems (GIS) and new technologies such as the Internet, the broader community is benefiting from access to datasets that were once used only by a small group of spe-

cialists. As such, there are now many more people making decisions based on information they perhaps know very little about—particularly in terms of its quality. In addition, we live at a time when there is less tolerance for poor decision-making and the consequences of ‘getting it wrong’ can be severe for individuals and organizations alike.

Of course, the accuracy issue was always in the minds of the map-makers who, as recently as 30 years ago, had sole responsibility for preparing our paper-based maps and charts. They met the accuracy-reporting challenge as best they could by providing estimates through reliability diagrams, special symbols, and accuracy statements based on testing to accepted industry standards. They knew their products were not perfect and users were also generally aware of this fact, so there was a degree of shared knowledge between the data collectors and users that has since disappeared with the advent of digital data.

With the introduction of digital mapping techniques in the 1960s and then GIS shortly afterwards, researchers realized that error and uncertainty in digital spatial data had the potential to cause problems that had not been experienced with paper maps (for example, see MacDougall 1975; Goodchild 1978; Blakemore 1984; Chrisman 1984; Robinson and Frank 1985; Burrough 1986; Bedard 1987; Epstein and Roitman 1987). On the other hand, the wider GIS community took far longer to realize the potential traps that existed for unwary users, and there is no doubt that the notion of ‘the computer must be correct’ held force for many years.

In conjunction with these warnings, an international trend started in the early-1980s to design and implement data transfer standards which would include data quality information that had disappeared from the margins of paper maps with the transformation to digital data products (Moellering 1991). The standard that clearly led the way in documenting data quality was the U.S. Spatial Data Transfer Standard (NCDCDS 1986; NIST 1992) which divided quality reporting into five parts, viz.: dataset lineage; positional accuracy; attribute accuracy; logical consistency and completeness (Chrisman 1991; Guptill and Morrison 1995). By and large, these elements have stood the test of time, although there have been recent additions and/or variants such as ‘semantic accuracy’ (as a broader alternative to attribute or thematic accuracy), ‘temporal accuracy’ (the accuracy of reporting time associated with the data), and ‘metaquality’ (data about the quality data, such as its reliability and confidence) (CEN 1998; ISO 2002, 2003a, 2003b).

By the late-1980s and early-90s, special attention was being given at international conferences and in the scientific literature to the importance and need for the proper treatment of spatial data quality, and the benefits that would come from its use. Moreover, leading international research centers in the US and Europe had identified key initiatives in spatial data

accuracy, the treatment of indefinite boundaries, and visualization of spatial data quality as being of fundamental scientific importance. There was even an international uncertainty visualization ‘challenge’ conducted in the U.S. Buoyed by this activity and the widespread publicity surrounding the topic, we believe spatial data users at the time expected they would soon be able to (1) easily interpret data producer’s quality statements, (2) understand the inherent strengths and limitations of a dataset in quantitative terms, and (3) translate that information to a form suitable for assessing whether or not they should use it for their decision tasks.

While this might have seemed a utopian view, it was in fact the complete expression of the solution to the spatial data quality research problem, and it was discussed to differing degrees by research leaders such as Openshaw (1989), Burrough (1991) and Goodchild (1992). Of course there were many assumptions underlying this perfect vision of the future. For instance, it was expected that: (1) data quality statements would have appropriate content and format; (2) data consumers would possess the requisite knowledge and skill to comprehend and translate these statements; (3) commercial product developers would write the software to analyze, portray and keep track of error; (4) innovative researchers would produce new error theories, models and methods; and (5) spatial data quality would be able to be expressed in terms of its quantifiable impacts upon intended decisions in a manner that would be obvious to all concerned.

Clearly, the achievement of all these tasks was always going to be difficult, and so just like other critical reviews that have recently been conducted into topics such as space-time data representation (Peuquet 2001), computational methods for representing geographical concepts (Egenhofer et al. 1999), and the integration of GIS and spatial analysis (Getis 2000; Goodchild 2000), in this paper we reflect upon the progress made in spatial data quality over the past 20-30 years. Certainly, standards have been implemented, many datasets now carry quality statements, and researchers continue to investigate models of error and its portrayal, however we suggest that several of the original goals are still to be met. Accordingly, in this paper we revisit the topic of spatial data quality to identify the problems that remain (Section 2) and the work that needs to be undertaken to bring the original vision to completion (Section 3).

2 Problems

In reflecting upon our present level of understanding of spatial data quality, we believe the problems still being experienced in this subject lie in five fundamental areas, viz.: data quality reporting, description and visualization;

error propagation; and the application of data quality information in practical decision-making environments.

2.1 Poor Quality Reporting

Starting with the issue of data quality information content, we believe the current lack of detail provided in many data quality statements makes them ineffective for any subsequent use. To demonstrate this point, the examples of poor data quality reporting presented in Table 1 have been selected from actual data quality statements recently collected.

Taking the positional accuracy examples first, obvious questions soon arise with these statements such as, “Exactly how variable are the observations?”, “Where are the 1000m errors located?”, “Where does the urban/rural quality transition occur in the dataset?” and “Where were the deliberate cartographic offsets made?” In other cases there may be little or no actual basis for making these statements—for instance when the positional accuracy of a very small sample of well-defined point features is tested, and the results are then inappropriately assigned as an accuracy indicator to all objects, regardless of their type (such as linear and areal features).

As for attribute accuracy, to state that this is not relevant for a vegetation map is clearly unacceptable, while claims of ‘high’ accuracy and ‘100%’ accuracy that carry no indication of what was tested, how it was tested or the sample size used, do little to inspire trust in a prospective data consumer. There are also other deficiencies with attribute accuracy reporting that are not listed in Table 1 and which need correction. Firstly, the accuracy of all attributes should be reported separately, since it is not possible to assign a single accuracy value to describe multiple attributes in a database (and indeed, if it were possible it would be an outstanding breakthrough in the data quality research agenda). Secondly, the scale of measurement for each attribute (for example, nominal, ordinal, interval and ratio) should be included in the data quality report as an aid to its later use in conjunction with error modeling and visualization tools.

Moving to the logical consistency examples in Table 1, data with different lineage should be tagged with appropriate identifiers if there are different accuracies present—so that users might learn which features can be expected to possess poorer logical consistency. In addition, there seems to be a common misconception that logical consistency consists only of ensuring polygon boundaries are closed and that linear features meet where intended, however in practice there are many different tests for logical consistency that need to be undertaken with spatial datasets. Reporting of completeness suffers similarly and data quality statements rarely state what information is actually missing. However, stating that some (unidentified)

features are missing or that “street address details are partially complete”, provides little useful information to potential users.

Table 1. Examples of poor data quality reporting.

Positional Accuracy
“Variable”, “100m to 1000m”
“+/- 1.5m (urban) to +/-250m (rural)”
“No feature in error by more than 100m”
“90% of all points lie within 1mm at plot scale”
“Cartographic offsets may be present”
Attribute Accuracy
“Not relevant” (for a vegetation map)
“100% accurate”
“High attribute accuracy”
Logical Consistency
“Between 1% (new data additions) and 5% (pre-maintenance contract data)”
Completeness
“Some features have been eliminated”
“Street address details partially complete”
Currency
“From aerial photography 1965-1992”

Finally there is the reporting of currency (temporal accuracy) and the example given in Table 1 would surely have a potential consumer wondering exactly which parts of the dataset referred to are derived from 40-year old photography and which ones have been updated from more recent material. In addition, currency tends to be described for datasets as a whole and not as it should be for each data quality element where appropriate. For example, the date at which a feature’s position is observed may often be different to the date that its attributes were recorded—and coupled with this is the need to record database transaction dates for maintenance and update purposes.

2.2 Incomplete Quality Descriptions

While the problems associated with poor data quality reporting are relatively minor, there are several other problems that will have greater impact in the future if left unresolved—and they relate to incompleteness in spatial data quality descriptions.

The first of these is that data quality information suffers generally from being presented at a generic global level rather than at more detailed levels of granularity. While global information will always be required in data quality statements, there is also a need to report any spatial variation in data quality. This divergence might reside naturally in the data, or else it

might come about as a result of spatial operations—such as when two datasets with different positional accuracies are overlaid or merged. There is also the need to report any spatial uncertainty or spatial correlation of local data quality. This is important, for instance, if the areas of continuous regions are to be estimated from raster data or when slope gradients, viewsheds or watersheds are computed from DEMs.

Another fundamental problem with data quality descriptions is that they tend to work far better with data representing crisply-defined objects usually found in the built environment, rather than with the more abstract and vaguely-defined data representing phenomena occurring in the natural environment (for instance, see Burrough and Frank 1996). This is hardly surprising since we are the ones who have designed the coordinate systems, built the technology to measure positions, and defined the terms and values used to describe their attributes. However, the natural world is not of our making, and trying to observe and represent its processes are difficult enough to achieve in practice without also having to describe the accuracy with which we define and model it. For example, when we perform soil sampling we must limit our testing to points to minimize soil damage, and then (to make the observations fit our relatively simple computational models) we allocate crisply-defined boundaries to polygons having homogeneous consistency to represent something that is inherently heterogeneous and known to possess widely varying transition zones. Describing this difference, between the models we use to depict the real-world and the real-world itself, is a continuing problem and continued research will clearly be required in this area.

Furthermore, for the estimation of error propagation to be successfully achieved (see section 2.4) we need considerably better information to be provided than we now receive. Taking DEMs as an example, the elevation error in a DEM is typically conveyed by a Root Mean Square Error (RMSE), however that on its own is not always sufficient. For error propagation to be estimated (such as when we derive a slope map or a viewshed from a DEM) we also need to know the spatial autocorrelation in the error. Ideally, we should have the full joint probability distribution but this is not available in practice so we tend to get, at best, a parameterized model of the joint probability distribution. This means that someone else has chosen a particular model, with its inherent assumptions, such as stationary random variables.

Finally, some comments should be made about error modeling, because if we cannot define error then we cannot measure it or describe it. Certainly, ten years ago few theoretical error models existed and Goodchild (1993) noted at the time that the known and accepted techniques we possessed for describing and measuring error were essentially limited to: the locational accuracy of a single point (through the circular normal model of

positional error); the accuracy of a single measured attribute; the probability that a point at a randomly chosen location on a map has been misclassified (through the misclassification matrix); the effects of digitizing error on measures of length and area; the propagation of error in raster-based area class maps through spatial operations such as overlay; and the error associated with measures of area derived from dot counting. Since then, the development of error models has progressed and numerous models have now been proposed in areas as diverse as: positional error in vector data; thematic uncertainty in the integration of Remote Sensing and GIS; the accuracy of TINs and Lattices; elevation error in DEMs; errors in point-in-polygon analysis; fuzzy representation of boundaries; field attribute error; errors in buffering operations; probabilistic viewsheds, and in cartographic generalization processes. However our knowledge of error remains relatively immature, although it is not due to lack of effort.

2.3 Barriers to Communicating Quality

Moving away from how we describe the fundamental spatial data quality elements, there is a range of issues relating to how quality is being communicated to spatial data users. While data producers have generally accepted the need for data quality reporting, consumers of their products do not seem to have embraced the spatial data quality issue to the same extent. This could be due in part to reasons such as: (1) the fact that many users have never received formal education in GIS; (2) that there is no commonly taught approach to the critical analysis of results in geographic information science (unlike in other disciplines such as surveying and geodesy); (3) that clients who commission a data product may not necessarily be interested in its quality; and, perhaps, (4) that users have become disillusioned with the lack of progress in this area. Of course, even if we were able to overcome each of these difficulties, there remains the issue of how to effectively communicate data quality to different types of users. For instance, while an analyst may readily understand the meaning of linear regression statistics, standard deviations and semi-variograms, such concepts can be bewildering for both the novice at one end of the user-spectrum and the senior decision-maker at the other.

Another communication problem, this time associated with spatial database structure and design, is that we do not possess the tools to manipulate, query, analyze or display data quality information—as we already do for spatial data. Similarly, we are unable to update data quality information in real-time as changes occur in a database. For example, while we can easily take two separate point datasets and combine them to form a new dataset through a simple ‘merge’ or ‘append’ operation, if they each have their

own data quality statements we are currently incapable of automatically integrating their respective data quality information to yield a new data quality report for the merged data product. Similarly, we are unable to produce a quality report for a slope or aspect map that might be derived from a DEM—even though the DEM will in all likelihood have its own quality information readily available (albeit in a relatively simple form such as a global RMSE). So while we can easily update spatial features and their attributes, it remains a challenge to researchers to provide an effective means of updating attached data quality information ‘on-the-fly’ when spatial processes are applied and new datasets are created—yet this is something that will obviously have to occur in future GIS.

Effective communication of data quality also means having the visualization tools to help do the job, and while we would appear to already have the techniques necessary to perform the task they have yet to be implemented in commercial GIS packages (although there are numerous examples of their implementation in proof-of-concept form). This is partly due to the fact that data quality information is not normally coupled with the data to which it refers, and so there is no capability for subsequently linking it to error modeling and visualization software. While the software developers are naturally the people best able to implement these visualization techniques, the task still does not seem to have a high priority for their companies. Informal discussions suggest there is still not a strong enough level of demand from the user community for this product functionality to justify the expense of incorporating it into commercial systems. On the other hand, the drive by vendors and third parties over the past five years to provide easy-to-use metadata entry tools has been rapidly achieved in response to demands by data producers (particularly government agencies)—so the industry has certainly demonstrated its capacity and technical skill to act quickly when the need arises.

2.4 Keeping Track of Error

Another key issue impediment in dealing with spatial data quality is that our knowledge is still deficient in the way error propagates in many spatial operations. Although we have approximate methods of error propagation in the area of quantitative modeling with continuous data based on the principle of propagation of variances (Heuvelink et al. 1989), and simulation methods in which the effect of perturbation of the input data is observed and quantified in the outputs, these are methods that become impractical when dealing with chains of complex operations and when dealing with categorical rather than continuous data. Furthermore, the error propagation techniques we do possess are inevitably applied by expert ana-

lysts, with the result that once the ‘average’ GIS user studies the data quality statement for a dataset there is little else that can be done to translate that initial information into quality descriptors for any secondary products they might create. So in essence our progress beyond the current body of knowledge in modeling, reporting and communicating spatial data quality lies frozen at this point.

2.5 Application of Data Quality Information

Finally, users are experiencing problems applying data quality information in real-world, everyday situations—and we should remember that the notion of quality is concerned with ‘fitness-for-use’ or suitability for a task, not just the degree of error in the source data. At the present time data quality reporting could be said to generally be characterized as governed by producer-driven standards and requirements rather than user applications. Of course, from a producer’s perspective this is reasonable since there is no way of controlling how users might apply their data. So their products are tailored to meet certain specifications to satisfy particular past user-demands, but these do not necessarily help other consumers assess whether an information product is actually suitable for a given function. On the other hand, we believe that users would like to be provided with the technical capability to take the initial data quality information and use it to determine what output accuracy will result from the use of a given set of inputs, models and spatial operations—preferably before the task is actually undertaken so that alternative data, algorithms and models can be investigated if required.

At present this has only been performed in a limited way by skilled analysts, and this functionality does not generally exist in commercial software packages. In particular, the problem of verifying model outputs is currently causing concern amongst leading scientists as they discover that governments are increasingly reluctant to commit to highly sensitive policy decisions without any knowledge of the validity of the models being used. For example, Beven (2000, p. 2605) reported that a proposal to establish an underground repository for radioactive waste at Sellafield in the UK was refused permission to proceed after the results from simulated groundwater flow studies “...differed drastically between modelers on both sides of the argument.” At the same time, there have been calls for new research efforts into how we might generally describe the quality of models, and how we might derive a set of model quality elements in a fashion similar to the data quality elements we now possess.

Even if we could propagate error in spatial data and quantitatively assess its effect upon derived outputs, ultimately what users really want to

know is what risk is associated with using information of a given quality—in other words “What can go wrong?” and “Will my decision remain unaltered?” The answer to these questions may well require users to be better trained in decision-making and risk management to foster a fundamental change in the way they perceive their information, and a more probabilistic approach will probably need to be adopted in terms of their interpretation of spatial information.

3 Future Prospects

If these are the fundamental problems associated with spatial data quality that remain after 30 years, then what are the prospects of overcoming them? Certainly, the problems do not rest solely with any particular sector of the spatial information industry and solutions must be found jointly by the data producers, consumers, system developers, educators and researchers. Indeed, there are valuable messages for each of these groups to take from the following discussion.

3.1 Enhanced Quality Reporting

Looking at the data quality reporting problem first, it is clear that poor reporting can be overcome relatively easily with experience and advice, and there are excellent examples of comprehensive data quality reports and technical user guides available on the Internet—such as the Geoscience Australia (2006) and Ordnance Survey (2009) digital data products. The latter provides a good illustration of completeness reporting in its user guide where it lists several pages of real-world features not included in its ‘MasterMap’ dataset (for example, buildings below a minimum size are not shown, telephone lines and poles are only shown when they are of outstanding significance, and roads on private property are only shown when longer than 100m).

This is what Brassel et al. (1995) would refer to as ‘model completeness’ information, as opposed to ‘data completeness’ information which would report whether all existing roads greater than 100m in length have actually been included in the database with their correct attributes. Similarly, for reporting logical consistency the user guide for the Geoscience Australia ‘TOPO 250K’ product details some 60 tests that are performed and reported (for example, label points have only one coordinate pair, road tunnels and bridges are coincident with nodes in the road network, coastline is cloned as a zero height contour, features labeled as an island or reef are completely surrounded by water), together with the test sample size

and the acceptable quality level for each test. So, as stated previously, the prospects for good quality reporting are excellent and will improve with time.

3.2 Improved Quality Descriptions

Moving to the next issue of how quality is described, we have already seen producers provide multi-level data quality information, and an example of this that has existed for over 10 years can be found in the product metadata described in Geoscience Australia (2006). In this instance, data quality information is presented at four levels for the product, viz.: the dataset; data layer, feature class and individual feature levels—with the quality information for the three highest levels being stored in hardcopy narrative form, while the quality information at the feature level is held in attribute form against each object.

As an exercise in describing this multi-level data quality information, Qiu and Hunter (2002) took a sample set of the Geoscience Australia product and its associated data quality information, converted all of the latter to digital form, and then attached it at all four levels within a commercial GIS (ArcView)—so that it became possible to select and display both the spatial data and the data quality information from within the GIS environment. While the trial was successful, in the longer-term they believe a more elegant solution would be achieved by adopting an object-oriented approach in order to make full use of the inheritance, classification and encapsulation capabilities available to more effectively model the data and its quality information. Subsequent research in this area has been conducted by Sadiq and Duckham (2009) who successfully implemented a data quality module in Oracle Spatial software to cater for individual feature- and even sub-feature-level quality variation reporting and querying. In this area, the proposed US ANSI Metadata Standard expected to be introduced in 2009 includes the provision for metadata descriptions at different data levels.

Moving next to the description of data quality particularly when we attempt to represent natural phenomena, some researchers are now suggesting that if we were to ‘step back’ and focus more on describing the quality of the original observations (instead of the models we consequently create from them), then we may well find it easier to provide data quality statements that meet our current standards. For instance, we could more ably define the positional accuracy of soil test sites and better state the variation that occurred in their observed attributes. In essence, we would simply provide source data that had been quality tested, and then let users take responsibility for what happens to it from that point onwards. However, this

assumes they possess the tools necessary to describe the quality of the outputs of subsequent spatial operations on the original data—and at this time they do not. Furthermore, it could be argued that providing only the original test site data may not be sufficient as we could only confirm how well the data used for calibration of the model were represented by the model. In the case of kriging, the fit would be perfect, which could lead a less-skilled user to assume the whole model was perfect. On the other hand it might be more useful to supply users with the model data plus a set of independent data test sites which would allow them to validate the model they propose using.

Alternatively, we could try to search for a more rigorous and exhaustive means of storing uncertainty information about the continuous and categorical forms of spatial data that tend to characterize natural resources. For this to succeed we will need to know not only the (spatially varying) variances but also the distribution type, the spatial autocorrelation, and any cross-correlation with errors in other attributes—and this information will also have to be derived for any new attributes that are created in the database. Taking all of this into consideration is clearly a major task and will impact significantly on the database design, which suggests we might need to settle for a less rigorous approach—but the question then arises “What choices do we make?” Unfortunately, when we come to deal with categorical variables the situation becomes worse—since not only is there the question of “How do we manage and control the parameters needed to fully convey uncertainty?” but more importantly “How can we estimate them?”

So we continue to experience considerable difficulty in describing and measuring error in these types of data, which might explain why many of the examples of (what is taken to be) good data quality reporting tend to relate to digital versions of traditional products such as topographic maps. However, these products are often not the ones used for decision-making, and instead are inclined only to be used along the way to derive secondary information which is what users will ultimately want to know more about in terms of quality. Of course, some of the deficiencies mentioned here are not just confined to data representing natural phenomena, and indeed information on distribution types, spatial autocorrelation and cross-correlation is required for all field data (as opposed to object data) regardless of what they represent.

3.3 More Effective Quality Communication

Dealing next with the problems relating to communicating data quality, clearly the next generation of spatial data consumers will be better educated in issues such as quality. Whereas 10 years ago many GIS courses

tended to focus on GIS technology and its applications, it is now quite common for students to be introduced at an early stage in their studies to the legal and institutional issues of GIS which inevitably connect with matters of quality. So in that respect, the problems associated with the lack of education and awareness in the subject can be expected to be overcome with the passage of time.

In addition, new forms of metadata description are now under development. For example, Boin and Hunter (2007a) have found that to begin with consumers have little or no understanding of the terms 'metadata' or 'data quality'—instead preferring the term 'product description'. Furthermore, their review of many hundreds of data complaint emails coupled with a detailed survey of data consumers, revealed a much greater need for information on what users could expect a dataset to contain. Thus, in terms of understanding how potential datasets are chosen for user applications, their surveys revealed that published metadata played little or no part in the selection process since its content was considered too technical in its nature—even by professionals such as engineers, architects and planners who are using spatial data on a daily basis. Instead they relied upon colleague opinions of which datasets to use for a given purpose or else learnt through experience which datasets could be trusted to meet their needs. The implication of this is that data producers are creating metadata and populating data directories around the world with information that has little or no benefits to data consumers. So in an effort to make the metadata that is collected more meaningful, Boin and Hunter (2007b) reports on the design and testing of a new graphical style of describing the contents of a spatial dataset which consumers found more interesting and informative. Devillers et al. (2007) have also been working in this area of communicating metadata in new ways and their dashboard-style of presentation is similarly finding interest amongst consumers.

A more serious impediment to better communication of data quality, however, is the fact that we are not making life any easier for data users by introducing notions of error and uncertainty given their negative connotations. For instance, urban and regional planners, civil engineers, real estate appraisers and others have all used soil maps for decades to make effective decisions without being aware of the uncertainties about inclusions and map unit variability—a view supported by the work of van Oort and Bregt (2005). Surprisingly, this approach has worked well for many years (for example, see Hudson 1990). So for groups such as these who have learned to live quite comfortably with the usual binary outcomes in GIS processes ('one' or 'zero', 'in' or 'out', 'yes' or 'no', 'black' or 'white'), we are not necessarily seen as doing them any favors by introducing grayness or a 'plus/minus' to their decision-making—even though we know it is something they should be taking into consideration.

What would undoubtedly help most in promoting the importance of data quality to users would be a series of well-documented case studies describing the perils associated with ignoring data quality. While there are already several well-known examples of proven legal liability associated with the provision of erroneous nautical charts and topographic maps, in such cases the relationships between data error and its adverse consequences tends strongly to be both obvious and severe in its impact. For example, if a reef in the middle of a shipping channel is missing from a nautical chart or assigned the wrong depth sounding, then it will only be a matter of time before a ship runs aground on it resulting in expensive claims for compensation being made against the data provider, and (all too often today) major environmental harm.

On the other hand, to the average GIS user the range of possible adverse consequences due to using poor quality data never seem to be quite as dramatic in terms of their impact. Of course for the mistakes that do happen, the reality is that their news tends to be suppressed, arising out of a sense of shame and often also as a condition of any out-of-court compensation payments made—which are preferable to the publicity of a court hearing and the establishment of a legal precedent. Thus, the effect is that any prospective authors who decide to report such cases in the literature do so at considerable risk of initiating legal action against themselves. So it seems unlikely that any “Greatest Failures in GIS” texts will ever appear, although we could certainly develop a ‘best practice’ handbook in spatial data quality which cites positive outcomes—similar in essence to what Marble (2000) called for to encourage greater interest and interaction between the GIS and spatial analysis communities. In addition, it is interesting to note that there is still no textbook dedicated solely to the way we treat spatial data quality, although a valuable and practical handbook on positional accuracy is available on the web (Minnesota Planning 1999).

3.4 Better Error Tracking

One way of increasing data quality awareness would be to communicate the changes that actually occur to quality in real-time as users combine and process datasets using GIS. However the methods that have been developed by researchers to date are invariably time consuming and complex to use. For example, the effort required to run a Monte Carlo uncertainty propagation analysis is at least an order of magnitude greater than running the basic analysis itself (not only in terms of computing time, but also in terms of parameter estimation and management of the process). In addition, these operations tend to have something of a prototype air about them, meaning ‘it only works when I run it’—which implies they are prob-

ably still far too context-specific to be of much value to the broader GIS community. Nevertheless, some of these communication problems have already been recognized by software developers—with the IDRISI product designers clearly taking the lead in the mid-1990s when they introduced a suite of uncertainty management and decision-making tools (Eastman 1997). Other products are also showing increased functionality in this area such as ESRI's ArcGIS software which caters for basic metadata management (which includes data quality) in its ArcCatalog module (ESRI 2000).

Geostatistics provide one means of quantifying certain types of uncertainty in a fairly complete fashion through the use of standard deviations, variograms and cross-variograms. It also offers possibilities for generating realisations of uncertain spatial attributes needed for Monte Carlo uncertainty propagation analysis (and here we could think of sequential Gaussian simulation as the simplest example). However, geostatistics deals primarily with quantitative field data, although there are some extensions to categorical field data using indicator approaches. When it comes to handling uncertainty in object data we lack an equivalent set of tools, although the spatial statistics software, S-Plus, has some functionality in this direction through point pattern analysis techniques. While there has been some introductory work in perturbing the locations of spatial boundaries (for instance, Hunter et al. 1999), it remains to be seen whether these techniques can be easily incorporated or connected to GIS packages.

One solution would be for commercial GIS to have a range of error models available to run on datasets, coupled with error propagation functions that would automatically operate whenever a spatial operation or model was initiated, plus a suite of error communication options, which would all work towards providing input for a set of decision management tools. Of course, such a solution could also exist within a third-party software product, and the move from closed proprietary GIS to open system architectures is an important advancement that will obviously facilitate this. This is starting to occur with links between statistical software and GIS modules (for example, through OLE/COM), and it is expected that a broader group of users will take advantage of tools that were previously restricted to specialists. Unfortunately, third-party products suffer from requiring a separate, deliberate purchase on the part of data users, and therefore may not become as widely adopted as the mainstream GIS package with which they operate in conjunction. Nonetheless, third-party products can eventually become indispensable components of popular software, and the spellcheckers we employ in our word-processing packages are an obvious example.

Clearly, the sensible approach would be for software developers to start small and introduce some simple error models. For instance, it is a comparatively easy task to calculate the standard deviations of polygon areas

from the horizontal positional error estimate in a dataset. Similarly, for grid-cell data an error propagation tool could be developed for the numerical modeling case, while for error communication in vector data a drop-down menu of visualization options could be made available. Such tools would still require users to have a sound knowledge of their application, in much the same way that the well-known Microsoft Excel Charting module offers a wide range of graph types but the responsibility for the outcome ultimately rests with the user. While some of these ideas were promoted over a decade ago by Burrough (1991), to this day they remain such an obvious part of the solution to the spatial data quality problem that they need to be repeated—although clearly the respective forces of demand (from users) and supply (from software developers) have been far too small to bring about their introduction.

3.5 Complete Utilization of Quality Information

Finally, there is the matter of how the data quality application problems described above might be overcome. While easy to express, they will most likely prove difficult to resolve in the short-term given that their solutions depend on how we deal with the more fundamental problems occurring with data quality description and communication. Nonetheless, some researchers suggest that risk management theory might be usefully applied here to assist decision-makers (Agumya and Hunter 1999), and case studies using spatial data are already starting to appear in the literature (De Bruin et al. 2001). Importantly, the risk management approach links into cost-benefit analysis so users can determine whether it is worth improving their data or else taking other provisions to cover their risks (for instance, they might limit the reliance placed upon the outputs of a GIS). Other researchers are combining uncertainty assessment with sensitivity analysis to estimate the comparative quality of different GIS-based models. For example, the excellent work of Crosetto and Tarantola (2001) describes a study in which 15 different types of error residing in seven separate datasets were assessed to judge the reliability of different hydrologic models for flood forecasting.

However, it is possible that these types of studies may be far more detailed than some consumers actually need, and in addition the majority of GIS users are neither experts nor highly-skilled analysts, so we must learn to cater for their needs. Indeed, there may be little point in ‘disturbing’ a large group of users with questions they can neither answer nor understand regarding the specification of data quality parameters. Instead, we should consider developing tools that cater for different user backgrounds and supply default parameter values for non-experts. These users can identify

themselves at the time of log-in and let the system interact in the most appropriate manner from then on.

Similarly, for users browsing spatial data directories on the Internet, we could have a system that requests information about the intended application and then consults a library of spatial data usage histories to suggest possible datasets, algorithms, models and methods to meet the user's need. Alternatively, in some situations it might save time and money if we were able to audit or pre-certify the quality of a spatial dataset as being suitable for a particular task. Users could then be provided with simpler product descriptions coming from organizations that they could trust without having to undertake their own detailed analysis. While data producers have traditionally been reticent to document the applications their digital data might possibly be used for, an agency such as the UK Ordnance Survey handles the matter quite openly and not only states for each of its products the professional groups who are likely to use it and the general application areas for which it is used, but it also provides Internet-based case studies of the actual application of the data (Ordnance Survey 2009).

3.6 Final Remarks

This paper now closes by posing several questions that might help us to understand why we have still not resolved our spatial data quality problems. Firstly, are our difficulties with spatial data quality the result of having to work with legacy system structures designed almost 40 years ago? In essence, our commercial software packages are still based on concepts derived in the 1960s and 1970s when the situation was one of applications searching for computer-based solutions. The tide then turned in the mid-1980s when the software and hardware we needed finally arrived, and continued to improve to the extent that by the 1990s we witnessed technology in search of applications. Perhaps the ebb and flow of scientific and technological development needs to turn again, and we need to see a second generation of geographic information science concepts and systems designed and developed that will handle spatial data in new ways—such as object-oriented, error-aware GIS (Duckham and McCreadie 1999).

Secondly, do we need an entirely new stimulus to drive the data quality issue to a satisfactory conclusion? One possible incentive could come from the many large spatial data infrastructure initiatives being developed worldwide at local, regional, national and global levels. As different agencies (often from different countries) contribute to the development of Geospatial Data Service Centers (GDSCs), Doucette and Paresi (2000) contend that data providers who have taken due care with their quality assurance methods are becoming anxious not to attract unnecessary liability

through their cooperative arrangements with other producers who may not have been so prudent. While there are potential rewards for the participants, the pitfalls are waiting there as well unless they can more effectively deal with data quality reporting and communication.

Finally, have we simply been expecting to achieve too much, too soon, with too many unexpected problems having occurred along the way (as Peuquet 2001 suggests may be the case with developments in space-time data representation)? Or (dare we ask) do we as a scientific community lack the necessary intellectual capacity to overcome our conceptual and technical problems? Certainly, there may be elements of truth in both these propositions—especially when compared to other scientific endeavors. In astronomy, for example, some of our planet's best minds have been continually engaged in refining our knowledge of the universe for several thousand years now, with many wrong assumptions and theories being proposed along the way. Yet it is only in the last 300-400 years that we have started to get things right—even though there are still many unexplained mysteries of the universe to be answered. If people like Galileo, Newton and Hawking—coupled with technology such as the Hubble Telescope—are needed to resolve some of astronomy's fundamental questions, perhaps we need our own equivalents to solve our more modest problems in GIScience concerning spatial data quality.

4 Conclusion

This paper has critically reviewed the problems and prospects associated with the treatment of spatial data quality during the past three decades. While the early years were characterized by warnings from leading researchers and the subsequent development of international standards that included data quality provisions, the original notion of having the tools and techniques needed to assess data quality has generally not yet been achieved. This paper has examined the current problems associated with the description of data quality, its communication and its application in real-life, and it is argued that that we still have a long way to go to fulfill our original visions in each of these areas. The solutions in some cases will slowly occur with time as user education and awareness grows with each generation of spatial data consumers. In other cases, however, greater cooperation and common focus will be needed between the different sectors of the spatial information community if we are to one day see the necessary tools and techniques either embedded in or attached to the commercial systems we now use.

References

- Agumya A, Hunter GJ (1999) A Risk-Based Approach to Assessing Fitness for Use of Geographical Information. *Journal of the Urban and Regional Information Systems Association* 11(1): 33–44
- Bedard Y (1987) Uncertainties in Land Information Systems Databases. In: *Proceedings of the ACSM-ASPRS Auto Carto 8 Conference*, Baltimore, Maryland, pp 175–84
- Beven K (2000) On Model Uncertainty, Risk and Decision Making. *Hydrological Processes* 14: 2605–2606
- Blakemore M (1984) Generalization and Error in Spatial Databases. *Cartographica* 21(2): 131–39
- Boin AT, Hunter GJ (2007a) Facts or Fiction: Consumer Beliefs About Spatial Data Quality. In: *Proceedings of the Spatial Science Institute Biennial International Conference (SSC 2007)*, Hobart, Tasmania, 14–18 May 2007, pp 721–727
- Boin AT, Hunter GJ (2007b) What communicates quality to the spatial data consumer? In: Stein A (ed) *Proceedings of the 2007 International Symposium on Spatial Data Quality (ISSDQ 2007)*, Enschede, The Netherlands, 8 pp
- Brassel K, Bucher F, Stephan E-M, Vckovski A (1995) Completeness. In: Guptill SC, Morrison JL (eds) *Elements of Spatial Data Quality*, International Cartographic Association (ICA) Commission on Spatial Data Quality, Elsevier Science, Oxford, pp 81–108
- Burrough PA (1986) *Principles of Geographic Information Systems for Land Resources Assessment*, Clarendon Press, Oxford
- Burrough PA (1991) The Development of Intelligent Geographical Information Systems. In: *Proceedings of the 2nd European Conference on GIS (EGIS '91)*, Brussels, Belgium, vol. 1, pp 165–174
- Burrough PA, Frank AU (eds) (1996) *Geographic Objects with Indeterminate Boundaries*, London, Taylor & Francis, 345 pp
- CEN (Comité Européen de Normalisation) (1998) *European Prestandard ENV 12656: Geographic Information - Data Description - Quality*, dated October 1998, CEN Secretariat, Brussels, 46 pp
- Chrisman NR (1984) The Role of Quality Information in the Long-term Functioning of a Geographic Information System. *Cartographica* 21(2 & 3): 79–87
- Chrisman NR (1991) The Error Component in Spatial Data. In: Maguire DJ, Goodchild MF, Rhind DW (eds) *Geographical Information Systems: Principles & Applications*, Longman, London, vol. 1, pp. 165–174
- Crosetto M, Tarantola S (2001) Uncertainty and Sensitivity Analysis: Tools for GIS-based Model Implementation. *International Journal of Geographical Information Science (IJGIS)* 15(5): 415–447
- De Bruin S, Bregt AK, Van de Ven M (2001) Assessing Fitness for Use: the Expected Value of Spatial Data sets. *International Journal of Geographical Information Science (IJGIS)* 15(5): 457–471
- Devillers R, Bedard Y, Jeansoulin R, Moulin B (2007) *Towards Spatial Data Quality Information Analysis Tools for Experts Assessing the Fitness for Use*

- of Spatial Data. *International Journal of Geographical Information Science (IJGIS)* 21(3): 261–283
- Doucette M, Paresi C (2000) Quality Management in GDI. In: Groot R, McLaughlin J (eds) *Geospatial Data Infrastructure: Concepts, Cases and Good Practice*, Oxford, London, pp 85–96
- Duckham M, McCreadie J (1999) An Intelligent, Distributed Error-Aware OOGIS. In: Shi W, Goodchild MF, Fisher PF (eds) *Proceedings of the 1st International Symposium on Spatial Data Quality*, pp 496–506
- Eastman JR (1997) *IDRISI for Windows: User's Guide Version 2.0*. Clark University, Worcester, Massachusetts
- Egenhofer MJ, Glasgow J, Gunther O, Herring J, Puequet DJ (1999) Progress in Computational Methods for Representing Geographical Concepts. *International Journal of Geographical Information Science (IJGIS)* 13(8): 775–796
- Epstein EF, Roitman H (1987) Liability for Information. In: *Proceedings of the URISA 1987 Annual Conference*, Fort Lauderdale, Florida, vol. 4, pp 115–25
- ESRI (2000) *Using ArcCatalog*. Environmental Systems Research Institute (ESRI), Redlands, California
- Geoscience Australia (2006) *Geodata TOPO 250K Series 3 User Guide*. Geoscience Australia website, <http://www.ga.gov.au/nmd/products/digidat/250k.htm>, accessed 1 January 2009
- Getis A (2000) Spatial Analysis and GIS: An Introduction. *Journal of Geographical Systems* 2: 1–3
- Goodchild MF (1978) Statistical Aspects of the Polygon Overlay Problem. In: Dutton G (ed) *Proceedings of the First International Advanced Study Symposium on Topological Data Structures for Geographic Information Systems*, Harvard University, Massachusetts, Vol. 6, 22 pp
- Goodchild MF (1992) Research Initiative 1: Accuracy of Spatial Databases - Closing Report. National Center for Geographic Information and Analysis (NCGIA), University of California, Santa Barbara, 19 pp
- Goodchild MF (1993) Data Models and Data Quality: Problems and Prospects. In: Goodchild MF, Parks BO, Steyaert LT (eds) *Environmental Modeling with GIS*, Oxford, New York, pp 94–103
- Goodchild MF (2000) The Current Status of GIS and Spatial Analysis. *Journal of Geographical Systems*: 2, pp 5–10
- Guptill SC, Morrison JL (eds) (1995) *Elements of Spatial Data Quality*, International Cartographic Association (ICA) Commission on Spatial Data Quality, Elsevier Science, Oxford
- Heuvelink GBM, Burrough PA, Stein A (1989) Propagation of Errors in Spatial Modeling in GIS. *International Journal of Geographical Information Systems (IJGIS)* 3(4): 302–322
- Hudson BD (1990) Concepts of Soil Mapping and Interpretation. *Soil Survey Horizons* 31(3): 63–72
- ISO 19113 (2002) *Geographic Information-Quality Principles*. International Organization for Standardization, Geneva, Switzerland
- ISO 19114 (2003) *Geographic information-Quality evaluation procedures*. International Organization for Standardization, Geneva, Switzerland

- ISO 19115 (2003) Geographic Information-Metadata. International Organization for Standardization, Geneva, Switzerland
- Hunter GJ, Qiu J, Goodchild MF (1999) Application of a New Model of Vector Data Uncertainty. In: Lowell K (ed) *Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources*, Ann Arbor Press, Michigan, pp 201–206
- MacDougall EB (1975) The Accuracy of Map Overlays. *Landscape Planning* 2: 25–30
- Marble DF (2000) Some Thoughts on the Integration of Spatial Analysis and Geographic Information Systems. *Journal of Geographical Systems* 2: 31–35
- Minnesota Planning (1999) Positional Accuracy Handbook, Minnesota Planning: Land Management Information Centre, <http://www.mnplan.state.mn.us/press/accurate.html>, accessed 12 November 2001
- Moellering H (ed) (1991) *Spatial Database Transfer Standards: Current International Status*, Elsevier, New York
- NCDCDS (1986) Working Group II on Data Set Quality: Testing the Interim Proposed Standard for Digital Cartographic Data Quality. In: Moellering H (ed) *Issues in Digital Cartographic Data Standards, Report #7*, U.S. National Committee for Digital Cartographic Data Standards (NCDCDS), The Ohio State University, Columbus, Ohio
- NIST (National Institute of Standards and Technology) (1992) *Spatial Data Transfer Standard*. Federal Information Processing Standard 173, US Department of Commerce, Washington, DC
- Openshaw S (1989) Learning to live with Errors in Spatial Databases. In: Goodchild MF, Gopal S (eds) *Accuracy of Spatial Databases*, Taylor & Francis, London, pp 263–276
- Ordnance Survey (2009) OS MasterMap Topography Layer. Ordnance Survey website, www.ordnancesurvey.co.uk/oswebsite/products/osmastermap/layers/topography/index.html, accessed 1 January 2009
- Qiu J, Hunter GJ (2002) A GIS with the Capacity for Managing Data Quality Information. In: Goodchild MF, Fisher PF, Shi W (eds) *Spatial Data Quality*, Taylor & Francis, London, pp 230–250
- Peuquet DJ (2001) Making Space for Time: Issues in Space-Time Data Representation. *GeoInformatica* 5(1): 11–32
- Robinson VB, Frank AU (1985) About Different Kinds of Uncertainty in Collections of Spatial Data. In: *Proceedings of the Seventh International Symposium on Computer-Assisted Cartography (Auto Carto 7)*, Washington, D.C., pp 440–449
- Sadiq Z, Duckham MD (2009) Integrated Storage and Querying of Spatially Varying Data Quality Information in a Relational Spatial Database. *Transactions in GIS* 13(1): 30–42
- Van Oort P, Bregt AK (2005) Do Users Ignore Spatial Data Quality? A decision-Theoretic Perspective. *Risk Analysis* 25(6): 1599–1610

Latent Analysis as a Potential Method for Integrating Spatial Data Concepts

Richard A. Wadsworth¹, Alexis J. Comber², Peter F. Fisher²

¹CEH Lancaster, Bailrigg, Lancaster, LA1 4AP, UK. E-mail: rawad@ceh.ac.uk

²Department of Geography, University of Leicester, Leicester, UK. E-mail: ajc36@le.ac.uk, pff1@le.ac.uk

Abstract

In this paper we explore the use of Probabilistic Latent Analysis and Latent Dirichlet Allocation (LDA) as methods of latent analysis to quantifying semantic differences and similarities between categories. The results are promising, revealing ‘hidden’ or not easily discernable data concepts. LDA provides a ‘bottom up’ approach to interoperability problems for users in contrast to the ‘top down’ solutions provided by formal ontologies. We note the potential for a meta-problem of how to interpret the concepts and the need for further research to reconcile the top-down and bottom-up approaches.

1 Introduction

Many workers have identified differences in data semantics as the major barrier to data integration and interoperability (Frank 2001; Harvey et al. 1999; Pundt & Bishr 2002) and as Frank (2007a) notes, “In order to achieve interoperability in GIS, the meaning of data must be expressed in a compatible description”. The crux of the problem is that the same real world features can be represented in many different ways. The suitability (quality) of a data set is therefore not static or absolute but depends on the

appropriateness of the representation in the context of the user's needs (Frank et al. 2004; Frank 2007b).

Many large datasets depend on multi-disciplinary teams whose members have different conceptualizations of the phenomena being recorded, and who are funded by Research Councils, Government Departments and Conservation Agencies etc who bring their own set of policy, scientific, financial and ethical concerns to the process. The difficulty in achieving interoperability in this context has not been helped by it becoming enmeshed in narrow technical issues related to discovery level metadata and metadata reporting standards.

A "top-down" approach to interoperability might start with the formal assertion that Newtonian physics and Euclidian geometry are sufficient (Frank 2003) and proceed to the development of ontologies, taxonomies and controlled vocabularies into which real data may be placed. We adopt a "bottom-up" approach and consider interoperability from the standpoint of a (naive) data user. We want to know; what the data "labels" *mean*, how the categories are related to each other and did the data producer have the same conceptual understanding of the phenomenon as the user? The final, and perhaps most difficult, task of bridging the top-down and bottom-up approaches has yet to be attempted within both formal ontology research activities such as OWL and emerging e-science infrastructures such as INSPIRE.

In particular we are concerned with *consistency* and *similarity* between data objects and how this affects a user's analysis (Comber et al. 2006). This paper proposes using a text mining approach called Latent Dirichlet Allocation (LDA) (Blei et al. 2003) (a development of Probabilistic Latent Semantic Analysis (PLSA) (Hofmann 1999a,b)) to extract or infer the data concepts contained in written descriptions of spatio-environmental information.

2 Estimating Semantic Consistency

Estimating Semantic consistency can be done in various ways:

- **Declarative Approaches:** Rules (typically *If ... Then ... Else ...*), are used to characterize relationships between objects. Generating rules is difficult, time consuming, and error prone. Rules may be inconsistent through error or because non-monotonic logic applies (consider the *if ... then ... else* rules of the children's game rock-paper-scissors).
- **Semantic Look Up tables:** Relationships are encoded in tables (matrices). Comber et al. (2004a,b; 2005a,b) used expert opinion to encode consistency as "expected", "uncertain" and "unexpected"

relationships in a successful attempt to compare two Land Cover Maps – a problem the data producers warned users was intractable. Wadsworth et al. (2005) decomposed land cover attributed into data primitives before re-integrating them to explore inconsistencies between three land cover maps of Siberia. Fritz & See (2005) used fuzzy logic to average the response of a group of experts.

- **Statistical approaches:** Foody (2004), Hagen (2003), Csillag & Boots (2004) used statistical analysis to compare alternative representations of the same phenomenon in attempts to highlight the locations where variables are incompatible. Kampichler et al. (2000), Maier & Dandy (2000), Guo et al. (2005) and Phillips et al. (2006) made use of Genetic Algorithms and Neural Networks for similar purposes. These approaches are not always robust in the face of “noise”.

The first two approaches (declarative and semantic) rely on the interaction with domain experts (knowledge engineering). The third method requires the user to already have a significant amount of both data sets, while we assume the user may want to perform an assessment before obtaining the data. As experts are not always available we want to try and extract the knowledge that they have “stored” in written descriptions. NLP (natural language processing) (Jurafsky and Martin 2008), especially of scientific texts, is a very complex problem but document categorization and information retrieval making the “bag-of-words” assumption is a much simpler problem. We adapted the work of Lin (1997) and Honkela (1997) to look at the similarity between categories rather than documents (Wadsworth et al. 2006). In an attempt to understand why two categories might be considered similar we are now investigating the potential of two Latent Semantic Analysis techniques (PLSA and LDA) (Hofmann 1999a,b; Blei et al. 2003).

3 Methods

In Latent Analysis the assumption is that there are underlying and unobserved variables (the latent variables) that can be used to explain an observed pattern. In Latent Semantic Analysis the pattern is the frequency of words in documents and the latent variables are concepts (ideas) described in the documents. We can observe the relationship between the documents and words and we want to uncover the latent concepts that can explain the distribution of words in documents. Probabilistic Latent Semantic Analysis (PLSA) was proposed by Hofmann (1999a,b) as a “generative” model of latent analysis; the joint probability that a word (w) and document (d) co-occur ($P(d,w)$) is a function of two conditional probabilities; that the

document contains a concept (z) ($P(z|d)$) and that the word is associated with that concept ($P(w|z)$) (equation 1)

$$P(d, w) = P(d) \sum_{z \in Z} P(w | z) P(z | d) \quad (1)$$

Because we know the frequency of the words in documents ($n(d,w)$) it is possible to rearrange the probabilities to develop an iterative expectation maximization scheme to estimate all the probabilities. The expectation step generates $P(z|d,w)$ while the maximization step calculates $P(w|z)$, $P(d|z)$ and $P(z)$.

When using PLSA there can be problems with “over-fitting” so Hofmann (1999a) proposes using a variation on simulated annealing (called tempered annealing) to prevent this. Unfortunately the tempered annealing requires a “hold out” of test data and most of our data sets are too short to allow this. An alternative approach is to assume that the very skewed frequency distribution of words in documents follow a known distribution; such an assumption leads to a technique called Latent Dirichlet Allocation (Blei et al 2003). Implementations of LDA are available for free in both C and Java (http://en.wikipedia.org/wiki/Latent_Dirichlet_allocation).

Deciding how many latent variables exist is analogous to determining how many classes exist in a fuzzy classification scheme (like c-means). Making the assumption that the probabilities are like membership functions then the indices proposed by Roubens (1982) can be applied. In our implementation of PLSA we used both the Fuzziness Performance Index (FPI) and the Modified Partition Entropy (MPE) to estimate the “best” number of classes from where the sum of the two indices is at a minimum. For consistency we used this “best” number in our explorations of LDA.

Because of restrictions on space we present the results of LDA for only three data sets; the Land Cover Map of Great Britain (LCMGB; Fuller et al. 1994) class descriptions, USDA Soil Orders (Soil Survey Staff 1999) and the abstracts of 677 refereed papers in the International Journal of Geographic Information Science (IJGIS).

4 Results

4.1 Number of Latent Variables in a Data Set

With the PLSA the optimum number of latent variables in the LCMGB land cover example is about 12 (there are 25 categories); this is the minimum of the combined FPI and MPE (Roubens 1982). Because the process may converge to a local minima several trials need to be conducted; Fig. 1 shows the results of five trials. Seven latent variables were found in 12

categories of soil orders; while ten themes were specified (not estimated) for the journal abstract example.

Estimating the “correct” number of latent variables in larger data sets and with LDA is more problematic. With large data sets like the abstracts from IJGIS a hierarchical approach might be preferred, with higher levels of the hierarchy showing the broad trends and lower levels breaking down the finer details and changes over time.

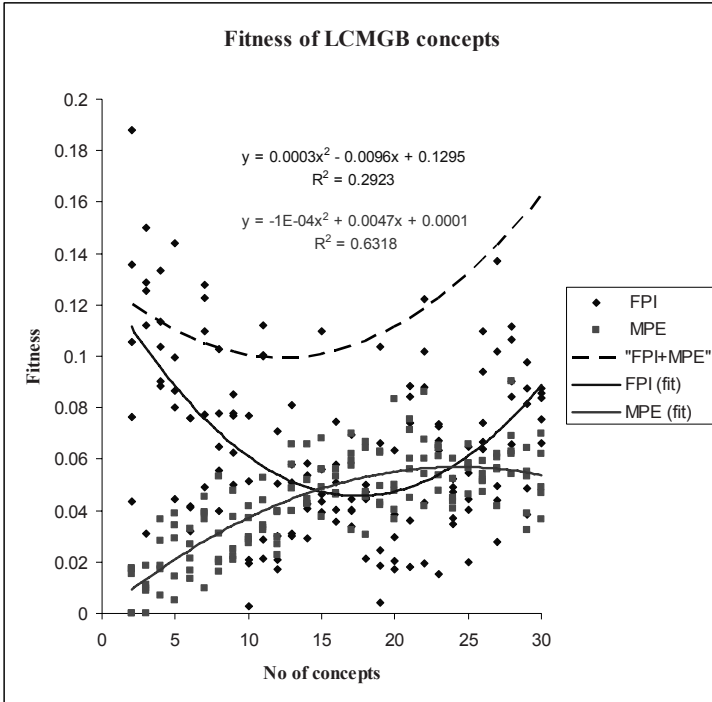


Fig. 1. Fitness measures used to determine optimum number of latent variables (concepts) in the LCMGB categories using PLSA.

4.2 Latent Variables Uncovered by Latent Analysis

Table 1 shows the relationship between the 25 LCMGB target classes and the twelve themes (latent variables) found by LDA.

Table 1. LDA of LCMGB class descriptions into 12 Topics

Topic	LCMGB class(es)	Distinctive Terms (the five words with the highest probabilities)
-------	-----------------	--

A	6 Mown Grazed Turf	Grazed Mown turf swards amenity
B	5 Grass Heath 8 Rough Marsh Grass 12 Bracken	Grassland Species Grass Lowland winter
C	20 Suburban & rural development 21 urban development	Rural Permanent vegetation development developments
D	1 sea estuary 2 inland water	Waters inland sea point water
E	3 coastal bare 4 Saltmarsh	High tides tide water lower
F	19 Ruderal weed 23 Felled Forest	Felled ruderal rough bare ground
G	7 Meadow Verge Semi natural swards	Swards grasslands semi-natural agrostis hay
H	15 Deciduous woodland 16 Coniferous evergreen woodland	Bare deciduous coastal evergreen woodland
I	10 Open shrub moor 11 Dense shrub moor 13 Dense Shrub Heath 25 Open Shrub Heath	Shrub grass dense heath moorland
J	0 unclassified	Cover types some data 25m
K	18 Tilled land (arable crops) 22 Inland bare ground	Bare ground land soil natural
L	14 Scrub Orchard 17 Upland bog 24 Lowland bog	Scrub upland grass lowland species

Table 2 shows the results of specifying 7 latent variables for the USDA Soils orders data sets, both the most probable terms and those with the highest fidelity are listed.

Table 2. Latent variables in the USDA Soil Orders.

Latent variable	Soil orders	10 most probable terms	10 most specific terms
A	Vertisols	vertisols, cracks, open, cropland, united, drainage, low, temperature, vegetation, xererts	vertisols, cracks, open, xererts, system, conductivity, hydraulic, installed, permeability, presents
B	Mollisols	mollisols, temperature,	ustolls, xerolls, cryolls, rich,

		vegetation, moisture, united, forest, grass, epipedon, cropland, plains	tall, addition, albolls, aquolls, drought, limestone,
C	Aridisols Entisols	united, aridisols, diagnostic, temperature, habitat, wildlife, moisture, rangeland, 100_cm, entisols,	psamments, aquents, orthents, recent, boundary, salts, arents, fluvents, sorted, weathering,
D	Gelisols Spodosols	spodic, organic, materials, spodosols, gelisols, united, permafrost, matter, under, alaska,	gelisols, aquods, cryoturbation, orthods, cryods, ice, gelic, iron, applied, good,
E	Histosols Ultisols	organic, united, vegetation, materials, forest, ultisols, cropland, moisture, drained, argillic,	histosols, bulk, density, udufts, ultisol, ustults, botanic, decomposed, fiber, fingers,
F	Andisols Inceptisols	epipedon, forest, united, temperature, vegetation, inceptisols, cambic, moisture, ochric, deposits,	almost, tightly, ustands, xerands, aquepts, plaggen, torands, udands, udepts, xerepts,
G	Alfisols Oxisols	united, moisture, oxisols, vegetation, forest, temperature, crops, association, udalfts, cropland,	oxisols, association, udalfts, xeralfs, alfisols, aqualfts, deciduous, rare, ustalfts, believed

Table 3 shows the results of applying the LDA to 677 abstracts of refereed papers in IJGIS.

Table 3. Abstracts from IJGIS grouped into 10 “themes” by LDA

Cluster	Typical title of a papers	Distinctive words (words with a high probability and high fidelity)
A	Accuracy assessment of digital elevation models using a non parametric approach	Error accuracy errors dem elevation flow propagation estimates interpolation mean input monte carlo terrain cent source positional average square confidence
B	Assessing farmland dynamics and land degradation on Sahelian landscapes using remotely sensed and socioeconomic data	Cover urban neural sdi imagery aggregation pixel agricultural metrics indices agent suitability social sds class city artificial nitrogen landscapes trend

C	A Voronoi based 9 intersection model for spatial relations	Relations topological voronoi tree query join boundary indexing intersection topology diagram relation metric hierarchical graph structures index building formal indices
D	A general model of watershed extraction and representation using globally optimal flow paths and up slope contributing areas	Terrain elevation visibility topographic parallel triangulation dems dem interpolation delaunay surfaces tin parameters variable triangulated irregular aspect channel radiation paths
E	Comparing area and shape distortion on polyhedral based recursive partitions of the sphere	Fuzzy query vague operators uncertain crisp precision arc text gps views membership info geo dbms position insurance soft view shell
F	Analysis of land use drivers at the watershed and household level: Linking two paradigms at the Philippine forest fringe	Soil forest crime erosion regression units risk fuzzy index factors vegetation kappa class moisture predictions loss landslide expert cover membership
G	TERRA VISION the integration of scientific analysis into the decision making process	Urban support criteria ca growth cellular automata factors suitability programming economic vulnerability sdss sensitivity group parameters making transition integrated sustainable
H	A proposed framework for feature level geospatial data sharing: a case study for transportation network data	Temporal spatio phenomena geography generic dynamics current census distributed events event internet matrix agents geospatial individuals behaviour spatiotemporal relationship transportation
I	Data gathering strategies for social behavioral research about participatory geographical information system use	Technology project national government state science community technical technologies current activities social benefits article discussion role countries support education efforts
J	Colour coded pixel based high-interactive Web mapping georeferenced data exploration	Cartographic generalization interactive path data-rough mining solution task ontology categorical original discovery categories interpretation display exploration facility base

5 Discussion and Conclusions

Although the description of the LCMGB classes are rather short Latent Analysis has managed to identify some reasonable concepts (reasonable in the eyes of a domain expert). Unfortunately, some of the concepts are rather more difficult to interpret and may reflect statistical artifacts or the lack of words to process. A problem with stochastic approaches like the PLSA is that repeated “runs” on the same data set do not always result in the same groups being “discovered”; in this respect LDA is much more stable. When applying the approach to other data sets we have had mixed results. In descriptions of soil orders the main problem facing the non-expert is being unfamiliar with the terms used, but, by using latent analysis first it helps the user to identify which of these unfamiliar terms are likely to be the most important and therefore should be “de-coded” first; for an

expert it may help understand the subtleties of the categorization process used. When applying the approach to the abstracts from the IJGIS the method produced apparently interpretable “clusters”, however, the size of the data set and the difficulty in deciding what constitutes an appropriate number of clusters suggests that a more hierarchical approach might be better.

Where human domain experts exist then knowledge engineering methods can codify their expertise in ways that make inter-operability a practical proposition. Domain experts may not exist or may not be accessible (through time constraints or geography) in those cases where domain experts have expressed their expertise through *long* textual descriptions text mining can produce acceptable estimates of semantic similarity. A “reconnaissance” assessment of PLSA and LDA suggests that LDA may go some way to explain why concepts are considered to be similar.

As yet the task of reconciling the top-down and bottom-up approaches to interoperability remain unexplored but the latent analysis approaches can be applied to more than one dataset to identify classes (i.e., documents) with shared concepts to facilitate data integration.

References

- Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet Allocation. *Journal of machine Learning Research* 3: 993–1022
- Comber A, Fisher P, Wadsworth R (2004a) Integrating land cover data with different ontologies: identifying change from inconsistency. *International Journal of Geographical Information Science (IJGIS)* 18(7): 691–708
- Comber AJ, Fisher PF, Wadsworth RA (2004b) Assessment of a Semantic Statistical Approach to Detecting Land Cover Change Using Inconsistent Data Sets, *Photogrammetric Engineering and Remote Sensing*, 70(8), pp 931–938
- Comber AJ, Fisher PF, Wadsworth RA (2005a) A comparison of statistical and expert approaches to data integration. *Journal of Environmental Management* 77, pp 47–55
- Comber AJ, Fisher PF, Wadsworth RA (2005b) Combining expert relations of how land cover ontologies relate. *International Journal of Applied Earth Observation and Geoinformation* 7(3): 163–182
- Comber AJ, Fisher PF, Harvey F, Gahegan, M, Wadsworth RA (2006) Using metadata to link uncertainty and data quality assessments. In: Riedl A, Kainz W, Elmes G (eds) *Progress in Spatial Data Handling, Proceedings of SDH 2006*, Springer, Berlin Heidelberg, New York, pp 279–292
- Csillag F, Boots B (2004) Toward comparing maps as spatial processes. In: Fisher P (ed) *Developments in Spatial Data Handling*, Springer, Berlin Heidelberg New York, pp 641–652
- Foody GM (2004) Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy. *Photogrammetric Engineering and Remote Sensing* 70 (5): 627–633

- Frank AU (2001) Tiers of ontology and consistency constraints in geographical information systems. *International Journal of Geographical Information Science (IJGIS)* 15(7): 667–678
- Frank AU (2003) A linguistically justified proposal for a spatio-temporal ontology. In: COSIT'03, Conference on Spatial Information Theory 24-28 September, Ittingen, Switzerland
- Frank AU (2007a) Towards a Mathematical Theory for Snapshot and Temporal Formal Ontologies. In: The European Information Society, Lecture Notes in Geoinformation and Cartography, Springer, Berlin Heidelberg New York, pp 317–334
- Frank AU (2007b) Incompleteness, error, approximation, and uncertainty: An ontological approach to data quality. In: Morris A, Kokhan S (eds) *Geographic Uncertainty in Environmental Security, Proceedings of NATO Advanced Research Workshop on Fuzziness and Uncertainty in GIS for Environmental Security and Protection*, Kyiv, Ukraine, JUN 28-JUL 01, 2006, pp 107–131
- Frank AU, Grum E, Vasseur R (2004) Procedure to select the best dataset for a task. In: *Proceedings of Geographic Information Science, Lecture Notes in Computer Science Vol 3234*, pp 81–93
- Fritz S, See L (2005) Comparison of land cover maps using fuzzy agreement. *International Journal of Geographical Information Science (IJGIS)* 19(7): 787–807
- Fuller RM, Groom GB, Jones AR (1994) The Land Cover Map of Great Britain: an automated classification of Landsat Thematic Mapper data. *Photogrammetric Engineering and Remote Sensing* 60: 553–562
- Guo QH, Kelly M, Graham CH (2005) Support vector machines for predicting distribution of sudden oak death in California. *Ecological Modelling* 182(1): 75–90
- Hagen A (2003) Fuzzy set approach to assessing similarity of categorical maps. *International Journal of Geographical Information Science (IJGIS)* 17(3): 235–249
- Harvey F, Kuhn W, Pundt H, Bishr Y, Riedemann C (1999) Semantic interoperability: A central issue for sharing geographic information. *Annals of Regional Science* 33(2): 213–232
- Hofmann T (1999a) Probabilistic latent semantic indexing. In: Hearst M, Gey F, Tong R (eds) *Proceedings of 22nd International Conference on Research and Development in Information Retrieval* University of California, Berkeley, California, Aug, 1999, pp 50-57
- Hofmann T (1999b) Probabilistic latent semantic analysis. In: Laskey KB, Prade H (eds) *Proceedings of 15th Conference on Uncertainty in Artificial Intelligence*, Royal Inst Technol, Stockholm, Sweden, Jul 30-Aug 01, 1999, pp 289–296
- Honkela T (1997) Self-Organising maps in natural language processing. PhD thesis, Helsinki University of Technology, Department of Computer Science and Engineering, <http://www.cis.hut.fi/~tho/thesis/>
- Jurafsky D, Martin JH (2008) *Speech and Language Processing (Second edition)*, Prentice Hall Series in Artificial Intelligence

- Kampichler C, Dzeroski S, Wieland R (2000) The application of machine learning techniques to the analysis of soil ecological data bases: relationships between habitat features and Collembola community characteristics. *Soil Biology and Biochemistry* 32: 197–209
- Lin X (1997) Map displays for information retrieval. *Journal of the American Society for Information Science* 48: 40–54
- Maier HR, Dandy GC (2000) Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications *Environmental Modelling and Software* 15: 101–124
- Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling* 190(3-4): 231–259
- Pundt H, Bishr Y (2002) Domain ontologies for data sharing—an example from environmental monitoring using field GIS. *Computers and Geosciences* 28(1): 95–102
- Roubens M (1982) Fuzzy clustering algorithms and their cluster validity. *Eur. J. Oper. Res.* 10: 294–301
- Soil Survey Staff (1999) *Soil Taxonomy A Basic System of Soil Classification for Making and Interpreting Soil Surveys*, 2nd Edition, Natural Resources Conservation Service Number 436. U.S. Government Printing Office Washington, DC 20402 (available at ftp://ftpfc.sc.egov.usda.gov/NSSC/Soil_Taxonomy/tax.pdf, accessed 7 October 2007)
- Wadsworth RA, Comber AJ, Fisher PF (2006) Expert knowledge and embedded knowledge: or why long rambling class descriptions are useful. In: Riedl A, Kainz W, Elmes G (eds) *Progress in Spatial Data Handling, Proceedings of SDH 2006*, Springer, Berlin Heidelberg New York, pp 197–213
- Wadsworth RA, Fisher PF, Comber A, George C, Gerard F, Baltzer H (2005) Use of Quantified Conceptual Overlaps to Reconcile Inconsistent Data Sets. In: *Proceedings of GIS Planet 2005, Session 13 Conceptual and cognitive representation*, Estoril Portugal 30th May - 2nd June 2005. ISBN 972-97367-5-8. 13pp

Stereology for Multitemporal Images with an Application to Flooding

Alfred Stein, Petra Budde, Mamushet Zewuge Yifru

ITC, PO Box 6, 7500 AA Enschede, The Netherlands

Abstract

This paper presents stereology for flooded areas observed on a multitemporal remote sensing image. Stereology is a mathematical method to quantify objects at one dimension from simulated objects at a lower dimension. It was initially developed for geological and soil objects. Here it is applied to objects on multitemporal remote sensing images, i.e. for image mining. Image mining considers the chain from object identification from remote sensing images through modeling, tracking a series of images and prediction, towards communication to stakeholders. The paper introduces the estimation of the area size of the same object observed at various moments in time. It is illustrated with a case study on flooding of the Tongle Sap lake in from Cambodia.

1 Introduction

Remote sensing images are available at an increasing frequency and spatial resolution and from a large variety of sensors. Monitoring is thus increasingly benefitting from the presence of satellites. Commonly a spatial object is identified, and several of its properties are followed in time. Examples are the monitoring of cities, monitoring forests and of deforestation,

and the monitoring of drought. Apparently, both spatial and temporal dimensions have their own uncertainties: observation of the area is usually not done continuously, whereas the object can be uncertain at each moment of observation. Simply following an object is usually not sufficient for decision making. A division can thus be made into the signaling of threshold exceedance and the spotting of unexpected events, such as landslides and outbreak of pests in an agricultural field. The information thus collected has as a requirement that it is communicated to stakeholders. Image mining has been shown to be a useful system for doing so. Despite this development, uncertainty is still largely present, and information should be communicated as condensed as possible. In this regard, uncertainty plays a critical role as well, being both a problem and an asset to spatial information (Frank 2008).

We thus increasingly see it as an important step to summarize the information of spatial objects in a quantitative way. For any spatial object both geometrical properties and attributes are to be considered. In the recent past, measuring the size of an object was an important and relevant step. Spatial data quality is a cornerstone in this area, quantifying lack of quality caused by mis-classification, poor object definition and timeliness of the data. This has resulted in the development of spatial data quality as a scientific discipline (Shi 2009; Stein et al. 2008). We realize, however, that with the increasing amount of information *in time* the need is increasing to better quantitatively assess this information. In a risk analysis, for example, a long lasting disaster may have a much larger impact, say on crop production, than a short one, although both can be equally devastating.

In this study, we use stereology for multitemporal remote sensing images, focusing on estimation of 2D objects. The more general approach to estimate $2D \times T$ objects will be postponed to another paper. Stereology is an unbiased and effective tool to obtain quantitative 2D geometric properties (e.g., number, length, area, volume etc.) from recorded series of sections in a lower dimension. Stereology can be stated as a method for solving the problem of measuring a physical object in n dimensions from random measured objects in less dimensions (Baddeley 1991; Baddeley and Vedel Jensen 2005). In an earlier paper (Stein et al., 2009) we focus on 2-dimensional objects on a single image. We now extend this approach towards multitemporal images. To do so, we turn towards image mining.

Image mining is a relatively new development focusing on extracting relevant information from large sets of remote sensing images (Stein 2008; Silva et al. 2008). It can be defined as “The analysis of (often large sets of) observational images to find (un)suspected relationships and to summarize the data in novel ways that are both understandable and useful to stakeholders”. Image mining in space has as a focus to combine a large set of similar images, in order to identify similar objects. In space, it concerns the

classification and segmentation, for example using textural image segmentation as a first step, to identify for example forest fires, flooded parcels, deforested patches or sub-areas of a rich biodiversity. Its main objective is to reduce uncertainty, allowing making better decisions (Frank 2008). An important step during image mining is thus the identification of spatial patterns and testing of their significance. Image mining recognizes issues of data quality as a crucial element. Data quality depends upon the fitness for use, i.e. the required decision and hence on the stakeholders' interests.

The aim of this paper is to present essentials of stereology towards analysis on multitemporal remote sensing images. The study is illustrated with an example from flooding in Cambodia.

2 Stereology

2.1 Basics

Stereology is an unbiased and effective tool to obtain quantitative geometric properties like number, length, area and volume from recorded series of sections of a lower dimension (Vedel Jensen 1998). It can be used to measure a physical object in n dimensions from randomly measured objects in fewer dimensions. It is based on Delesse (1848) who applied it to determine the volume of a number of minerals in rocks. He showed that the areal fraction occupied by a given mineral on the section of a rock AA is proportional to the volume fraction of the mineral within the rock volume VV . This notation identifies A or V as the object dimension of interest, and the subscripts as the dimension of the sampling unit (Baddeley 1991; Baddeley and Vedel Jensen 2005). A basic principle of stereology is Cavalieri's principle (Cavalieri 1635): "if two solid objects have equal plane sections on all the intersecting planes ($A_h, h \in T$), then the objects have equal volume". The principle of stereological measurement is by taking the geometry and probability statistics of an object into consideration. One can consider estimation of the area of an object within a window by allocating, say, m random test points to that window and using the fraction of points that falls on the object as a measure of the proportion of the object covering the window, and hence as a measure for the area of the object. Similarly, one can randomly distribute m random test lines to the window and the fraction of those lines within the object gives an estimate of the proportion of the object falling within the window. These two examples are typically two-dimensional, but stereology is applicable to any set of dimensions. For a 1-dimensional line object, containing a line fragment of interest, one may allocate random points and count the fraction of points falling on the segment and have that as the fraction of the line

fragment to the line object. In a 3 dimensional window, interest may be on a 3D object, covering a part of that window and one may wish to estimate the fraction of the window occupied by that object. This volume could be either estimated by allocating random points, or random lines, or random planes inside the window. Similarly, when interest concerns a surface, the area of the surface could be estimated by allocating random lines in the window and identifying the number of intersecting points.

Two important concepts in stereology are its unbiasedness and the way of taking samples. Unbiasedness implies that by taking enough samples the calculated property of the volume gives the population value. Stereological estimators are unbiased estimators. Sampling concerns the choice of the sampling objects within the window. This gives a distinction between classical and modern stereology. On the one hand, classical stereology, or model-based stereology, is based on the assumption that the material is homogeneous in composition, i.e. a geometric assumption about the structure of the object (Baddeley 1991; Ross and Dehoff 1999). On the other hand, design-based stereology requires no assumptions regarding the geometric aspect of the feature of interest, such as its shape, its size, and its orientation, as well as by the use of systematic random sampling procedures. Design-based stereological sampling relies mainly on systematic sampling (Glaser and Glaser 2000). For example, a test line is selected randomly and the sample is assumed to be arbitrary and fixed. These ensure that each feature of interest in the specimen has an equal probability of being sampled (Baddeley and Vedel Jensen 2005). Design-based stereology is effective and suited in estimating global and nonhomogeneous populations (Baddeley and Vedel Jensen 2005). Like model-based stereology it results in an unbiased estimate of a high precision of geometric quantities like volume, surface area or perimeter of an object.

Applications of stereological techniques are useful in a broad variety of disciplines, such as, in biological, medical, material science, food science, and other fields. So far, much research has been done in the field of biomedical science to estimate quantitative information about 3D microscopic structures, based on 2D observations, especially to obtain a deeper understanding of the structure and function of human body cells and a more objective diagnosis of progressive disease assessment (Roberts et al. 2000). Stereology has also been used as a precise, simple and efficient means of quantifying three-dimensional microscopic structures from two-dimensional quantitative sections. Stereological volume estimations are also applied as sums of measurements of serial sections performed on sampling areas measured on thin sections (Dorph-Petersen et al. 2000; Kötzer 2006). But so far, applications of stereology to remote sensing images have been missing and it is little known in geoinformation science (see Stein et al. 2009 for a short introduction).

2.2 Identifying 2D Objects from Images

In this study we aim to use stereology within an image mining as stereology may serve as an opportunity to handle the large amount of data thus collected. In our study we consider 2D objects of which we estimate the area using random straight lines, i.e. an L_L estimation.

Stereology extracts structural quantities from measurements made on 2D images: the surface area (S) of a 2D object, the length of lines (L) and number of points (n). We apply the standard notation of the objects together with their subscripts indicate the way that the measurements are made with respect to this object. Subscripts indicate the type of measurements that have been used, for example, measurements on an image per volume area have a subscript $_A$. Stereology thus considers S_V as a specific surface area of a volume, for example the surface of a 3D object per unit volume of this object. Similarly, L_V is a specific line length of (length per unit volume) of a curve or line structure. This applies as well to the length per area L_A as the length per unit area for a 2D object, the number of points per area unit n_A , n_L as the number of points per linear unit and A_A as the (dimensionless) area fraction per unit area.

Table 1. Notation for geometrical properties (after Baddeley 1991)

Dimension	set X	Symbol	Meaning
Space (2D)	Plane domain	A	Area
	Curve	L	Curve length
	Finite set of points	P	Number of points
	Finite set of objects	Q	Number of objects
	Curve	C	Total curvature

We next move towards estimation and we do this by the use of test lines. In doing so, we may consider linear objects, for example the length of a perimeter of an object on an image can be determined from the number of times it meets a straight line and the length of a curve that can be determined from the number of intersection points it makes with random lines. Common stereology estimators are denoted as \hat{L} as an estimator for the length of a line per unit area, \hat{A}_A as the surface area estimator per unit area, \hat{L}_L as the area estimator for the length per unit length, etc. and \hat{V} as the volume estimator. Table 2 shows basic stereological formulae for expected measurable parameters.

Table 2. Basic stereological formula for expected measurable parameters.

	Dimension			
	3	2	1	0
0	\hat{V}_V	\hat{A}_A	\hat{L}_L	\hat{P}_P
1	\hat{S}_V	$\frac{4}{\pi} \hat{L}_A$	$2\hat{P}_L$	
2	\hat{L}_V	$2\hat{Q}_A$		

2.3 Remote Sensing Images

We now consider an individual remote sensing image and we identify a region of interest with the object and the image in full with the window. A sharp and unambiguous identification of the object is determined by different factors, such as atmospheric conditions, time of observation. Moreover, the spatial resolution of an image is limited, leading to a blocky structure when zooming in, and also the definition of the object may be such that it is only vaguely defined. Moreover, the object may show internal variation and its identification and classification is subject to various uncertainties. At this stage, however, we will not venture too much into issues of data quality, but will focus on an object that is clearly identifiable, relatively homogeneous, of a sufficiently high resolution, and where atmospheric distortion is virtually absent. Interest will focus on determining the size of the area, and estimation will be done by random lines. When using a stereological estimator, the number of lines has to be decided upon in advance. This depends largely upon the desired quality. An increasing number of lines will lead to an increasingly precise estimate. As no general statements are given, we will make it part of this research.

2.3.1 Image Mining

Image mining (Stein 2008; Pereira dos Santos Silva et al. 2008), considers the analysis of observational images to find (un)suspected relationships and to summarize the data in novel ways that are both understandable and useful to stakeholders (Fig. 1). Objects representing earth processes discernable on those images can be either crisp or fuzzy and vague.

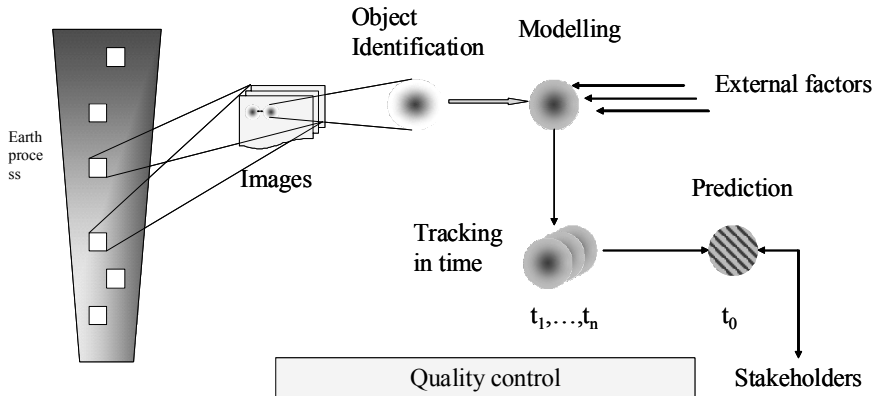


Fig. 1. Image mining of spatial and temporal earth processes.

We distinguish five key steps in image mining: identification, modeling, tracking, prediction and communication with stakeholders. All these steps will be briefly discussed below. On top of this, we notice aspects of spatial data quality in each of these steps.

Identification of an object requires making the step from raster to objects by grouping grid cells with similar digital numbers into one object. That is usually done by applying an image segmentation technique. Procedures for image segmentation are well documented, e.g. based on mathematical morphology, edge detection, identifying homogeneity in one band or in a set of bands, and on texture based segmentation (Glasbey and Horgan 1995; Ojala and Pietikäinen 1999; Lucieer et al. 2005). This may include as well uncertainty values. Segmentation is followed by a classification towards a set of object classes (Richards et al. 1998). Considering 2 classes at the moment (the pixel belongs to the object, or the pixel does not belong to the object), this results in an object X and the remainder of the window X^c . Image mining next considers modeling of this object. As our primary focus is on geometrical aspects, however, we will not proceed further here (see Stein 2008). Also the tracking of objects, i.e. following of the object in time, will not be explored in all detail. For multitemporal images, it is sufficient at this stage to restrict ourselves to objects that are well identified and correspond physically to objects at earlier stages. In fact, the objects are observed as snapshots of a process. It may be important in image mining to predict properties of an object at a moment t_0 beyond the observation period. This can be done by a parametric curve for the centroid and other parameters, e.g., by a linear statistical model (Rajasekar et al. 2006). The final stage concerns communication to stakeholders. Various ways exist here, ranging from simple visualization towards assessments of costs and benefits. Issues from decision support typically are required here (Van

de Vlag and Stein 2006). In our study exploring the possibilities of stereology, we will aim to communicate directly the area of an object as a number, possibly summarized by some simple graphical output.

2.3.2 Application

This application considers flooding of the Tonle Sap Great Lake in Cambodia (Fujii et al. 2003). The study area lies on the lower part of Mekong region of Cambodian floodplain following the Mekong river (Fig. 2). The country is characterized by five distinct topographic features: the sandstone Dangrek range in the north, forming the border with Thailand, the granite Cardamom Mountains with peaks of over 1500 m in the southwest, the Darlac Plateau which rises to over 2700 m and in the north-east and the Central Plains between 10 and 30 m above sea level, which form 75% of the Cambodian land area. Flooding due to high water levels of the Mekong River and its tributaries recurs yearly. It causes a considerable damage on human settlements, agricultural activities and infrastructures of the surrounding area. The Tonle Sap Great Lake covers a relatively small area in the dry season and increases to three to four times this area during the wet season. The study area is bounded within a box with the coordinates $12^{\circ}06'25''$ N to $13^{\circ}55'56''$ N and $102^{\circ}29'13''$ E to $104^{\circ}28'51''$ E.

Application of remote sensing to a flooded area in a densely vegetated area identifies the object of study by reflectance values that deviate from those in its neighborhood. Starting from May 2001 to January 2002, a series of 9 Landsat 7 ETM+, multi-spectral images were used. The area extent of the lake varies depending on the observation time. Images were collected on the following dates: 31 May (Early flood), 16 June (Early flood), 2 July (Early flood), 18 July (Start of rising flood), 3 August (Rising flood), 4 September (Rising to peak flood), 20 September (Peak flood), 23 November (Peak to falling flood) and 10 January (Falling flood stage). These moments were determined by visibility of the lake from the Landsat 7 ETM+ sensor: only images that contained less than 15% clouds were useful.

The stereological test system for this study consists of test lines, a known frame (i.e., the full images) and 9 incidentally recorded satellite images for information extraction. In this particular stereological test system, we equate the flood to our feature of interest Y_t , $t = 1, \dots, 9$ in the spatial domain \mathbb{R}^2 , and we equate the survey area with X_t , $t = 1, \dots, 9$. Apparently, $X_t \subset \mathbb{R}^2$ contains a subset $Y_t \subset X_t$.

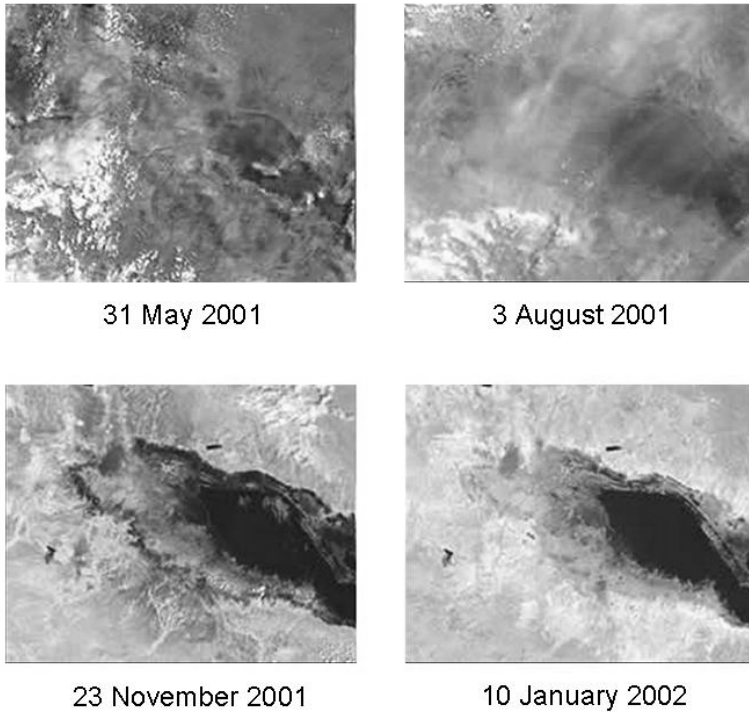


Fig. 2. Four Landsat images of the process of flooding around the Tonle Sap Great Lake in Cambodia during the 2002 rainy season (Yifru 2006).

We describe image mining for ‘flooded area’ as follows. At $t_1 = \text{May } 31^{\text{st}}$ we observe n_1 objects: one large object and a series of $n_{1,i}$ small objects, characterized as $X_{1,i}$, $i = 1, \dots, n_1$ and we have $X_1 = \bigcup_{i=1}^{n_1} X_{1,i}$ (Fig. 3). There is little change as compared to $t_2 = \text{June } 16^{\text{th}}$. The objects $X_{1,i}$ can simply be tracked to objects $X_{2,i}$, equating objects at similar position but at different moments in time with each other. As before, $X_2 = \bigcup_{i=1}^{n_2} X_{2,i}$. At $t_3 = \text{July } 2^{\text{nd}}$, some of the smaller objects expand, and some objects merge yielding one large and several small objects. Understanding the flooding process requires a careful tracking to relate the n_3 objects at t_3 with the n_2 objects at t_2 . We also notice the birth of several new objects. At $t_4 = \text{July } 18^{\text{th}}$, the large object is increasing further and one of the objects at t_3 is decreasing and splitting into some smaller objects. Changes from t_4 to t_5 are small, but we

observe a decrease in the size of the largest object. At $t_6 = \text{September } 4^{\text{th}}$ we notice that the objects have now merged into a single large object, labeled as $X_{6,1}$, which we also observe as the single object $X_{7,1}$ at $t_7 = \text{September } 20^{\text{th}}$, being of a somewhat smaller size. During the next period the flooding apparently reduces and at $t_8 = \text{November } 23^{\text{rd}}$ we notice a reduction in the size of $X_{7,1}$ to become $X_{8,1}$ and the birth of several, say $n_8 - 1$ smaller objects. At $t_9 = \text{January } 10^{\text{th}}$, we notice that the largest object, now labeled $X_{9,1}$ is of a comparable size as at t_1 but of a somewhat different shape. At all each moment t the object of interest equals $X_t = \bigcup_{i=1}^{n_t} X_{t,i}$.

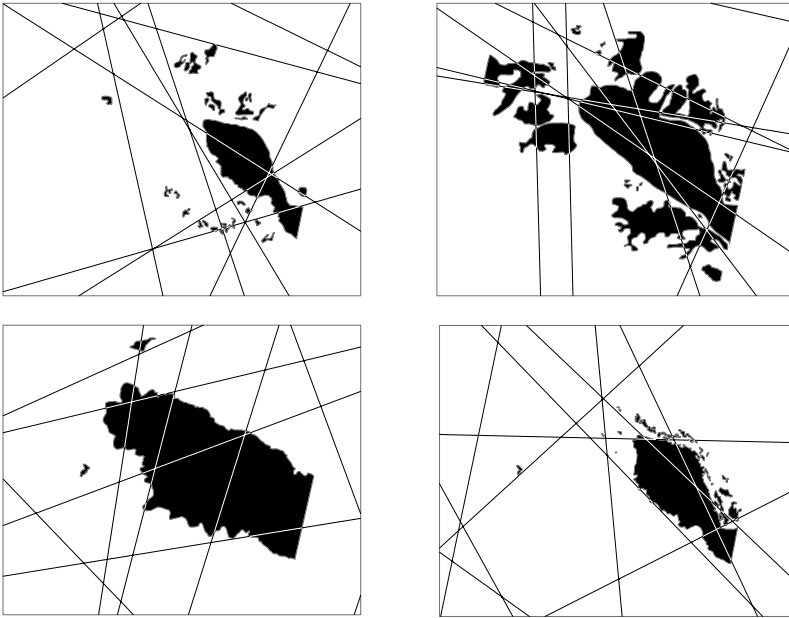


Fig. 3. Four flooding objects extracted from the Landsat images around the Tonle Sap Great Lake in Cambodia during the 2002 rainy season. Random lines are drawn to estimate the area size.

We use stereology to identify the size of the different lakes on each of the images. Initially we used $m = 1000$ lines for the estimation. As an example, we show 10 of such liens drawn at random at each of the four instances (Fig.4).

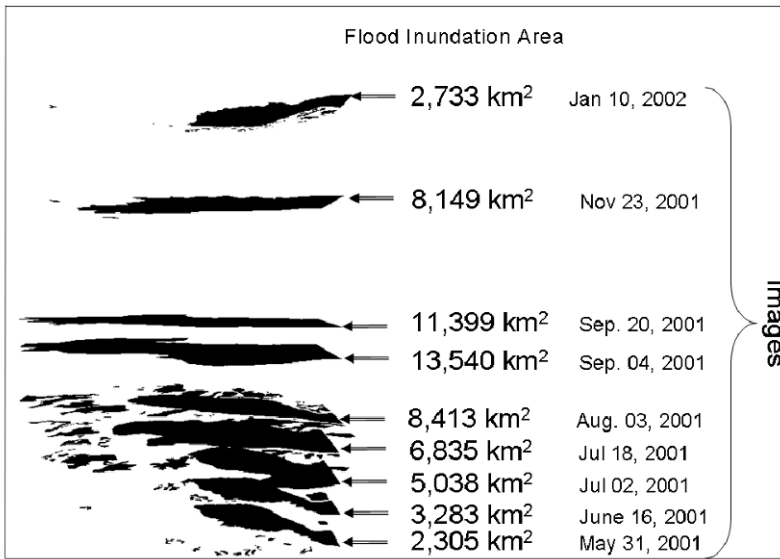


Fig. 4. Flooding objects identified in the Tonle Sap Great Lake area at the 9 moments of observations.

	31-5	6-6	2-7	18-7	3-8	4-9	20-9	23-11	10-1
Area									
Mean	0.0534	0.0767	0.1171	0.1558	0.1943	0.3128	0.2638	0.1887	0.0634
Mean (km ²)	2337	3334	5078	6757	8440	13497	11409	8154	2752
Min	0.0516	0.0737	0.1131	0.1500	0.1884	0.3062	0.2529	0.1820	0.0601
Max	0.0556	0.0805	0.1217	0.1621	0.2014	0.3220	0.2737	0.1997	0.0663
Max (km ²)	2396	3472	5248	6989	8687	13889	10908	8611	2590
Sd.	0.0009	0.0013	0.0017	0.0024	0.0024	0.0031	0.0033	0.0027	0.0013
Observed	0.0542	0.0773	0.1177	0.1567	0.1957	0.3130	0.2645	0.1891	0.0638
Error									
Mean (km ²)	36	29	28	39	60	9	30	16	20
Max (km ²)	59	138	170	232	247	392	501	457	162

Estimates of the area size obtained with 1000 random lines as proportion of the window and in km²; the error is expressed as the difference between expected and worst estimate of 100 simulations.

We notice that the estimates are precise, with standard deviations equal to approximately 2 % and measured and average of the simulations very close (within the mean $\pm 2 \times$ sd). This applies to all of the different dates for which observations were available. As the size of the lake is big, however, we notice that the differences as expressed in km² are substantial: on the average (for the 100 simulations) it ranges to areas of 9 to 40 km², whereas maximum values equal to more than 500 km² may emerge.

Fig. 5 shows for each time the empirical densities of the distribution of the surface areas and the $N(\mu, \sigma)$ densities at $t = 1, \dots, 9$ obtained with 100 estimates. We notice that differences can be substantial, for example at September 20th the empirical distribution shows more variation than could be represented by a normal distribution. It leads to the conclusion that a set of 100 lines (or less) may not be adequate to derive properties of the area.

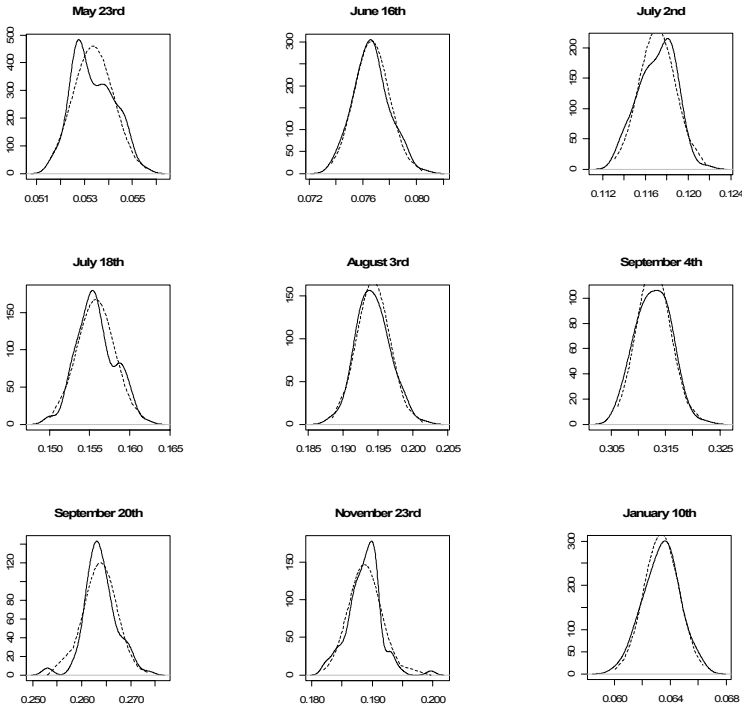


Fig. 5. Estimated densities and corresponding Normal distributions obtained with 100 estimated lines.

One question that we tried to answer with this study was the number of simulations necessary to do these calculations. Hence we did the calculations also with less random lines, leading to the results shown in Fig. 6.

We notice an expected gain in precision, for example expressed by the length of the whiskers. The number of $m = 1000$ lines may be sufficient for various purposes, but to obtain even better estimates this number may be further increased. We have to realize, though, that the limited resolution of the images may put a limit to the increase in precision. A number of m below 1000 seems not to be very attractive, and may result into unnecessary uncertainty, in particular as computation time for these objects is short.

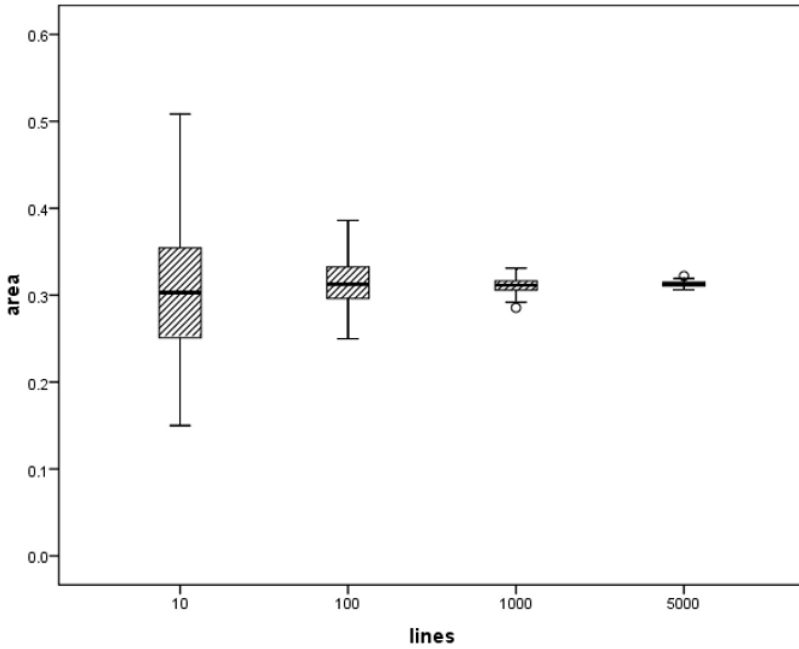


Fig. 6. Effect of the number of lines on the precision of the areal estimate.

3 Discussion

Stereology as it was developed during the last decades seems to be of interest to be applied to multitemporal images. It offers many possibilities to estimate geometrical aspects of spatial objects. With the further sophistication of image processing software, the possibilities might be further explored. In this study, we applied random lines for area estimates, but similar estimates might have been applied to a range of different other parameters, like the perimeter of an object, the length of the largest diameter, or simply the number of sub-objects constituting the single object as analyzed so far.

Stereological estimates may also play a role in data quality issues. Spatial data quality affects GIS based decision making activity (Shi et al. 2002). When using data as an input in numerical models, data precision should be taken care of, in particular if it propagates towards the decision. In this study, the accuracy of stereologically estimated 2D area by passing different set of lines (10, 100, 1000, and 5000) depends upon the number of

lines drawn to quantify the area and the lines passing through the image should be selected at random. It remains to the user of information, however, which precision is required. A land planner may be satisfied with a relatively coarse number, whereas an individual inhabitant of the area might require a much higher precision to have information about his ground at a sufficiently precise level.

Spatial data quality refers to various aspects of data quality as can be identified for in geographical objects. Although various aspects can be considered, we restrict ourselves to the two main components in this paper, namely positional accuracy and attribute accuracy. Apparently the two are closely related to each other. For uncertain objects, positional accuracy is dealt with by characteristics of the membership function, such as the support of a fuzzy set, the shape of the membership function and characteristics of its α -shapes. The attribute accuracy is identified by the content of the membership function, i.e. its relation towards the object under study. It basically answers the question to which degree the membership function expresses the concepts that are displayed. Extension of stereology to these objects has to be done.

The algorithm has been implemented in a Python environment. Estimators in the implemented algorithms from the competition of fraction of each line lying within the object representing the flooded area. To estimate a 2D area it counts the pixels on those lines for the case of 2D area estimation. The general trend of the stereologically estimated 2D area enable us at what time interval a significant change in the pattern occurs.

Flooding apparently is one of the phenomena that we can characterize using stereology in an image mining context. It starts at some moment in time, it may be poorly visible at several moments in time, because of (partial) cloud cover, it increases in size, it may split, several objects may merge, and after the river withdraws, the flood ends and the object reduces to its original size. Moreover, also the boundaries between flooded and non-flooded land are difficult to draw, if possible at all and (partial) cloud cover may prohibit its precise and detailed observation. Flooding can further be interpreted from information from various perspectives like natural, ecological, environmental or socio-economical information. It is a further challenge to include this all into the stereological context.

4 Concluding Remarks

We conclude from this study that stereology presents a set of estimators that are very helpful to estimate geometrical properties of objects that are clearly and unambiguously identifiable from remote sensing images. Thus

stereology could provide a valuable set of tools to be generally applicable in image mining. It has been shown to be applicable for area estimation by random line objects, whereas an extension to other properties by different estimators is applicable as well. In particular, estimation of area sizes by allocating random points might be simply applicable. Of slightly more interest may be the estimation of the length of linear objects, such as roads and rivers, by applying random lines. This would require a careful analysis of multiple hits, but we believe that it is solvable. Estimated values can be of any precision, restricted by the limited resolution of the images and by computation time, which has been short for this study.

For the application considered in this study, there is a clear indication that the number of applied random lines is satisfactory when considering the fraction of the area that is flooded, but that the size of that area expressed in km² may be of a too low precision for many practical purposes. In the future we aim to further explore possibilities for extending stereological estimators into the space-time domain.

At this stage, we could imagine that research questions to be resolved are primarily of a topological nature, i.e. the lengths and areas of objects. But after some extension, it might be possible as well to consider uncertain objects and research could extend towards identifying fuzziness of such objects, for example by selecting α shapes for several values of α , and translating areas (or lengths) thus identified towards object sizes for the same values of α . Further research is needed in these directions. A possibly interesting research direction could also be towards separability of classes when the number of classes is larger than 2. It is not difficult to imagine that class overlap may lead to ambiguity in topological properties, and robust solutions should be provided.

References

- Baddeley A (1991) Stereology. In: *Spatial Statistics and Digital Image Analysis*, chapter 10, National Research Council USA, Washington DC
- Baddeley A, Vedel Jensen EB (2005) *Stereology for statisticians*, Chapman and Hall/CRC 395, Boca Raton London
- Cavalieri B (1635) *Geometria indivisibilibus continuorum*. Bononi: Typis Clementis Ferronij, Reprinted 1966 as *Geometria Degli Indivisibili*, Unione Tipografico-Editrice Torinese, Torino
- Delesse MA (1848) Procédé mécanique pour déterminer la composition des roches. *Annales des Mines* 13, Quatrième série: 379–388
- Dorph-Petersen KA, Gundersen HJG, Jensen EBV (2000) Non-uniform systematic sampling in stereology. *Journal of Microscopy* 200: 148–157
- Frank AU (2008) Analysis of Dependence of Decision Quality on Data Quality. *Journal of Geographical Systems* 10(1): 71–88

- Fujii H, Garsdal H, Ward P, Ishii M, Morishita K, Boivin T (2003) Hydrological roles of the Cambodian floodplain of the Mekong River. *International Journal of River Basin Management* 1(3): 1–14
- Glaser J, Glaser EM (2000) Stereology, morphometry, and mapping: the whole is greater than the sum of its parts. *Journal of Chemical Neuroanatomy* 20: 115–126
- Glasbey C, Horgan R (1995) *Image analysis for the biological sciences*, Wiley, Chichester
- Kötzer S (2006) Geometric identities in stereological particle analysis. *Image Analysis and Stereology* 25(2): 63–74
- Lucieer A, Stein A, Fisher P (2005) Multivariate Texture Segmentation of High-Resolution Remotely Sensed Imagery for Identification of Fuzzy Objects. *International Journal of Remote Sensing* 26: 2917–2936
- Ojala T, Pietikäinen M (1999) Unsupervised texture segmentation using feature distributions. *Pattern Recognition* 32: 477–486
- Pereira dos Santos Silva M, Camara G, Sobral Eescada MI, Modesto de Souza RC (2008) Remote-sensing image mining: detecting agents of land-use change in tropical forest areas. *International Journal of Remote Sensing* 29(16): 4803–4822
- Rajasekar U, Stein A, Bijker W (2006) Image mining for modeling of forest fires from Meteosat images. *IEEE Transactions on Geoscience and Remote Sensing* 45(1): 246–253
- Richards JA, Jia X (1998) *Remote sensing digital image analysis - third edition*, Springer, Berlin Heidelberg New York
- Roberts N, Puddephat MJ, McNulty V (2000) The benefit of stereology for quantitative radiology. *British Journal of Radiology* 73: 679–697
- Shi W (2009) *Principles of modelling uncertainties in spatial data and spatial analysis*, CRC Press, Boca Raton
- Shi W, Fisher PA, Goodchild M (2002) *Spatial Data Quality*, Taylor & Francis, London
- Stein A (2008) Modern developments in image mining. *Science in China Series E: Technological Sciences* 51: 13–25
- Stein A, Shi W, Bijker W (2008) *Quality aspects in spatial data mining*, CRC Press, Boca Raton
- Stein A, Budde P, Yifru MZ (2009) A stereological estimator for the area of flooded land, *International Symposium for Spatial Data Quality*
- Van de Vlag D, Stein A (2006) Uncertainty Propagation in Hierarchical Classification using Fuzzy Decision Trees. *IEEE Transactions on Geoscience and Remote Sensing* 45(1): 237–245
- Vedel Jensen EB (1998) *Local Stereology*. World Scientific Publishing, Singapore
- Yifru MZ (2006) *Stereology for Data Mining*. Unpublished MSc thesis, ITC International Institute for Geoinformation Science and Earth Observation, Enschede, The Netherlands

Modeling Spatiotemporal Paths for Single Moving Objects

Kathleen Stewart Hornsby¹, Naicong Li²

¹Department of Geography, The University of Iowa, Iowa City, IA 52242

²The Redlands Institute, University of Redlands, Redlands, CA 92373
kathleen-stewart@uiowa.edu; naicong_li@spatial.redlands.edu

Abstract

In this work, we focus on modeling paths of movement that an individual moving object follows in space and time. We introduce a set of basic components for paths that serve as the basis for formalizing movement paths. We introduce a typology of paths that describes a classification of paths as open or closed paths. A broader set of path patterns is further investigated by varying temporal granularity between paths traveled on the same day, to paths taken on different days. Distinguishing the different path patterns that are possible for single moving objects provides a basis for searching and retrieving different kinds of spatiotemporal behaviors from collections of moving object data. Based on this work, it is also possible to analyze how patterns of movement may be decomposed to sets of these elemental paths in order to give a clearer understanding of the nature of movement of objects.

1 Introduction

The topic of understanding, modeling, and representing dynamics of geographic domains has been a major focus of research in the field of GIS-science (see, for example, Drummond et al. 2006; Stewart Hornsby and Yuan 2008). At a University Consortium of Geographic Information

Science (UCGIS) meeting on computation and visualization for understanding dynamics in geographic domains held in 2006, a series of research challenges were identified by participants at the workshop (http://www.ucgis.org/dynamics_workshop/). The list of topics included:

- data modeling for dynamic geographic domains,
- computation requirements for dynamics,
- visualization for geographic dynamics,
- spatiotemporal knowledge discovery,
- geographic dynamics over multiple granularities,
- spatiotemporal uncertainty and accuracy,
- dynamic social networks, and
- feature extraction and analysis of images, video, and other unstructured dynamic information sources (Yuan and Stewart Hornsby 2007).

Many of these topics are broad and include numerous subtopics that are still open for investigation. In this paper, the focus is on modeling paths of movement that an individual moving object follows in space and time. This topic cuts across several of the above areas, for example, data modeling for dynamic geographic domains, geographic dynamics over multiple granularities, and visualization for geographic dynamics. Geographic domains refer to geographic spaces such as urban or natural areas where interactions and movements between entities foster happenings that are dynamic and commonly result in change of some type.

The automated collection of movement data from mobile devices captures different kinds of spatiotemporal behaviors of individuals. In this paper, we investigate a set of possible movement patterns associated with individual moving objects. We introduce a typology of different kinds of spatiotemporal paths that are associated with a single object. This typology is based on a classification of paths into *open* or *closed* paths, where movement begins at one location and ends at a different location (open path type), or *backtracking* or *looping* paths (closed path types) where the origin and destination locations are the same. Many common movements can be described based on either of these basic path types, or on a combination of types. A closer examination of open paths exposes a set of possible movement or path patterns. This set of path patterns is complementary to the work of Dodge et al. (2008) where a taxonomy of movement patterns is discussed including individual vs. group movements, generic vs. behavioral patterns, and primitive vs. compound patterns. The patterns described in this paper highlight the characteristics of individual movements in more detail based on elements of spatiotemporal paths and by considering different temporal granularities.

The rest of this paper is structured as follows: section 2 examines the topic of spatiotemporal paths and gives examples of path characteristics that researchers have studied. Section 3 introduces a formalization for spatiotemporal paths of single moving objects. A typology of paths is introduced focusing on open and closed paths. Section 4 presents a set of possible path patterns by varying temporal granularity from paths traveled on the same day to paths taken on different days. A summary and discussion of future work are presented in section 5 of the paper.

2 Modeling Spatiotemporal Paths of Moving Objects

Considerable research on modeling moving objects has focused on methods for describing the paths that a moving object follows in space-time (see for example, Forlizzi et al. 2000, Ding and Güting 2004, Du Mouza and Rigaux 2005, Güting and Schneider 2005, Dodge et al. 2008) and yet open research questions persist. A path commonly describes a spatiotemporal ordering of locations encountered by a moving object or event (Stewart Hornsby and Cole 2007). Paths are often associated with graph representations, and also with networks since graphs are frequently viewed as the more general form of a network. A path captures, for example, a sequence of nodes and edges in a geospatial network used to represent the route traversed by a moving vehicle on a road network. Common path operations include shortest path computations where a route is returned based on the shortest possible path through the network (for an overview of related research on shortest path computations, see Shirabe 2005), or computing the simplest path where the complexity of instructions is minimized (Duckham and Kulik 2003, Richter and Duckham 2008). In time geography research, analyses focus on space-time paths in order to understand patterns of peoples' activities along these paths (see for example, Miller 2008; Raubal et al. 2004; Shaw et al. 2008; Kwan and Ren 2008). Paths are also the basis for modeling movements that are not constrained by networks, such as the paths followed by birds and animals (Laube et al. 2005; Laube et al. 2007) or ships on open water (Cole and Hornsby 2005; Stewart Hornsby and Cole 2007). Paths not only model sequences of space-time locations, but they serve as a form of aggregation where individual locations visited by moving objects or events are abstracted into a path. Paths may sometimes be uncertain (Shokri et al. 2006). Using a path as a basic element of movement allow computations to be made that return solutions for queries such as *Where should I next turn?* or *How much farther is it to the Italian restaurant?* (Güting et al. 2000).

As described above, paths can be computed based on any number of attributes, for example, the path that maximizes privacy, the most scenic

path, the fastest path, etc. In this way, different paths of movement suggest different semantics as for example, the semantics associated with the path a ship takes from offshore to its destination in the harbor (i.e., an expected path) (Fig. 1a) that can be contrasted with a path where a vessel returns to the offshore zone without reaching any destination zone such as the ferry landing (i.e., an unexpected path) (Fig. 1b) (Cole and Hornsby 2005).

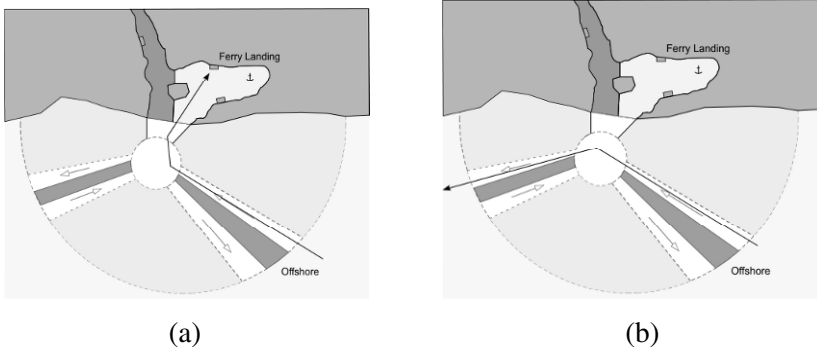


Fig. 1. (a) Expected ship movement from offshore to the ferry landing and (b) an unexpected path where the ship moves away from the harbor without arriving at any destination (after Cole and Hornsby 2005).

Other paths may correspond to more than one moving object and together these paths offer meaningful insights into movement. Such is the case, for example, where multiple ships are moving such that they are shown to converge in the same zone of the harbor, perhaps going to the aid of another ship or multiple ships are leaving an area, perhaps avoiding some event that has occurred (Stewart Hornsby and Cole 2007). In all of these cases, both for the single moving object and for multiple objects, paths can model expected or unexpected deviations in spatiotemporal characteristics of movement or interesting *spatiotemporal behaviors*. In this paper, we examine the topic of spatiotemporal behaviors further by deriving a typology of paths that describes the movements of single moving objects in more detail. To expand on the typology, the temporal granularity is varied from movements on the same day to movements on different days in order to investigate the range of possible path patterns and understand how they vary according to different temporal granularities. Adopting this framework for modeling paths may impose a certain perspective on movement related, for example, to studying movement patterns of people's daily activities. But of course, a modeler could also choose to select a temporal granularity that avoids this type of discretization and corresponds, for example, to discrete timestamps. Here, our interest is in developing a classification of paths and so we explore different temporal choices.

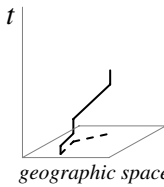
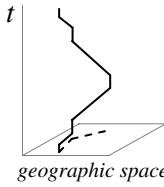
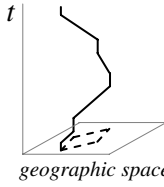
3 Modeling Spatiotemporal Paths of Movement: Open and Closed Paths

To facilitate the discussion of paths, we introduce a formalism where the notation ${}_{id_j}^{date_k}P_i$ is used to represent paths P_i where $i = 0, 1, \dots, n$. The notation id_j where $j = 0, 1, \dots, m$ distinguishes different moving objects that travel along a path on a given $date_k$, where $k = 0, 1, \dots, r$. Note that in this work we distinguish both date and time separately such that it is possible to describe varying temporal semantics including day, week, month, and year (i.e., dates) and also hours, minutes, and seconds (i.e., times). A path can be modeled as a set of locations visited by a moving object over time such that ${}_{id_j}^{date_k}P_i = \{loc_{0,t_0}, loc_{1,t_1}, \dots, loc_{n,t_n}\}$ where $t_0 < t_1 < \dots < t_n$, and $<$ refers to precedence (i.e., $t_0 < t_1$ means that time t_0 is before t_1). In this work, loc_{0,t_0} is the *source* or origin location at time t_0 (start_time) and loc_{n,t_n} is the *destination* location at time t_n (end_time). In addition to the source and destination, a path may have *route* components that model a set of key locations that distinguish the path (i.e., locations where the moving object turned, paused, stopped for a period of time, or was simply being recorded as having been there by, for example, a GPS tracking system). These route elements correspond, for example, to any of the locations visited between locations loc_0 and loc_n such as loc_j . This treatment of route elements may differ from formalizations used for other modeling tasks involving moving objects, but our focus here is on an abstract or higher-level view of paths in order to develop a classification of paths and their patterns of occurrence.

Such a classification begins with paths that correspond to the movement of a single moving object moving from a source location to a (different) destination location (i.e., an *open* path). These paths correspond to movements over a set of locations that range from $0, 1, \dots, n$ where n is the number of locations visited beginning with the source location and ending with the destination location (e.g., Table 1, *P1-I*), and the number of observed time steps that range from $0, 1, \dots, n$ where n corresponds to the last location that is visited, such that $loc_{0,t_0}(P_i) \neq loc_{n,t_n}(P_i)$ (i.e., source and destination of a path are *not* the same) (Table 1, *P1-I*). This common type of path can be contrasted with a *closed* path, the case where a moving object moves some distance from a source location and returns to that same location for its destination, i.e., $loc_0(P_i) = loc_n(P_i)$. These two basic path types have also been distinguished in studies on geospatial lifelines, where, for example, modelers are interested in capturing all the possible locations a moving object may visit when travelling between a source and destination (Hornsby and Egenhofer 2002; Miller 2006). The routes of closed

paths can take different forms resulting in two types of closed paths (Table 1, *PI-2* and *PI-3*). In Table 1, the second column contains graphical representations of these path patterns. The vertical axis t represents the temporal dimension, and the other two axes are the X and Y axes of the geographic space. The solid line in the graph represents the movement path in time and geographic space, and the dashed line represents the projection of this path on the surface of the geographic space. A vertical segment in the graph that projects to a point on the geographic space represents the elapsed time that the moving object spends at a stopping point (i.e., route element). In the second case, the route out and back is the same (capturing the semantics of backtracking), i.e., for this case, locations range from $0, 1, \dots, n$ such that loc_0 is the source location, $loc_{n/2}$ models the location that represents the farthest distance travelled before turning back, and loc_n models the location of the destination that is the same as the origin (e.g., path *PI-2*). For this case, the temporal properties of the path has times that range from $0, 1, \dots, n$, i.e., $t_{n/2}$ models the time that corresponds to the location of the farthest point visited before turning back on the route and completing the movement at time t_n . The third case of closed paths is where the route out and back are different to each other, forming a closed loop that ends back at the source location (e.g., *PI-3*). For this case, the source location is the same as the destination location (i.e., $loc_0 = loc_n$) although locations in between are different from each other (i.e., $loc_1 \neq loc_2 \neq \dots \neq loc_{n-1}$) and the times associated with the path vary from $0, 1, \dots, n$. Also for this case, there is usually no need to distinguish the point of maximal travel, unlike the backtracking case where it is useful to distinguish the point where the moving object turns back.

Table 1. Patterns of paths for single moving objects

Pattern		Spatial properties	Temporal properties
P1-1		$loc_{0,\dots,n}$ where $loc_0 \neq loc_1 \neq \dots \neq loc_n$	$t_{0,\dots,n}$
P1-2		$loc_{0,\dots,n}$ where $loc_0(P_i) = loc_n(P_i)$ and $loc_x = loc_{n-x}$ where $x=1, \dots, n-1$ and $loc_{n/2}$ is the last location visited in the set of loca- tions before turning back	$t_{0,\dots,n}$
P1-3		$loc_{0,\dots,n}$ where $loc_0(P_i) = loc_n(P_i)$ and $loc_1 \neq loc_2 \neq \dots \neq loc_{n-1}$	$t_{0,\dots,n}$

Given these two basic types of paths, open and closed, it is possible to describe paths that are based on a combination of both these path types. There may be combinations that are *open-open*, *open-closed*, *closed-open*, and *closed-closed*, as well as sequences of these types (e.g., *open-closed-open*). The first combination, *open-open*, refers to a single path that is combined with another single path. This is a common setting for moving objects and could describe pairings of paths that a single moving object follows. In this case, spatial granularity or level of detail is obviously an important consideration since an *open-open* combination emphasizes the path characteristics as being composed of two distinct paths where the destination of the first path equals the source of the subsequent path.

Given this understanding about granularity, the *open-closed* combination could describe the case where a person drives to a trailhead and then hikes a loop trail that end up back at the trailhead. A *closed-open* combination describes a drive around a loop road in a park followed by a drive to a restaurant. *Closed-closed* combinations model successive loops such as made by vehicle circling the block looking at a house for sale (and possi-

bly doing a different sized or shaped loop each time), searching for a parking space by circling the block, or could also describe separate repeated trips up and down a street. For example, within one day, sets of paths may correspond to $P1-3$ and $P1-2$ and a combination of $P1-1$ and $P1-3$ (Fig. 2). The role of time is important for distinguishing these different semantics and capturing the notion of *repeated* (i.e., same path at different times or dates) or *successive* trips (different paths that follow on from each other, e.g., in the course of one day).



Fig. 2. Paths of a single object over one day. One path corresponds to a closed path (P1-3), another path corresponds to an open-closed combination (P1-1 and P1-3) and a third path (P1-2) corresponds to a closed path with backtracking.

From the basic path types, open or closed, the classification can be extended by considering the set of possible open paths that occur when the start and end times of the paths are different (but, for example, the date is held constant). Since the source and destination locations of a closed path are the same, the focus here is on a classification of different kinds of open path patterns. A set of path patterns based on the components of *source*, *destination*, and *route* for different combinations of start and end times, exposes a number of spatiotemporal path characteristics. These different patterns have both spatial *and* temporal properties that set one pattern apart from another. In this work, we focus particularly on patterns that correspond to the movements of individual objects. The movements of groups of objects are not considered further in this study. In this way, we distinguish the range of path patterns possible for single moving objects and show how these paths change when the spatiotemporal parameters vary.

4 Possible Path Patterns for Single Moving Objects

We investigate possible patterns of paths by looking at paths over different temporal granularities, within the same day and over different days (and the temporal granularity can be altered here to other date types (e.g., week, month, year)). This allows for a wider range of assumptions relating to start and end times. For example, for the first set of paths, movements are assumed to occur during the course of one day (i.e., paths are for the same day, $date(P_i) = date(P_j)$), although not *all* the movements an object makes during a day are assumed. Grey and black paths represent different paths of the same moving object on the same granularity of date (e.g., day) (Table 2). For the graphics in Table 2 (and in subsequent tables), the vertical axis represents time (t). A spatiotemporal path is represented symbolically using only three segments, with the starting and ending points of the middle segment as an extremely simplified representation of the route of the path. Since for this group of paths, all movements are assumed to happen on the same date, it is understood that the start times (and end times) of different paths must be distinct from each other and path P_i is before path P_2 (i.e., $t_n(P_i) < t_o(P_2)$). One pattern of paths is where the source, destination, and routes are the same ($P2-1$), for example, a parent driving to a child's school using the same route each time at different times during the day. Another case is where the source and destination are the same, but the routes taken are different (e.g., travel to a child's school more than once in a day but taking different routes to get there each time) ($P2-5$). Another pattern, perhaps less common, has an object moving at different times such that the source and destination locations of the paths are different, but the route is common for both paths ($P2-4$). A different case is where the source is common for different paths, but the destination and route are changed ($P2-7$). This could correspond to paths taken to fulfill different errands, where the distinct destination for each trip causes a different route to be chosen. Path pattern $P2-8$ describes the case where the individual movements are independent, i.e., source, destination, and routes vary for each path. This set of eight path patterns represent a basic set of movements for single moving objects and our studies reveal that many of these patterns are prototypical for movements at other temporal granularities.

If we examine the set of path types that correspond to cases of a single moving object over *different dates* with the same or different route, additional patterns are revealed. Investigating paths over different dates now

Table 2. Open paths for a single moving object on same date with different start and end times.

No.	Spatial properties	
P2-1	$loc_0(P_1) = loc_0(P_2)$ $loc_n(P_1) = loc_n(P_2)$ $loc_{1,\dots,n-1}(P_1) = loc_{1,\dots,n-1}(P_2)$	
P2-2	$loc_0(P_1) \neq loc_0(P_2)$ $loc_n(P_1) = loc_n(P_2)$ $loc_{1,\dots,n-1}(P_1) = loc_{1,\dots,n-1}(P_2)$	
P2-3	$loc_0(P_1) = loc_0(P_2)$ $loc_n(P_1) \neq loc_n(P_2)$ $loc_{1,\dots,n-1}(P_1) = loc_{1,\dots,n-1}(P_2)$	
P2-4	$loc_0(P_1) \neq loc_0(P_2)$ $loc_n(P_1) \neq loc_n(P_2)$ $loc_{1,\dots,n-1}(P_1) = loc_{1,\dots,n-1}(P_2)$	
P2-5	$loc_0(P_1) = loc_0(P_2)$ $loc_n(P_1) = loc_n(P_2)$ $loc_{1,\dots,n-1}(P_1) \neq loc_{1,\dots,n-1}(P_2)$	
P2-6	$loc_0(P_1) \neq loc_0(P_2)$ $loc_n(P_1) = loc_n(P_2)$ $loc_{1,\dots,n-1}(P_1) \neq loc_{1,\dots,n-1}(P_2)$	
P2-7	$loc_0(P_1) = loc_0(P_2)$ $loc_n(P_1) \neq loc_n(P_2)$ $loc_{1,\dots,n-1}(P_1) \neq loc_{1,\dots,n-1}(P_2)$	
P2-8	$loc_0(P_1) \neq loc_0(P_2)$ $loc_n(P_1) \neq loc_n(P_2)$ $loc_{1,\dots,n-1}(P_1) \neq loc_{1,\dots,n-1}(P_2)$	

allows for start times (or end times) to be the same, unlike the cases shown in Table 2. The first set of paths that correspond to a single object moving

on different dates describes paths where the paths have the same source and same destination and occur along the same or different route (Table 3, where black and grey paths represent paths taken by the same object over the same temporal period on *different* dates). Some of these paths evoke spatiotemporal behaviors that are common or even routine, and are elemental for a population of moving objects over time. Patterns *P3-1*, *P3-2*, and *P3-3* describe cases where the same path is taken on different days but with varying start and end times. *P3-1* where the two paths completely overlap each other, captures routine movements, for example, traveling to work or school where movement is from the same source to the same destination along the same route at the same times each day. *P3-2* involves different start times (leaving from a source earlier or later), while *P3-3* involves different end times (getting to a destination earlier or later). Patterns *P3-4*, *P3-5*, and *P3-6* describe cases where the path of movement on different days occurs along *different routes* and over varying start and end times. Perhaps some local factor (e.g., construction or a festival) requires a different route to be used by the moving object on different days. These combinations of paths are common to many types of moving objects. *P3-4* describes the case where start and end times are the same each day. Path patterns *P3-5* and *P3-6* capture cases where the start times and end times respectively are different. It should be noted that combinations of paths where $t_0(P_1) \neq t_0(P_2)$ and $t_n(P_1) \neq t_n(P_2)$ over different dates are not presented in Table 3 as they are not significantly different to the patterns *P2-1* and *P2-5* already identified in Table 2. *P2-1* and *P2-5* are basic path patterns that hold over multiple temporal granularities.

When paths have different sources, destinations, or both, a new set of possible patterns emerges (Table 4). Path patterns *P4-1* through *P4-9* describe cases where the source, destination, or both the source and destination vary, but follow a common route (i.e., $loc_{1,\dots,n-1}(P_1) = loc_{1,\dots,n-1}(P_2)$). The temporal characteristics of these combinations of paths vary for each case. This set of patterns may not be as common as other path types for certain domains. Pattern *P4-4*, for example describes the case where the sources and routes are the same, as are the starting and ending times for each path, but the destinations are different. *P4-5* allows for different starting times, while *P4-6* captures different ending times. Again, for this set of path combinations, the complete set of possible cases includes three cases that are omitted from Table 4 as they are not significantly different from *P2-2*, *P2-3*, and *P2-4* (although they occur over different dates), and have already been presented as part of Table 2. These cases have the characteristics, $t_0(P_1) \neq t_0(P_2)$ and $t_n(P_1) \neq t_n(P_2)$.

Table 3. Open paths for a single moving object on different dates with same sources and destinations

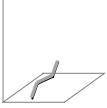
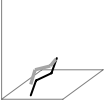
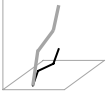
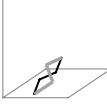
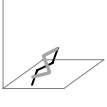
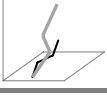
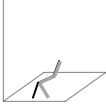
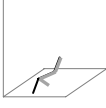
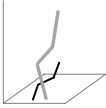
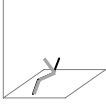
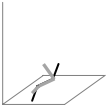
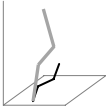
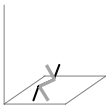
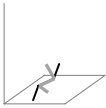

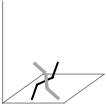
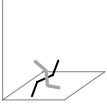
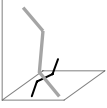
No.	Spatial properties	Temporal properties	
P3-1	$loc_{0,\dots,n}(P_1) = loc_{0,\dots,n}(P_2)$	$t_0(P_1) = t_0(P_2)$ $t_n(P_1) = t_n(P_2)$	
P3-2	$loc_{0,\dots,n}(P_1) = loc_{0,\dots,n}(P_2)$	$t_0(P_1) \neq t_0(P_2)$ $t_n(P_1) = t_n(P_2)$	
P3-3	$loc_{0,\dots,n}(P_1) = loc_{0,\dots,n}(P_2)$	$t_0(P_1) = t_0(P_2)$ $t_n(P_1) \neq t_n(P_2)$	
P3-4	$loc_0(P_1) = loc_0(P_2)$ $loc_n(P_1) = loc_n(P_2)$ $loc_{1,\dots,n-1}(P_1) \neq loc_{1,\dots,n-1}(P_2)$	$t_0(P_1) = t_0(P_2)$ $t_n(P_1) = t_n(P_2)$	
P3-5	$loc_0(P_1) = loc_0(P_2)$ $loc_n(P_1) = loc_n(P_2)$ $loc_{1,\dots,n-1}(P_1) \neq loc_{1,\dots,n-1}(P_2)$	$t_0(P_1) \neq t_0(P_2)$ $t_n(P_1) = t_n(P_2)$	
P3-6	$loc_0(P_1) = loc_0(P_2)$ $loc_n(P_1) = loc_n(P_2)$ $loc_{1,\dots,n-1}(P_1) \neq loc_{1,\dots,n-1}(P_2)$	$t_0(P_1) = t_0(P_2)$ $t_n(P_1) \neq t_n(P_2)$	

Table 4. Open paths for a single moving object following the same routes on different dates with different sources and destinations

No.	Spatial properties	Temporal properties	
P4-1	$loc_0(P_1) \neq loc_0(P_2)$ $loc_n(P_1) = loc_n(P_2)$	$t_0(P_1) = t_0(P_2)$ $t_n(P_1) = t_n(P_2)$	
P4-2	$loc_0(P_1) \neq loc_0(P_2)$ $loc_n(P_1) = loc_n(P_2)$	$t_0(P_1) \neq t_0(P_2)$ $t_n(P_1) = t_n(P_2)$	
P4-3	$loc_0(P_1) \neq loc_0(P_2)$ $loc_n(P_1) = loc_n(P_2)$	$t_0(P_1) = t_0(P_2)$ $t_n(P_1) \neq t_n(P_2)$	
4-4	$loc_0(P_1) = loc_0(P_2)$ $loc_n(P_1) \neq loc_n(P_2)$	$t_0(P_1) = t_0(P_2)$ $t_n(P_1) = t_n(P_2)$	
P4-5	$loc_0(P_1) = loc_0(P_2)$ $loc_n(P_1) \neq loc_n(P_2)$	$t_0(P_1) \neq t_0(P_2)$ $t_n(P_1) = t_n(P_2)$	
P4-6	$loc_0(P_1) = loc_0(P_2)$ $loc_n(P_1) \neq loc_n(P_2)$	$t_0(P_1) = t_0(P_2)$ $t_n(P_1) \neq t_n(P_2)$	
P4-7	$loc_0(P_1) \neq loc_0(P_2)$ $loc_n(P_1) \neq loc_n(P_2)$	$t_0(P_1) = t_0(P_2)$ $t_n(P_1) = t_n(P_2)$	
P4-8	$loc_0(P_1) \neq loc_0(P_2)$ $loc_n(P_1) \neq loc_n(P_2)$	$t_0(P_1) \neq t_0(P_2)$ $t_n(P_1) = t_n(P_2)$	
P4-9	$loc_0(P_1) \neq loc_0(P_2)$ $loc_n(P_1) \neq loc_n(P_2)$	$t_0(P_1) = t_0(P_2)$ $t_n(P_1) \neq t_n(P_2)$	

Another set of possible paths highlights independent movements of a single object in more detail. For these cases sources, destinations, and routes are all different (i.e., $loc_{0,\dots,n}(P_1) \neq loc_{0,\dots,n}(P_2)$) (Table 5). Pattern *P5-1* describes the case where independent movement occurs, but the start and end times are the same. Other path types have different start time (*P5-2*) or end times (*P5-3*). For these cases, therefore, the temporal characteristics, rather than the spatial characteristics become the basis for comparison.

Table 5. Open paths for a single moving object on different dates with different sources, different destinations, and different routes.

No.	Spatial properties	Temporal properties	
5-1	$loc_{0,\dots,n}(P_1) \neq loc_{0,\dots,n}(P_2)$	$t_0(P_1) = t_0(P_2)$ $t_n(P_1) = t_n(P_2)$	
5-2	$loc_{0,\dots,n}(P_1) \neq loc_{0,\dots,n}(P_2)$	$t_0(P_1) \neq t_0(P_2)$ $t_n(P_1) = t_n(P_2)$	
5-3	$loc_{0,\dots,n}(P_1) \neq loc_{0,\dots,n}(P_2)$	$t_0(P_1) = t_0(P_2)$ $t_n(P_1) \neq t_n(P_2)$	

5 Summary and Future Work

Distinguishing the different path types that are possible for single moving objects provides a basis for searching and retrieving different spatiotemporal behaviors from collections of moving object data. If these patterns are considered as basic movements (i.e., movements common to a wide range of moving objects), it is also possible to analyze how patterns of movement may be decomposed to sets of these elemental paths to give a clearer understanding of possible movements. This paper presents a basic set of path types, open and closed, where open paths capture movements where the destination and origin are different from one another. Closed paths refer to cases where the origin and destination are the same, common in round trips or looping movements. Although many movements can be characterized by a combination of open and closed paths, this study examines cases of open paths further to expose a larger set of possible patterns of paths for single moving objects. Paths may be described as they occur over a single day or over different days. Examining paths over

different days allows the start and end time of the paths to be the same (not plausible for paths on the same day). For example, paths may be taken at the same time on different days, or paths may follow a different route to the same destination on different days. Certain paths may appear to be characteristic of a moving object over a time period and then movement may change, evoking a different path type. In this way, the framework can serve as a basis for searching for sudden or gradual changes in paths and movement patterns. Further work will focus on the formalizations necessary for modeling combinations of open and closed paths, as well as studying additional parameters for paths that support more semantics, for example, distance and speed, where shifting between shorter or longer paths may suggest new spatiotemporal behaviors.

Acknowledgments

Kathleen Stewart Hornsby's research is supported in part by grants from the U.S. Department of Defense HM1582-08-2001, HM1582-05-1-2039 and HM1582-08-1-0013. Naicong Li's research is supported in part by the Army Research Office under contract number W911NF-07-1-0392.

References

- Cole S, Hornsby K (2005) Modeling noteworthy events in a geospatial domain. In: Rodriguez MA, Cruz I, Egenhofer MJ, Levashkin S (eds) Proceedings of the First International Conference on Geospatial Semantics, GeoS 2005, Lecture Notes in Computer Science, 3799, Springer, Berlin Heidelberg New York, pp 77–89
- Ding Z, Güting R (2004) Managing moving objects on dynamic transportation networks. In: Proceedings of the 16th International Conference on Scientific and Statistical Database Management, Fernuniversitat Hagen, Germany, pp 287–296
- Dodge S, Weibel R, Lautenschütz A-K (2008) Towards a taxonomy of movement patterns. *Information Visualization* 7: 240–252
- Drummond J, Billen R, Joao E, Forrest D (eds) (2006) Dynamic and mobile GIS: investigating changes in space and time, CRC Press, Boca Rotan
- Du Mouza C, Rigaux P (2005) Mobility patterns. *GeoInformatica* 9(4): 297–319
- Duckham M, Kulik L (2003) Simplest paths: Automated route selection for navigation. In: Kuhn W, Worboys M, Timpf S (eds), Proceedings of COSIT 2003, Lecture Notes in Computer Science, 2825, Springer, Berlin Heidelberg New York, pp 169–185
- Forlizzi L, Güting R, Nardelli E, Schneider M (2000) A data model and data structures for moving objects databases. In: Proceedings of the 2000 ACM

- SIGMOD international Conference on Management of Data, Dallas, TX USA, pp 319–330
- Güting R, Schneider M (2005) Moving objects databases, Morgan Kaufmann Publishers
- Güting R, Böhlen M, Erwig M, Jensen C, Lorentzos N, Schneider M, Vazirgiannis M (2000) A foundation for representing and querying moving objects. *ACM Transactions on Database Systems* 25: 1–42
- Hornsby K, Egenhofer MJ (2002) Modeling moving objects over multiple granularities. *Ann. Math. Artif. Intell.* 36(1-2): 177–194
- Kwan M-P, Ren F (2008) Analysis of human space-time behavior: Geovisualization and geocomputational approaches. In: Stewart Hornsby K, Yuan M (eds) *Understanding Dynamics of Geographic Domains*, CRC Press, New York, pp 93–113
- Laube P, Imfeld S (2002) Analyzing relative motion within groups of trackable moving point objects. In: Egenhofer MJ, Mark D (eds), *Proceedings of GIScience 2002, Lecture Notes in Computer Science*, 2478, Springer, Berlin Heidelberg New York, pp 132–144
- Laube P, Imfeld S, Weibel R (2005) Discovering relative motion patterns in groups of moving point objects. *International Journal of Geographical Information Science (IJGIS)* 19(6): 639–668
- Laube P, Dennis T, Forer P, Walker M (2007) Movement beyond the snapshot - dynamic analysis of geospatial lifelines. *Computers, Environment, and Urban Systems* 31(5): 481–501
- Miller HJ (2006) Modeling accessibility using space-time prism concepts with geographical information systems: Fourteen years on. In: Fisher P (ed.) *Classics from IJGIS*, Taylor and Francis, pp 175–179
- Miller HJ (2008) Time geography. In: Shekhar S, Xiong H (eds.), *Encyclopedia of GIS*, Springer, Berlin Heidelberg New York
- Raubal M, Miller H, Bridwell S (2004) User-centered time geography for location-based services. *Geografiska Annaler-B* 86: 245–265
- Richter K-F, Duckham M (2008) Simplest instructions: finding easy-to-describe routes for navigation. In: Cova T, Miller H, Beard K, Frank AU, Goodchild MF (eds) *Proceedings of GIScience 2008, Lecture Notes in Computer Science*, 5266, Springer, Berlin Heidelberg New York, pp 274–289
- Shaw S-L, Bombom L, Yu H (2008) A space-time GIS approach to exploring large individual-based spatiotemporal datasets. *Transactions in GIS* 12(4): 425–441
- Shirabe T (2005) Shortest path search from a physical perspective. In: *Proceedings of COSIT 2005*, Buffalo, NY, pp 83–95
- Shokri T, Delavar M, Malek M, Frank AU, Navratil G (2006) 3D modeling moving objects under uncertainty conditions. In: Abdul-Rahman A, Zlatanova S, Coors V (eds), *Innovations in 3D Geo Information Systems*, Kuala Lumpur, Malaysia, pp 138–149
- Stewart Hornsby K, Cole S (2007) Modeling moving geospatial objects from an event-based perspective. *Transactions in GIS* 11(4): 555–573
- Stewart Hornsby K, Yuan M (eds) (2008) *Understanding Dynamics of Geographic Domains*, CRC Press, New York, NY

Yuan M, Stewart Hornsby K (2007) *Computation and Visualization for Understanding Dynamics in Geographic Domains: A Research Agenda*, CRC Press, New York, NY

Moving Objects in Databases and GIS: State-of-the-Art and Open Problems

Markus Schneider

University of Florida, Department of Computer & Information Science &
Engineering
Gainesville, FL 32611, USA
mschneid@cise.ufl.edu

1 Introduction

The field of *moving objects databases* (Güting and Schneider 2005) has received a lot of research interest in recent years. This technology enables the user to model, store, retrieve, and query the movements of spatial objects over time, called *moving objects*, and to ask queries about such movements in a database context. A moving object represents the *continuous* evolution of a spatial object over time. In some cases, only the time-dependent locations are of interest, and we speak of *moving points*. Examples are mobile phone users, whales, ships, planes, terrorists, cars, spacecrafts, satellites, and missiles. In other cases, also the time-dependent shape and/or areal extent, which can grow or shrink, need to be handled, and we speak of *moving regions*. Examples are hurricanes, lakes, forest fires, oil spills, and the spread of diseases. In some rarer cases, the time-dependent shape and/or linear extent, which can lengthen or shorten, is of interest, and we speak of *moving lines*. Examples are snakes, the slowly retreating front of an army or a glacier, and the boundaries of moving regions in general. Much interest in moving objects databases has been spurred by current trends in consumer electronics, wireless communications, positioning technologies, and location-based services (Schiller and Voisard 2004). Corresponding hardware like wireless networking enabled and position-aware (i.e., GPS equipped) devices such as PDAs, on-board

units in vehicles, sensors, or even special mobile phones have become affordable and will be in widespread use in the near future and trigger many new kinds of mobile, geographical applications. These applications will produce a huge volume of movement information that has to be managed and analyzed in database systems and be made available for spatiotemporal analysis in Geographic Information Systems (GIS). Unfortunately, current database technology and GIS technology are far away from being able to perform this task and thus require new data management and processing concepts and techniques.

The goal of this paper is to give an overview of the current state-of-the-art of moving objects databases and, in particular, to identify open research problems and indicate possible solutions.

Section 2 deals with moving objects in unconstrained environments. These are spatial objects that can freely change their location, shape, and extent. Section 3 describes moving objects in constrained environments. These are spatial objects whose temporal evolution is bounded due to spatial limitations like networks or labyrinths. In Section 4, we finally draw some conclusions.

2 Moving Objects in Unconstrained Environments

In the spatiotemporal database community, two main modeling paradigms have been proposed to characterize the movement of spatial objects. The first paradigm supports an *orthogonal* view. The concept is to handle multidimensional issues by decomposing them into different facets so that each facet (dimension) can be handled independently from the other. The benefit is that understanding and solving one facet at a time is much simpler than solving all facets together. For spatiotemporal issues this means that time and space are considered as separate facets and modeled as a part of type $time \times space$. An approach supporting this paradigm is *MADS (Modeling Application Data with Spatio-temporal Features)* (Parent et al. 2006). The second paradigm supports a uniform view and emphasizes the functional dependence between time and space in spatiotemporal issues. This means that time and space are handled in an integrated way and modeled in terms of functions of type $time \rightarrow space$ as a particular subtype of $time \times space$ with special features like continuity. The author has contributed to and advocates an approach supporting this paradigm (Güting and Schneider 2005) since movement denotes the evolution of spatial objects over time.

Moving objects in unconstrained environments are either not impeded by spatial constraints (for example, the extension of hurricanes), or we are not

interested in their spatial constraints (for example, the route of whales in oceans). The research literature has made a separation of this kind of moving objects into historical moving objects (Section 2.1) and predictive moving objects (Section 2.2).

2.1 Historical Moving Objects

Historical moving objects describe the temporal evolution of spatial objects *in the past* and are leveraged for spatiotemporal analysis. We first give an overview of available approaches and then identify some open research problems.

2.1.1 Overview

A widely accepted approach to modeling historical moving objects in databases introduces the fundamental concept of *spatiotemporal data types* (Erwig et al. 1999a). These data types enable the user to describe the continuous, dynamic, and time-dependent behavior and location change of spatial objects over time and to perform spatiotemporal analysis. That is, the spatial objects move, and they are therefore called *moving objects*. They are stored in special spatiotemporal databases called *moving objects databases* (Güting and Schneider 2005) and are designed as abstract data types, embedded as attribute types into a DBMS data model (relational, object-oriented, etc.), and equipped with a comprehensive collection of operations and predicates. Spatiotemporal data types are available for *moving points* (type *mpoint*), *moving lines* (type *mline*), and *moving regions* (type *mregion*). In case of moving regions, besides the movement aspect, one can also represent the change of their extent and shape over time. Conceptually, a moving point is a function $f: \text{time} \rightarrow \text{point}$, a moving line is a function $f: \text{time} \rightarrow \text{line}$, and a moving region is a function $f: \text{time} \rightarrow \text{region}$. For example, for a moving region this means that at each time instant an object of type region has to be returned. The general ideas of this model have been presented in (Erwig et al. 1998, 1999a, 1999b). A formal specification of the spatiotemporal data types and operations has been given in (Güting et al. 2000).

Spatiotemporal predicates (Güting and Schneider 2005) describe changing topological relationships of moving objects over time. In the same way as spatial objects can change over time, the topological relationships between them can change over time. An example of such a predicate is the term *Disjoint* \gg *meet* \gg *Inside* \gg *meet* \gg *Disjoint* that is composed of a *temporal sequence* of the basic spatio-temporal predicates *Disjoint* and *Inside* as well as the topological predicate *meet*. The *temporal composition*

operator is indicated by the symbol \gg . This predicate could, for example, ask for a spatiotemporal pattern whether a plane is disjoint from a hurricane for some period, then meets the boundary of the hurricane at a time instant, is inside the hurricane for some period, meets the boundary of the hurricane again at a time instant, and is disjoint again from the hurricane for some period. The alternating sequence of topological predicates that hold for some period or for some time instant is characteristic for composite spatiotemporal predicates. A *spatiotemporal query language* called *STQL* has been introduced in (Erwig and Schneider 1999), and a *visual query language* called *Query-By-Trace* has been proposed in (Erwig and Schneider 2000, 2003).

Regarding implementation, data structures for moving objects have been presented in (Forlizzi et al. 2000) and algorithms for spatiotemporal operations have been designed in (Cotelo Lema et al. 2003). The author has co-authored a book titled *Moving Objects Databases* (Güting and Schneider 2005) covering all aforementioned topics.

Since then, the field has blossomed out, and much work has been done especially on implementation issues, e.g., on developing index structures (Aragwal et al. 2000; Hadjieleftheriou et al. 2002; Kollios et al. 1999; Pfooser et al. 2000), processing continuous queries (Song and Roussopoulos 2001; Tao and Papadias 2003), studying similarity of trajectories (Yanagisawa et al. 2003), developing test data generators (Theodoridis et al. 1999), to name only some of the areas.

2.1.2 Open Research Problems

From a modeling perspective, the author sees a general need to perform research on *spatiotemporal predicates* as the temporally lifted versions of spatial predicates. In a spatial database context, these predicates are needed as filter conditions for spatiotemporal joins and selections. While a model for the evolution of *topological relationships* over time is available (Erwig and Schneider 2002), models concerning the evolution of *directional relationships* (*cardinal directions*) over time have not been proposed. One reason might be that the purely spatial problem of modeling directional relationships in the two-dimensional space and the three-dimensional space has not been solved satisfactorily so far. In a spatiotemporal context, the problem is to detect directional patterns of moving objects. For example, a group of whales could have moved from location X in northwestern direction for a while, then turned to the south, then moved to the east, and finally moved to the north. It seems that from this directional information we can neither derive the overall cardinal direction seen from the starting point of the route nor the location of the whales.

From an implementation perspective, further work is needed to develop appropriate database-enabled data structures and algorithms for moving objects. Efficient algorithms are needed for spatiotemporal predicates of all kinds. For example, it is unclear how topological predicates over time can be evaluated (efficiently). Traditionally, work is needed that focuses on spatiotemporal index structures.

2.2 Predictive Moving Objects

Predictive moving objects describe the predicted temporal evolution of spatial objects *at the present time* and *in the near future*. We first give an overview of available approaches and then identify some open research problems.

2.2.1 Overview

The data model *MOST* (*Moving Objects Spatio-Temporal model*) (Sistla et al. 1997, 1998) is the only model so far that is able to describe current and expected future movement in a database context. All known application-specific models are independent of a database context. MOST enables the user to keep track of a set of time-dependent locations (i.e., moving points like mobile phone users) in a database in terms of *location-based management*. The model is based on the observation that one should not keep the positions directly in the database, leading to a very high volume of updates, but rather represent them by a *motion vector*. Only when the object's position predicted by the motion vector deviates from the real position by more than a threshold, an update needs to be transmitted to the database.

The fundamental idea is to introduce so-called *dynamic attributes* which change their values automatically with time. Not all attribute types are eligible to be dynamic. Such a type must have a value 0 and an addition operation. The dynamics is given by linear functions that describe *motion vectors* and avoid frequent database updates. Examples are the types *dynamic integer* and *dynamic real*. Unfortunately, there is no concept of dynamic spatial data types so that the only option to represent a “moving point” is to model it as a pair (x : *dynamic real*, y : *dynamic real*). Dynamic lines or regions cannot be modeled, and there is no concept of spatiotemporal data types available. If a query refers to a dynamic attribute A, its dynamic value is meant and used in the evaluation. Hence, the result depends on the time when the query is issued. If such a query is reevaluated on each clock tick, this query is called *continuous*.

A query language called *Future Temporal Logic (FTL)* (Ststla and Wolfson 1995; Sistla et al. 1997, 1998) allows one to express conditions

about the future and specify temporal relationships between objects in queries. It especially introduces the temporal modal operators *until* and *nexttime* from which a collection of other temporal operators can be derived. This approach is restricted to moving points. An implementation of this model has been provided in a prototype called DOMINO and described in (Wolfson et al. 1998a, 1998b, 1999).

Another model uses cylindrical volumes to represent the uncertainty of the *trajectory* of a moving object (Trajcevski et al. 2002, 2004). Uncertainty is an inherent feature of current and near future locations of moving objects. This model takes into account the *temporal uncertainty* and the *spatial uncertainty* of objects. One can then ask for the objects that are inside a query region *sometimes* or *always* during a time interval (temporal uncertainty). Similarly, one may ask for the objects that are *possibly* or *definitely* inside a query region. A combination of both uncertainty aspects leads to a particular kind of *spatiotemporal predicates* like *PossiblySometimeInside* or *AlwaysDefinitelyInside*. Related work on uncertain trajectories includes (Mokhtar et al. 2002, Pfoser and Jensen 1999).

2.2.2 Open Research Problems

Despite some progress, research on predictive moving objects is still restricted. Current models do only refer to moving points which are modeled implicitly through dynamic attributes. Hence, the first research issue relates to the design and rigorous definition of a comprehensive data model for the predictive and near future evolution of moving objects. A requirement is that this data model supports a data type view on predictive moving objects, that is, predictive spatiotemporal data types, so that these data types can be later integrated into databases. A main challenge is the treatment of the inherent uncertainty that affects near future moving objects. A second issue refers to the design of spatiotemporal predicates for predictive moving objects. The main question here is to explore how moving objects with an individual, predicted behavior may behave towards each other. Such a behavior may comprise topological, directional, and distance relationships between these objects. Again, a challenge is the treatment of the uncertainty that also affects these relationships. A third issue is the design and implementation of effective and versatile data structures for predictive spatiotemporal data types and efficient algorithms for their operations and predicates. A fourth issue is the embedding of these novel concepts into a database query language and its integration into a database system. A fifth issue refers to a seamless integration of the data models and query languages for historical and predictive moving objects into database systems. The ultimate goal must be to obtain a data model that is capable of modeling historical *and* predictive movement in a single, seam-

less, consistent, and homogeneous framework. This will lead to a much more complete and powerful data model and allow queries that span the past and the future. An example is the query where a hurricane has been located since yesterday and where it will be in two days. Since all concepts are intended for a use in a database context, the issue is how the concepts for historical and predictive moving objects can be smoothly integrated into a database query language like SQL.

The author has started research on these issues which is led by the statement that it is *not* the task of a database system to predict the future movement of a moving object. The reason is that prediction models are tailored to specific application domains and based on external factors or domain-specific parameters which may significantly affect the future movements of moving objects. For instance, information such as atmospheric pressures, temperature zones, and wind and ocean currents plays a major role in predicting the future evolution of a hurricane. This requires highly specialized and sophisticated prediction models and algorithms beyond those in which only the past and current object movements are considered as system parameters. In fact, the development effort of such prediction methods is a whole discipline by itself, and this task belongs to domain experts, in this case meteorologists. Further, prediction models for different application domains are usually different. Therefore, a moving objects database system as a general-purpose tool should provide application-neutral modeling and persistence support for storing and retrieving predicted moving objects data and offer querying capability for retrieving such data. This means that it is solely reasonable to perform a separation between moving object models and prediction models with respect to representing future evolutions of moving objects. Since this separation has not been realized so far in existing future movement models, each of these models has only dealt with a specific problem area or object motion type while neglecting the problem of how a moving object database can model the future movements of moving objects in general.

The author proposes a novel algebra or type system called *Moving Balloon Algebra* for moving objects in unconstrained environments. So-called *balloon data types* enable the representation of both *historical*, *predictive*, and *time-spanning* (= historical + predictive) moving objects. The term *balloon* is used as a metaphor to model the nature of a moving object evolving in the past *and* extending to the future for a given current time t^{now} . The *string* and the *body* of a *balloon object* represent its past and future movement respectively. For example, the past, known movement and the future, predicted movement of the eye of a hurricane are usually illustrated by using a shape that resembles a balloon. The past movement of the eye (a moving point) can be seen over time as the movement along a line or a curve which resembles the string of a balloon. On the other hand, the

position of the eye at a time instant in the future can be anywhere within an area of uncertainty. Thus, the future movement of the eye can be seen as a moving region of uncertainty which resembles the body of a balloon. Furthermore, the connection point between the string and the body of a balloon object represents the present state of the moving object at t^{now} . From a data type perspective and ignoring the uncertainty aspect for a moment, we can describe the string by the spatiotemporal data type *mpoint* and the body by the type *mregion*. But this is not the only possible type combination. For example, assume a car driving on a road network has taken a particular route so far. At a junction where it is now at time t^{now} , it has several options since several roads may emanate from this junction. We assume that it is not exactly known which route the car will take. This situation can be modeled by a single-component *mpoint* object for the past and a multiple-component *mpoints* object for the possible future routes. Note the difference between the names and the semantics of the types *mpoint* and *mpoints* here. An interesting observation is that the dimension of the historical moving object representing the past of a time-spanning moving object is always less than or equal to the dimension of the predictive moving object representing the future. Due to uncertainty, the predictive object part cannot always be modeled with the same precision as the past object part.

So far, we have described that the set of potential future positions or the extent of a moving object can be modeled by a spatiotemporal data type like *mpoint* or *mregion*. However, this concept does not specify the relative change or degree of *confidence* with which a potential future position will eventually be the position of the moving object. To do this, we propose to employ a concept that we call *confidence distribution* such that each potential future position is associated with a degree of confidence. It is the task of an application-specific prediction model to provide an appropriate confidence distribution. Since the prediction model is unknown to a DBMS, our view is that the DBMS regards such a model as an *oracle* that can be consulted on request. This enables us for a future time instant to model the set of potential positions or the possible extent of an object by imposing a confidence distribution on a spatial object representing the future positions. We use the term *confidence* because we do not know how an application models uncertainty and because we permit any theory like probability theory, fuzzy set theory, or rough set theory. We assume that a confidence distribution is given by a function $f: \mathbb{R}^2 \rightarrow [0, 1]$.

An interesting observation is that balloon objects are *static* although they describe the dynamic, temporal evolution of a moving object. The reason is that they represent snapshots describing what the past development and the predicted development of a time-spanning moving object at a time t^{now} is. Since the current time moves on, we obtain a time-spanning

moving object for each current time instant t_i^{now} . This movement represents the dynamics, and we speak of *moving balloon objects*. So-called *spatio-temporal balloon data types* are supposed to enable us to represent them. If we compare t_i^{now} and t_{i+1}^{now} , at t_{i+1}^{now} we know more about the past. That is, the known past at t_i^{now} is contained in the past known at t_{i+1}^{now} . On the other hand, we will probably obtain different predictions at different current times. This is interesting from a querying perspective since it enables us to compare predictions at different current times and to make statements about their validity. Examples of interesting new queries are: (1) What area will potentially be affected by the hurricane XYZ in 10 hours from now? (2) What was thought 12 hours ago about the potentially affected area of XYZ 22 hours later? (3) Assess the match or similarity of both predictions. (4) What is the chance that XYZ will traverse the city ABC 16 hours from now? (5) Determine all cities for which the degree of confidence that they will be hit by XYZ is larger than 70%. (6) Identify all flight routes that are potentially affected by XYZ in the next 24 hours. (7) What is the development of XYZ from three days ago until in 8 hours?

3 Moving Objects in Constrained Environments

Many spatial objects like cars, planes, trains, and people in buildings are restricted in their possible motions since they move in *spatially embedded networks* like roads, highways, railways, lanes in factories for robots, buildings, or airplane routes. We speak of *moving objects in constrained environments*. Therefore, as a conclusion, spatial networks should be taken into account in a data(base) model and database query language for moving objects. This would make it possible to describe movement relative to a network rather than the general 2D space. It would also enable an easier formulation of queries and more efficient representations and indexing of moving objects for network-based database applications.

Section 3.1 deals with research on static spatial networks. Section 3.2 focus on moving objects in these networks.

3.1 Spatial Networks

Spatial(ly embedded) networks or *spatial graphs* are an important spatial concept in the geosciences and have been widely discussed in the literature. They describe a *spatial connectivity structure* (another important one would be *spatial partitions* (Erwig and Schneider 1997, McKenney and Schneider 2007)) and consist of a set of point objects representing their nodes and a set of line objects describing the geometry of their edges.

Examples are transportation networks and supply networks. However, the modeling, representation, and integration of *spatial networks in databases* are a rather open research issue. We first give an overview of available approaches and then identify some open research problems.

3.1.1 Overview

Without going into detail, the manipulation of abstract, non-spatial graphs (networks) in databases has received quite a bit of attention in the database field. However, research on *spatial(ly embedded) graphs* or *spatial networks* is rather limited. Spatial graphs are graphs whose nodes and edges are annotated with a geometry and are thus embedded in the Euclidean space. Chapter 6 of (Shekar and Chawla 2003) gives a good overview of the current state-of-the-art of representing, querying, and implementing spatial graphs in databases. The current representation strategy of spatial graphs is the same as in the non-spatial case: Networks are modeled at a very low abstraction level (based on identifiers and (x, y)-coordinates) by explicit node and edge relations but they are not visible as self-contained entities in the database. Further, there is no explicit support of paths, which the author regards as important. The formulation of queries in SQL is cumbersome, even if a special *connect* clause in SQL2 and a *recursive* clause in SQL3 enable the recursive traversal of a graph structure by deriving the transitive closure of a relation. Commercial solutions like the ESRI Network Data Model (Zeiler 1999) or the Oracle Network Data Model (Oracle Corporation 2005, 2006) follow exactly that strategy. The consequence for applying network operations is that the network data have to be loaded from the node, edge, and many other database tables into a middleware layer outside the database system and that network operations are implemented and executed in this middleware layer. That is, mass data are handled outside the database, network data are kept twice, and updates have to be reported to the database in order to maintain consistency. All this causes a lot of unnecessary overhead. Core database functionality like transactions, query processing, concurrency control, and recovery cannot or can only very limitedly be performed or used in the middleware layer. The reason of this development is the fact that graphs or networks are not (but should be) first-class concepts in DBMSs. The only approaches that go into his intended direction of treating graphs or networks as first-class objects can be found in (Amann and Scholl, 1992, Güting 1991, 1994). In (Güting 1991) relations and graphs coexist as modeling facilities but a graph consisting only of nodes is practically the same as a relation. If several graphs occur in a database, it is hard to separate them in the design. The approach in (Amann and Scholl 1992) offers node and edge objects but no path objects. The proposal in (Güting 1994) is the most interesting

one and offers node, edge, and path objects in an object-oriented setting. A number of special implementation problems have been considered like query processing in spatial network databases (Papadias et al. 2003, Shekhar et al. 1993), matching different road networks (Chen et al. 2006), clustering objects in spatial networks (Yiu and Mamoulis 2004), processing of spatial network queries (Huang et al. 1996, 1997; Shaw et al. 2006), and k-nearest neighbor search in spatial networks (Almeida and Güting 2006).

3.1.2 Open Research Problems

From a conceptual standpoint, a first research issue is the modeling and formal definition of spatially embedded networks and operations on them. Standard databases only allow the user to model spatial networks by the standard facilities of a DBMS data model. This is rather problematic since such a model is unable to design networks as self-contained entities, to represent their spatial connectivity, and to specify and process operations on them (like shortest path). Spatial databases, which are an improvement regarding spatial data handling and offer spatial data types (Schneider 1997), are only able to represent the geometry (points, lines) of the components of spatial networks but do also not have a concept of their connectivity and are also unable to specify operations on them. Hence, the requirement is that such a data model should have an explicit concept of a network embedded in space. That is, spatial networks should be first-class objects in spatial databases. Fulfilling this requirement would lead to easier and more powerful formulations of queries. A second issue is that the classical modeling of networks through nodes and edges might be too simplistic. *Paths* over networks (graphs) are often, or even more, the main conceptual entities of interest. For example, in a road network, roads represent paths over such a network. Of less importance are nodes representing their junctions and edges corresponding to the parts between junctions. A third issue is that networks should not only represent geometric information but also be labeled by thematic information. Further, the thematic information about a network should be extensible and leverage the standard facilities of the DBMS data model. In a relational environment we should be able to create relations to add information relative to a network. For example, it should be possible to label the components of a highway network with information describing motels or speed limits. A fourth issue is the exploration of operations and predicates within a single network and between different networks. Operations on single networks include shortest path computations based on network distance or user-defined weights. Interesting questions on different networks are how they can be combined (union, intersection, difference) and which relationships we can find between

them. A fifth issue is the design of a suitable DBMS query language that incorporates the operations and predicates on networks.

From an implementation standpoint, the first research issue relates to the integration of spatial networks into DBMS. Due to the fact that spatial networks are not first class objects in databases, available implementations are nowadays only provided as middleware layers *outside* the DBMS. The role of databases is limited to delivering basic geometric and thematic components of networks so that networks can be constructed and network operations can be executed in the middleware layer. However, this means that already available standard and spatial database functionality has to be sourced out into the middleware layer, network data have to be duplicated, large volumes of data have to be transferred, and network operations have to be processed in the middleware layer. All this overhead can be avoided if spatial networks and their operations are part of the spatial DBMS. A second issue refers to the design of effective data structures for spatial networks and efficient algorithms for network operations. Classical data structures for networks like the *doubly-connected edge list* (DCEL) or the *winged-edge data structure* are mainly main-memory structures. They have the negative features that they have to be stored in a number of tables and that they require random array access. This is difficult to implement in a DBMS context. Further, they do not harmonize with the concept of spatial data types. A persistent version of these data structures makes it very difficult to support the algorithms implementing the network operations. Hence, the task is to find a better representation of these data structures or even other data structures that support these operations better. Due to their importance, especially paths in networks should be supported. A third issue is the design of algorithms for operations and predicates between different networks.

The author has started some research on this topic and proposes a novel type system or algebra called *Spatial Network Algebra* (*SNETALG*) for modeling and implementing spatial networks in databases. This concept defines networks on the basis of *routes* and *junctions* (crossings) (and not nodes and edges). Routes correspond to roads, streets, or highways in real life and to paths in a graph. A route itself can have several properties. It can be bi-directional (dual route), one-directional (directed route), or the direction does not matter (simple route, example: pedestrian zone). Further, it can have several lanes, and the number of lanes can change within the same route. This is, e.g., interesting for traffic management. Junctions correspond to intersection or meeting points of routes. They play a special role since they are responsible for the connectivity in a network and determine the allowed motion directions ((no) U-turn, one-way). Routes usually have names (like road names) but junctions and parts of roads between junctions usually do not have names.

Abstract data types like *network*, *npoint* (network point), and *nline* (network line) are proposed to represent the network, a position within the network, and a section of the network respectively. A comprehensive collection of operations and predicates includes construction and transformation (import and export) operations from and to predefined network relations, operations on networks (e.g., shortest path), between networks (e.g., union), and between networks and spatial data types (e.g., part of network intersected by region).

3.2 Moving Objects in Spatial Networks

Adding movement relative to a spatial network leads us to *moving objects in networks*. That is, the motion of spatial objects (often point objects) is constrained by a static network. This shows great promise for interesting, new operations and predicates and thus new and, compared to the unconstrained case, different kinds of queries. We first give an overview of available approaches and then identify some open research problems.

3.2.1 Overview

Data models for *moving objects in networks* are rare. The MOST model (Sistla et al. 1997, 1998) uses a particular dynamic location attribute which includes a polyline representing a path over the network plus some additional information like start time, start position, and speed. However, the underlying network is not modeled in any way. Hence, there is no explicit relationship to the network any more. The problem of finding relationships between moving objects and the network in queries is not addressed. The approach in (Vazirgiannis and Wolfson 2001) considers modeling and querying moving objects in road networks. The network model is a relation representing “blocks” that are the edges of the network graph. Each tuple contains a polyline describing the geometry of the edge. The model basically corresponds to an undirected graph where nodes are street crossings and edges are city road blocks. The model is not defined formally but rather an application-specific model. The approaches in (Hage et al. 2003; Jensen et al. 2003b; Speicys et al. 2003) look at data modeling issues for spatial networks regarding possible uses for location-based services. These interesting application studies emphasize the large complexity of real road networks that cannot be adequately modeled by simple directed graph models; they are a good motivation for our planned work. The approach in (Güting et al. 2006) is so far the only one proposing data types for moving objects in networks. Implementation-oriented aspects have, e.g., dealt with index structures for moving objects in networks (Renzos 2003; Pfoser and

Jensen 2003), query processing algorithms for networks (Jensen et al. 2003a; Shahabi et al. 2003), and building test data generators for network-based moving objects (Brinkhoff 2002).

3.2.2 Open Research Problems

Movement in networks can be mainly evoked by two kinds of spatial objects. Either point objects like cars or robots move, or line objects like traffic jams evolve in a network. In this case, we may use our well known spatiotemporal data types for modeling these moving objects, only with the restriction that their motion is constrained to a network.

Considering moving objects in networks results in a number of conceptual research issues. The first issue refers to the problem that both spatial networks and moving objects model geometry and locations. But it is sufficient to store geometry only once. Hence, the question is how moving objects should be modeled in a spatial network without redundant geometric information. This seems to indicate that a new and, compared to the unconstrained case, different data model of moving objects is needed in the network case. A second issue is that the model should allow the user to describe static or moving objects relative to the network, such as static positions (e.g., motels, restaurants, gas stations), static regions (e.g., construction sites, speed limit zones), moving positions (e.g., vehicles), and moving regions (e.g., traffic jams, network parts affected by inundations or hurricanes). A third issue relates to the possible transfer or adaptation of available operations and predicates for moving objects in unconstrained environments to the network case and to the identification and design of possibly new operations and predicates for moving objects in networks. For example, the Euclidean distance concept for unconstrained environments is not meaningful in spatial networks. A fourth issue is the design of a query language for moving objects in networks. A fifth issue is the seamless integration of the data models and query languages for moving objects in unconstrained and constrained environments into a homogeneous and consistent framework.

The consideration of moving objects in networks also leads to implementation issues. A first issue is that if moving objects in networks should have to be modeled differently, different data structures will be needed for them. The consequence would be that such moving objects would be bound to a network and could not be used in an unconstrained environment. A second issue is the immediate impact on the design of algorithms for the operations and predicates on such moving objects. A third issue is the consistent integration of moving objects in networks with their unconstrained counterparts as well as spatial networks. A fourth issue is the embedding of all these concepts into databases.

The author has started some research on this topic and proposes a novel type system or algebra called *Moving Objects in Networks Algebra* (*MONET Algebra*) to model and implement moving objects in spatial networks. The first task is to design appropriate abstract data types for the objects that can move in a network. We have seen that applying the temporal type constructor $\tau(\alpha) = \text{time} \rightarrow \alpha$ to the spatial data types $\alpha \in \{\text{point}, \text{line}\}$ leads to the spatiotemporal data types $\tau(\text{point}) = \text{mpoint}$ and $\tau(\text{line}) = \text{mline}$. In the same way, we can now obtain the spatiotemporal data types *mnpoint* and *mpline* whose movement is constrained to a network by applying τ to the network types *npoint* and *nline*. That is, $\text{mnpoint} = \tau(\text{npoint})$ and $\text{mpline} = \tau(\text{nline})$. The second task is to devise meaningful operations and predicates on these types. An issue is whether some of the operations on moving objects in unconstrained environments can be transferred to the network case. The third task is the design and implementation of effective data structures for the data types and efficient algorithms for the operations and predicates. The fourth task is the design of a query language that is in accord with the query language for networks. The fifth task is the database and query language integration of all concepts in accordance with the integration of spatial networks.

Interesting and new queries for spatial networks and moving objects in them will be possible. Example queries are: (1) Find all sections (edges) of University Avenue that are longer than one mile. (2) Determine the part of the Gainesville road network located east of Main Street. (3) Determine the location of postman Miller at 3pm last Friday. (4) In which part of the network did the postman deliver letters between 10 am and 1pm of last Monday? (5) How many cars have passed Main Street in the last five hours? (6) Which postman is currently nearest to Main Street and moving into that direction (to give him a letter)? (7) Compute a list that determines the current number of construction sites (modeled as *nline* objects) of all U.S. interstates in decreasing order. (8) What are currently the ten longest traffic jams (modeled as *mpline* objects) on the interstate network? (9) Which parts of the network are affected by fog?

The problem that both networks and moving objects in them have absolute locations is solved by only maintaining the absolute locations of networks and by modeling the positions of moving objects relative to the length of a route (rather than an edge of a graph). This is very similar to the concept of *linear referencing* widely used in the *GIS in Transportation* literature and available in commercial database products such as Oracle Spatial (Oracle Corporation 2000). If positions are given relative to edges, then, e.g., a car along an interstate at constant speed needs a change of description at every exit and junction because the edge identifier changes. If positions are given relative to routes, then the description only needs to change when the car changes or leaves the interstate.

4 Conclusions

Moving objects are ubiquitous in our life and a promising concept to adequately model and represent the changing space-time behavior of geometric objects in spatiotemporal applications. From a conceptual standpoint, this chapter gives an overview of the state-of-the-art of moving objects technology in the context of databases and GIS. It especially distinguishes moving objects in unconstrained environments (historical moving objects, predictive moving objects) and constrained environments (spatial networks, moving objects in spatial networks) and sketches a number of open problems in these research fields.

Acknowledgements

This work was partially supported by the National Science Foundation under grant numbers NSF-CAREER-IIS-0347574 and NSF-IIS-0812194.

References

- Agarwal PK, Arge L, Erickson J (2000) Indexing Moving Points. In: ACM Symp. on Principles of Database Systems, pp 175–186
- de Almeida VT, Güting RH (2006) Using Dijkstra's Algorithm to Incrementally Find the k-Nearest Neighbors in Spatial Network Databases. In: ACM Symp. on Applied Computing, pp 58–62
- Amann B, Scholl M (1992) Gram: A Graph Data Model and Query Language. In: European Conf. on Hypertext Technology
- Brinkhoff T (2002) A Framework for Generating Network-Based Moving Objects. *GeoInformatica* 6(2): 153–180
- Chen C-C, Shahabi C, Knoblock CA, Kolahdouzan M (2006) Automatically and Efficiently Matching Road Networks with Spatial Attributes in Unknown Geometry Systems. In: Int. Workshop on Spatio-Temporal Database Management
- Cotelo Lema JA, Forlizzi L, Güting RH, Nardelli E, Schneider M (2003) Algorithms for Moving Objects Databases. *Computer Journal*: 680–712
- Erwig M, Schneider M (1997) Partition and Conquer. In: 3rd Int. Conf. on Spatial Information Theory (COSIT'03), LNCS 1329, Springer, Berlin Heidelberg New York, pp 389–408
- Erwig M, Schneider M (1999) Developments in Spatio-Temporal Query Languages. In: IEEE Int. Workshop on Spatio-Temporal Data Models and Languages, pp 441–449

- Erwig M, Schneider M (2000) Query-By-Trace: Visual Predicate Specification in Spatio-Temporal Databases. In: 5th IFIP 2.6 Working Conf. on Visual Database Systems
- Erwig M, Schneider M (2002) Spatio-Temporal Predicates. *IEEE Trans. on Knowledge and Data Engineering* 14(4): 1–42
- Erwig M, Schneider M (2003) A Visual Language for the Evolution of Spatial Relationships and its Translation into a Spatio-Temporal Calculus. *Journal of Visual Languages and Computing* 14(2): 181–211
- Erwig M, Güting RH, Schneider M, Vazirgiannis M (1998) Abstract and Discrete Modeling of Spatio-Temporal Data Types. In: *ACM Symp. on Geographic Information Systems*, pp 131–136
- Erwig M, Güting RH, Schneider M, Vazirgiannis M (1999a) Spatio-Temporal Data Types: An Approach to Modeling and Querying Moving Objects in Databases. *GeoInformatica* 3(3): 265–291
- Erwig E, Schneider M, Güting RH (1999b) Temporal Objects for Spatio-Temporal Data Models and a Comparison of Their Representations. In: *Advances in Database Technologies (ER'98 Workshop on Spatio-Temporal Data Management)*, LNCS 1552, Springer, Berlin Heidelberg, New York, pp 454–465
- Fenzos R (2003) Indexing Moving Objects on Fixed Networks. In: *Int. Symp. on Spatial and Temporal Databases*, pp 289–305
- Forlizzi L, Güting RH, Nardelli E, Schneider M (2000) A Data Model and Data Structures for Moving Objects Databases. In: *ACM SIGMOD Int. Conf. on Management of Data*, pp 319–330
- Güting RH (1991) Extending a Spatial Database System by Graphs and Object Class Hierarchies. In: *Int. Workshop on Database Management Systems for Geographical Applications*
- Güting RH (1994) GraphDB: Modeling and Querying Graphs in Databases. In: *Int. Conf. on Very Large Data Bases*, pp 297–308
- Güting RH, Schneider M (2005) *Moving Objects Databases*. Morgan Kaufmann Publishers
- Güting RH, Böhlen MH, Erwig M, Jensen CS, Lorentzos NA, Schneider M, Vazirgiannis M (2000) A Foundation for Representing and Querying Moving Objects. *ACM Trans. on Database Systems* 25(1): 1–42
- Güting RH, de Almeida VT, Ding Z (2006) Modeling and Querying Moving Objects in Networks. *VLDB Journal* 15(2): 165–190
- Hadjieleftheriou M, Kollios G, Tsotras VJ, Gunopulos D (2002) Efficient Indexing of Spatiotemporal Objects. In: *Int. Conf. on Extending Database Technology*, pp 251–268
- Hage C, Jensen CS, Pedersen TB, Speicys L, Timko I (2003) Integrated Data Management for Mobile Services in the Real World. In: *Int. Conf. on Very Large Data Bases*, pp 1019–1030
- Huang YW, Jing N, Rundensteiner EA (1996) Path Queries for Transportation Networks: Dynamic Reordering and Sliding Window Paging Techniques. In: *ACM Symp. on Geographic Information Systems*, pp 9–16
- Huang YW, Jing N, Rundensteiner EA (1997) Integrated Query Processing Strategies for Spatial Path Queries. In: *Int. Conf. on Data Engineering*, pp 477–486

- Jensen CS, Kolavr J, Pedersen TB, Timko I (2003a) Nearest Neighbor Queries in Road Networks. In: ACM Symp. on Geographic Information Systems, pp 1–8
- Jensen CS, Pedersen TB, Speicys L, Timko I (2003b) Data Modeling for Mobile Services in the Real World. In: Int. Symp. on Spatial and Temporal Databases, pp 1–9
- Kollios G, Gunopulos D, Tsotras VJ (1999) On Indexing Mobile Objects. In: ACM Symp. on Principles of Database Systems, pp 261–272
- McKenney M, Schneider M (2007) Spatial Partition Graphs: A Graph Theoretic Model of Maps. In: 10th Int. Symp. on Spatial and Temporal Databases, LNCS 4605, Springer, Berlin Heidelberg New York, pp 167–184
- Mokhtar H, Su J, Ibarra OH (2002) On Moving Object Queries. In: ACM Symp. on Principles of Database Systems, pp 188–198
- Oracle Corporation (2000) Oracle Spatial Linear Referencing System Release 8.1.6. User's Guide
- Oracle Corporation (2005) Oracle Database 10g: Oracle Spatial Network Data Model. An Oracle Technical White Paper
- Oracle Corporation (2006) Oracle Spatial Topology and Network Data Models 10g Release 2. Oracle Spatial User's Guide and Reference
- Parent C, Spaccapietra S, Zimányi E (2006) Conceptual Modeling for Traditional and Spatio-Temporal Applications : The MADS Approach. Springer, Berlin Heidelberg New York
- Papadias D, Zhang J, Mamoulis N, Tao Y (2003) Query Processing in Spatial Network Databases. In: Int. Conf. on Very Large Data Bases, pp 802–813
- Pfoser D, Jensen CS (1999) Capturing the Uncertainty of Moving-Object Representations. In: Int. Symp. on Advances in Spatial Databases, pp 111–131
- Pfoser, Jensen CS (2003) Indexing of Network Constrained Moving Objects. In: ACM Symp. on Geographic Information Systems, pp 25–32
- Pfoser D, Jensen CS, Theodoridis Y (2000) Novel Approaches in Query Processing for Moving Object Trajectories. In: Int. Conf. on Very Large Data Bases, pp 395–406
- Schiller J, Voisard A (eds) (2004) Location-Based Services. Morgan Kaufmann Publishers
- Schneider M (1997) Spatial Data Types for Database Systems – Finite Resolution Geometry for Geographic Information Systems, LNCS 1288. Springer, Berlin Heidelberg New York
- Shahabi C, Kolahdouzan MR, Sharifzadeh M. A Road Network Embedding Technique for K-Nearest Neighbor Search in Moving Object Databases. *GeoInformatica* 7(3): 255–273
- Shaw K, Sample J, Ioup E, Abdelguerfi M, Mansion O (2006) Graph processing for spatial network queries. In: Int. Conf. on Information and Knowledge Engineering, pp 3–9
- Shekar S, Chawla S (2003) Spatial Databases: A Tour. Prentice Hall
- Shekhar S, Kohli A, Coyle M (1993) Path Computation Algorithms for Advanced Traveller Information Systems. In: Int. Conf. on Data Engineering, pp 31–39
- Sistla AP, Wolfson O. Temporal Triggers in Active Databases. *IEEE Trans. on Knowledge and Data Engineering* 7(3): 471–486

- Sistla AP, Wolfson O, Chamberlain S, Dao S (1997) Modeling and Querying Moving Objects. In: *Int. Conf. on Data Engineering*, pp 422–432
- Sistla AP, Wolfson O, Chamberlain S, Dao S (1998) Querying the Uncertain Position of Moving Objects. In: Etzion O, Jajodia S, Sripada S (eds) *Temporal Databases: Research and Practice*, LNCS 1399, Springer, Berlin Heidelberg New York, pp 310–337
- Speicys L, Jensen CS, Kligys A (2003) Computational Data Modeling for Network-Constrained Moving Objects. In: *ACM Symp. on Geographic Information Systems*, pp 118–125
- Song Z, Roussopoulos N (2001) K-nearest Neighbor Search for Moving Query Point. In: *7th Int. Symp. on Spatial and Temporal Databases*, pp 79–96
- Tao Y, Papadias D (2003) Spatial Queries in Dynamic Environments. *ACM Trans. on Database Systems* 28(2): 101–139
- Theodoridis Y, Silva JRO, Nascimento MA (1999) On the Generation of Spatio-temporal Datasets. In: *Int. Symp. on Advances in Spatial Databases*, pp 147–164
- Trajcevski G, Wolfson O, Hinrichs K, Chamberlain S (2004) Managing Uncertainty in Moving Objects Databases. *ACM Trans. on Database Systems*: pp 463–507
- Trajcevski G, Wolfson O, Zhang F, Chamberlain S (2002) The Geometry of Uncertainty in Moving Objects Databases. In: *Int. Conf. on Extending Database Technology*, volume LNCS 2287, Springer, Berlin Heidelberg New York, pp 233–250
- Vazirgiannis M, Wolfson O (2001) A Spatiotemporal Model and Language for Moving Objects on Road Networks. In: *Int. Symp. on Spatial and Temporal Databases*, pp 20–35
- Wolfson O, Chamberlain S, Dao S, Jiang L, Mendez G (1998a) Cost and Imprecision in Modeling the Position of Moving Objects. In: *Int. Conf. on Data Engineering*, pp 588–596
- Wolfson O, Xu B, Chamberlain S, Jiang L (1998b) Moving Objects Databases: Issues and Solutions. In: *Int. Conf. on Scientific and Statistical Database Management*, pp 111–122
- Wolfson O, Sistla AP, Chamberlain S, Yesha Y (1999) Updating and Querying Databases that Track Mobile Units. *Distributed and Parallel Databases* 7: 257–287
- Yanagisawa Y, Akahani J, Satoh T (2003) Shape-based Similarity Query for Trajectory of Mobile Objects. In: *Int. Conf. on Mobile Database Management*, pp 63–77
- Yiu ML, Mamoulis N (2004) Clustering Objects on a Spatial Network. In: *ACM SIGMOD Int. Conf. on Management of Data*, pp 443–454
- Zeiler M (1999) *Modeling Our World: The ESRI Guide to Geodatabase Design*. ESRI Press

The Degree Distribution of Random Planar Graphs

Michael Drmota

Institute of Discrete Mathematics and Geometry, Vienna University of Technology, Wiedner Hauptstrasse 8-10, A-1040 Vienna, Austria

Abstract

The degree distribution is a very important characteristic of a network and has been well studied in various kinds of real-world networks (like the internet or social networks) and also from a theoretical point of view in several probabilistic network models. In this paper we present an overview over recent results on the degree distribution of planar graphs and maps, where one assumes that every planar graph with n vertices is equally likely. The main result says that for every fixed $k \geq 1$ the average number of vertices of degree k in a planar graph of size n is (asymptotically) proportional to $d_k n$, where $d_k > 0$.

1 Introduction

Let $G = (V, E)$ be a undirected graph without loops and multiple edges. The degree $d(v)$ of a vertex $v \in V$ is defined as the number of neighbors of v or equivalently as the number of edges that are linked to v . Let n_j be the number of vertices in G of degree j and $n = |V|$ the total number of vertices in G . Then the *degree distribution* of G is given by the sequence

$$d_0 = \frac{n_0}{n}, d_1 = \frac{n_1}{n}, d_2 = \frac{n_2}{n}, \dots$$

The value d_j might be interpreted as the probability that a randomly selected vertex in G has degree j .

The degree distribution of several kinds of networks and graphs is in fact a very important characteristic. So-called small-world networks are usually *scale-free* (Watts and Strogatz 1998), which means that the degree distribution d_j follows (at least asymptotically) a power law

$$d_j \sim cj^{-\alpha} \quad (j \rightarrow \infty)$$

for some $\alpha > 1$. In particular this means that the probability of observing large degrees is relatively large. One of the most prominent examples of scale-free graphs is the Albert-Barabási graph growth model (Albert and Barabási 2002), where a new node is linked to an existing one with probability proportional to the degree. The idea behind this construction principle is to model various (real-world) networks like the internet or social networks. These kinds of models are widely and also well studied, see (Bollobás and Riordan 2004; Bollobás et al. 2001). By the way if the probability of adding a node is equal for all existing nodes then one observes a geometric degree distribution.

However, in this article we focus on planar graphs which do not behave as scale-free networks. Planar graphs are graphs that can be embedded into the plane (or into the 2-sphere) without edge crossings. It should be mentioned that certain planar graphs like Voronoi diagrams and Delaunay triangulations are frequently used in GIS.

Let \tilde{G} denote (one of) the embedding(s) of a planar graph G into the plane \mathbb{R}^2 . The complement of the embedding $\mathbb{R}^2 \setminus \tilde{G}$ consists then of several connected components, the so-called *faces* of \tilde{G} . For example, if G is connected then the number of vertices $|V|$, the number of edges $|E|$ and the number of faces $|F|$ satisfy the (Euler) relation

$$|V| - |E| + |F| = 2$$

In particular this means that the number of faces is independent of the embedding.

If there is a unique embedding into the plane then one can introduce the *dual graph* G^* . The vertices of G^* are the faces of G , and two faces of G are joined by an edge (in G^*) if they have a common edge (in G). G^* is then another planar graph with a unique embedding and G^{**} is isomorphic to G .

One also distinguishes between connected, 2-connected, and 3-connected (planar) graphs. A graph is 2-connected if it is connected and one has to remove at least two vertices (and all incident edges) to disconnect it. Similarly, a graph is 3-connected if it is 2-connected and one has to

remove at least three vertices to disconnect it. Fig. 1 shows a connected, a 2-connected and a 3-connected planar graph. 3-connected planar graphs are of special interest since Whitney's theorem (Whitney 1932) says that they have a unique embedding.

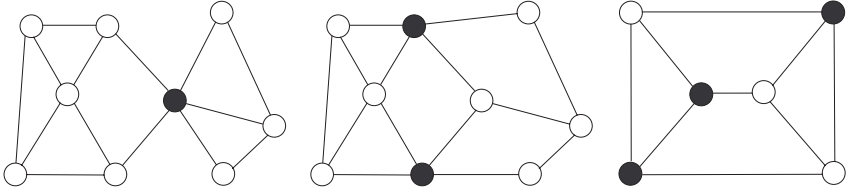


Fig. 1. A connected, a 2-connected and a 3-connected planar graph. If the black vertices are removed then the corresponding graphs get disconnected.

The main part of this paper is devoted to the description of a recent result on the degree distribution of random planar graphs that is due to Drmota, Giménez and Noy (Drmota et al. 2008). We survey the history, discuss the counting problem and describe the degree distribution analytically in terms of generating function.

Finally we compare these results with corresponding results for planar maps (already embedded planar graphs) and for Erdős-Rényi random graphs.

2 Planar Graphs

The counting problem of several classes of planar graphs resp. planar maps goes back to Tutte (Tutte 1962, 1963; Brown and Tutte 1964). Interestingly enough, the study of random vertex labelled planar graphs is a recent one. Recall that vertex labelled means that a graph with n vertices is labelled with $1, 2, \dots, n$. In this context we consider two graphs as different if they have different labellings even if they are isomorphic as unlabelled graphs. Furthermore we do not distinguish between different embeddings of a planar graph, see Fig. 2. The choice of unlabelled graphs makes the problem feasible to a theoretical analysis. A heuristic reason for this observation is that symmetries that appear in an unlabelled graph are not taken into account in the labelled case. Nevertheless one expects qualitatively the same kinds of results for unlabelled planar graphs which might be more natural in the context of GIS.

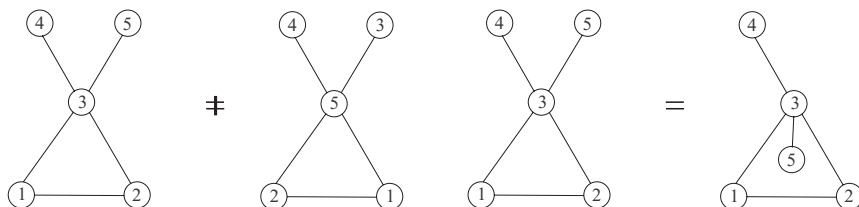


Fig. 2. The two vertex labelled graphs on the left hand side are different due to their different label distribution. They are certainly isomorphic—as unlabelled graphs—but no isomorphism transfers all labels. However, the two graphs on the right hand side are isomorphic as labelled planar graphs although their embeddings are different.

Let \mathfrak{R}_n denote the set of all different vertex labelled planar graphs with n vertices. If we assume that each graph in \mathfrak{R}_n appears equally likely then we call it a *random planar graph*. This kind of notion was introduced by Denise et al. (Denise et al. 1996), and since then random planar graphs have been widely studied. Several natural parameters defined on \mathfrak{R}_n have been studied, starting with the number of edges, which is probably the most basic one. Partial results were obtained in (Denise et al. 1996; Gerke and McDiarmid 2004; Osthus et al. 2003; Bonichon et al. 2006), until it was shown by Giménez and Noy (2009) that the number of edges in random planar graphs obeys asymptotically a normal limit law with linear expectation and variance. The expectation is asymptotically κn , where $\kappa = 2.21326\dots$ is a well-defined analytic constant¹. This implies that the average degree of the vertices is $2\kappa = 4.42652\dots$ McDiarmid et al. (2005) showed that with high probability a planar graph has a linear number of vertices of degree k , for each $k \geq 1$.

In what follows we present here a recent approach to random (vertex labelled) planar graphs that is based on generating functions and indicate how corresponding counting problems and distributional results can be obtained.

We first provide a system of equations for the corresponding generating functions (see Bender et al. 2002; Giménez and Noy 2009)².

¹ Here and in what follows the notation *well-defined analytic constant* means that the constant is defined by an analytic equation where only elementary functions occur. From these equations—which are quite involved in the present context—one can compute approximate values to any prescribed precision.

² A *generating function* of a sequence a_n is usually a power series of the form $A(x) = \sum_n a_n x^n$ or $\hat{A}(x) = \sum_n a_n x^n / n!$. These functions capture all information of the sequence a_n in one mathematical object. Similarly a double sequence

Theorem 1. Let $b_{n,m}$ denote the number of 2-connected labelled planar graphs, $c_{n,m}$ the number of connected labelled planar graphs and $g_{n,m}$ the number of all labelled planar graphs with n vertices and m edges. Furthermore, let

$$B(x, y) = \sum_{m,n \geq 0} b_{n,m} \frac{x^n}{n!} y^m, \quad C(x, y) = \sum_{m,n \geq 0} c_{n,m} \frac{x^n}{n!} y^m,$$

$$G(x, y) = \sum_{m,n \geq 0} g_{n,m} \frac{x^n}{n!} y^m$$

the corresponding generating functions. Then these functions are determined by the following system of equations:

$$G(x, y) = \exp(C(x, y)),$$

$$\frac{\partial C(x, y)}{\partial x} = \exp\left(\frac{\partial B}{\partial x}\left(x \frac{\partial C(x, y)}{\partial x}, y\right)\right), \tag{2.1}$$

$$\frac{\partial B(x, y)}{\partial y} = \frac{x^2}{2} \frac{1 + D(x, y)}{1 + y},$$

$$\frac{M(x, D(x, y))}{2x^2 D(x, y)} = \log\left(\frac{1 + D(x, y)}{1 + y}\right) - \frac{x D(x, y)^2}{1 + x D(x, y)},$$

$$M(x, y) = x^2 y^2 \left(\frac{1}{1 + xy} + \frac{1}{1 + y} - 1 - \frac{(1 + U)^2 (1 + V)^2}{(1 + u + v)^3} \right), \tag{2.2}$$

$$U(x, y) = xy(1 + V(x, y))^2,$$

$$V(x, y) = y(1 + U(x, y))^2.$$

Note that the number of edges have to be taken into account, too, the equations (2.1) and (2.2) could not be stated without the variable y . Nevertheless, we can set $y=1$ after all and obtain generating functions for the numbers $b_n = \sum_{m \geq 0} b_{n,m}$ etc., see (Bender et al. 2002; Giménez and Noy 2009).

Theorem 2. The numbers b_n , c_n and g_n of labelled 2-connected resp. connected resp. all planar graphs are asymptotically given by

$a_{n,k}$ can be related to a power series—or generating function—in two variables. The well-known Z-transform $X(z) = \sum_n x(n)z^{-n}$ in signal processing is just a variant of a generating function.

$$b_n = b \cdot n^{-\frac{7}{2}} \rho_1^n n! \left(1 + O\left(\frac{1}{n}\right)\right),$$

$$c_n = c \cdot n^{-\frac{7}{2}} \rho_2^n n! \left(1 + O\left(\frac{1}{n}\right)\right),$$

$$g_n = g \cdot n^{-\frac{7}{2}} \rho_2^n n! \left(1 + O\left(\frac{1}{n}\right)\right),$$

where $\rho_1 = 0.03819\dots$, $\rho_2 = 0.03672841\dots$ and
 $b = 0.3704247487\dots \cdot 10^{-5}$,
 $c = 0.4104361100\dots \cdot 10^{-5}$,
 $g = 0.4260938569\dots \cdot 10^{-5}$

are well-defined analytic constants.

It is much more difficult to get a precise description of the singular behaviour of the above generating functions than in the previous examples. Nevertheless, it turns out that $B(x,y)$, $C(x,y)$, and $G(x,y)$ have a representation of the form

$$g(x,y) + h(x,y) \left(1 - \frac{x}{\rho(y)}\right)^{5/2}$$

with certain analytic function $g(x,y)$, $h(x,y)$, $\rho(y)$. Of course, this implies Theorem 2. Furthermore, we directly obtain a central limit theorem for the number X_n of edges where expected value and variance are asymptotically proportional to n :

$$EX_n = \mu n + O(1) \quad \text{and} \quad VX_n = \sigma^2 n + O(1).$$

For example, for connected resp. all planar graphs we have $\mu = 2.2132652\dots$ and $\sigma^2 = 0.4303471\dots$ (compare with Giménez and Noy 2009).

By extending the above procedure one can also get a description of the degree distribution of planar graphs. This has been worked out recently by Drmota, Giménez and Noy (Drmota et al. 2008).

Theorem 3. *Let $d_{n,k}$ be the probability that a random node in a random planar graph \mathfrak{R}_n has degree k . Then the limit*

$$d_k := \lim_{n \rightarrow \infty} d_{n,k}$$

exists. The probability generating function $p(w) = \sum_{k \geq 1} d_k w^k$ can be explicitly computed. The first few values are given in the following table and asymptotically we have

$$d_k \sim ck^{-1/2}q^k,$$

where $c = 3.0826285\dots$ and $q = 0.6734506\dots$ are well-defined analytic constants.

d_1	d_2	d_3	d_4	d_5	d_6
0.0367284	0.1625794	0.2354360	0.1867737	0.1295023	0.0861805

The proof is based on the generating functions $B^*(x,y,w)$, $C^*(x,y,w)$ and $G^*(x,y,w)$ that correspond to 2-connected, connected resp. all planar graphs, where one vertex is marked and the exponent of w counts the degree of the rooted vertex. The corresponding system of equations is now the following one (see Drmota et al. 2008):

$$\begin{aligned} G^*(x,y,w) &= \exp(C(x,y,1))C^*(x,y,w), \\ C^*(x,y,w) &= \exp(B^*(xC^*(x,y,1),y,w)), \\ w \frac{\partial B^*(x,y,w)}{\partial w} &= xyw \exp\left(S(x,y,w) + \frac{1}{x^2 D(x,y,w)} \times T^*\left(x, D(x,y,1), \frac{D(x,y,w)}{D(x,y,1)}\right)\right) \\ D(x,y,w) &= (1+yw) \exp\left(S(x,y,w) + \frac{1}{x^2 D(x,y,w)} \times T^*\left(x, D(x,y,1), \frac{D(x,y,w)}{D(x,y,1)}\right)\right) - 1 \\ S(x,y,w) &= xD(x,y,1)(D(x,y,w) - S(x,y,w)), \\ T^*(x,y,w) &= \frac{x^2 y^2 w^2}{2} \left(\frac{1}{1+wy} + \frac{1}{1+xy} - 1 \right. \\ &\quad \left. - \frac{(U+1)^2 \left(-w_1(U,V,w) + (U-w+1)\sqrt{w_2(U,V,w)} \right)}{2w(Vw+U^2+2U+1)(1+U+V)^3} \right), \\ U(x,y) &= xy(1+V(x,y))^2, \quad V(x,y) = y(1+U(x,y))^2 \end{aligned}$$

with polynomials $w_1 = w_1(U,V,w)$ and $w_2 = w_2(U,V,w)$ given by

$$\begin{aligned} w_1 &= -UVw^2 + w(1 + 4V + 3UV^2 + 5V^2 + U^2 + 2U + 2V^3 \\ &\quad + 3U^2V + 7UV) + (U+1)^2(U + 2V + 1 + V^2), \\ w_2 &= U^2V^2w^2 - 2wUV(2U^2V + 6UV + 2V^3 + 3UV^2 + 5V^2 + U^2 \\ &\quad + 2U + 4V + 1) + (U+1)^2(U + 2V + 1 + V^2)^2. \end{aligned}$$

It turns out that that the singular behaviour of the functions $B^*(x,y,w)$, $C^*(x,y,w)$ and $G^*(x,y,w)$ is of the form

$$g(x, y, w) + h(x, y, w) \left(1 - \frac{x}{\rho(y)}\right)^{\frac{3}{2}},$$

i.e., the singularity does not depend on w . With help of these kinds of representations the degree distributions can be characterised.

3 Random Planar Maps

The difference between planar maps and planar graphs is that a planar map is an already embedded planar graph. If a planar graph has several non-equivalent embeddings on the 2-sphere then each of them corresponds to different planar maps although the underlying planar graph is the same. In this context it is usual to consider unlabelled graphs and to fix the number of edges and to distinguish one of the edges on the outer face which is called *root edge*. For example, Fig. 3 shows a planar map which is also a triangulation of an n -gon.

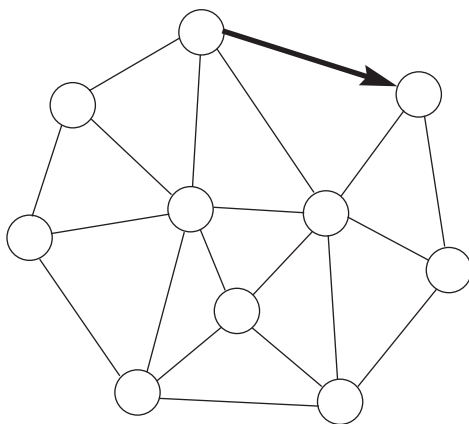


Fig. 3. Triangulation of an n -gon with an oriented root edge on the outer face.

As already mentioned the counting problem of (several classes of) planar maps goes back to Tutte (Tutte 1962, 1963; Brown and Tutte 1964). Also the counting problems (and their asymptotics) is very well known, see (Bender and Richmond 1986). The asymptotic number m_n of planar maps (in a certain class) with n edges follows (usually) a pattern of the form

$$m_n \sim cn^{-5/2} p^n \quad (n \rightarrow \infty)$$

This is in accordance with the results of Theorem 2. Of course, there is no factorial factor $n!$ since the vertices are unlabelled. Furthermore, the factor $n^{-5/2} = n^{-7/2}$ stems from the fact that one edge is distinguished.

The degree distribution d_k of several classes of random planar maps is also well studied (see the survey paper Liskovets 1999). As in the case of planar graph we observe that

$$d_k \sim ck^{-1/2}q^k \quad (k \rightarrow \infty)$$

for certain constants $c > 0$ and $0 < q < 1$. This is true in particular for the class of all planar maps, for Eulerian planar maps (where one restricts on even degrees), for 3-connected planar maps, or for 2- and 3-connected triangulations.

Finally we mention that the dual M^* of a planar map M is again a planar map and the degree of a vertex in M translates into the valency of the corresponding face of M^* . Thus, one observes that the valency distribution of faces in random planar maps obeys the same behaviour as the degree distribution.

4 The Random Graph Model of Erdős and Rényi

In this final section we want to compare random planar graphs with the random graph model of Erdős and Rényi (1959, 1960). This model is defined as follows: Given are n vertices with labels $1, \dots, n$ and a probability p . Then each of the $\binom{n}{2}$ possible edges are included in the edge set of the graph with probability p , where all edges are treated independently. The resulting graph is denoted by $G(n, p)$. In particular it follows that the expected number of edges is given by $p\binom{n}{2}$ and the probability d_k that a random vertex has degree k equals

$$d_{n,k} = \binom{n-1}{k} p^k (1-p)^{n-1-k}.$$

If we want to compare these two models we have to assure that the number of edges is approximately the same. In random planar graphs the number of edges is approximately κn , where $\kappa = 2.21326\dots$. Thus, we have to assume that $p = c/n$ with $c = 2\kappa$ so that there are $\approx cn/2 = \kappa n$ edges in the Erdős-Rényi model. However, in this case we get (as $n \rightarrow \infty$)

$$d_{n,k} = \binom{n-1}{k} p^k (1-p)^{n-1-k} \sim e^{-k} \frac{c^k}{k!} = d_k,$$

which means that the degree distribution d_k follows a Poisson law with parameter c . Therefore the behaviour of Erdős-Rényi random graphs with respect to the degree distribution is (again) completely different from random planar graphs.

Finally we comment on one interesting phenomenon of Erdős-Rényi random graphs, namely the emergence of a giant component. If we let $n \rightarrow \infty$ and set (again) $p=c/n$, then for $c < 1$ a typical graph consists of small and simple components, i.e., each component has a typical size of $O(\log n)$ and does not contain many cycles. If $c > 1$, then a typical graph consists of one giant component which comprises roughly $2/3$ of all vertices and many small and simple components. Actually one can observe a similar phenomenon for random planar graphs. It is either connected or consists of one giant component together with a few nodes in small planar components, see (McDiarmid 2008).

5 Conclusions

In this paper we have studied the degree distribution of large (random) planar graphs. Special kinds of planar graphs like Voronoi diagrams and Delauney triangulations are frequently used in GIS. The main result (Theorem 3, fully proved in Drmota et al. 2009) says that the probability d_k that a randomly chosen vertex has degree k is asymptotically given by

$$d_k \sim ck^{-1/2}q^k$$

for certain constants $c > 0$ and $0 < q < 1$. This means that the degree distribution follows almost a geometric law. This is in accordance with similar models like planar maps (see Section 3) and should be also typical for planar graphs that appear in GIS.

This property is in contrast to the observed degree distribution of so-called scale-free networks like the internet, where the degree distribution is asymptotically a power law which has a *large tail*. There is also a big difference to Erdős-Rényi random graphs, where the degree distribution follows asymptotically a Poisson law with a subexponential tail.

References

- Albert R, Barabási A-L (2002) Statistical Mechanics of complex networks. *Rev. Mod. Phys.* 74: 47–97
- Bender EA, Richmond LB (1986) A survey of the asymptotic behaviour of maps. *J. Combin. Theory Ser. B* 40: 297–329
- Bender EA, Gao Z, Wormald NC (2002) 2-connected labelled planar graphs. *Elec. J. Combinatorics* 9(#43)
- Bollobás B, Riordan O (2004) The diameter of a scale-free random graph. *Combinatorica* 24: 5–34
- Bollobás B, Riordan O, Spencer J, Tusnády G (2001) The degree sequence of a scale-free random graph process. *Random Structures Algorithms* 18: 279–290
- Bonichon N, Gavoille C, Hanusse N, Poulalhon D, Schaeffer G (2006) Planar graphs, via well-orderly maps and trees. *Graphs and Combinatorics* 22: 185–202
- Brown WG, Tutte WT (1964) On the enumeration of rooted non-separable planar maps. *Can. J. Math.* 16: 572–577
- Denise A, Vasconcellos M, Welsh DJA (1996) The random planar graph. *Congr. Numer.* 113: 61–79
- Drmotá M, Giménez O, Noy M (2008) Degree distribution in random planar graph. In: *DMTCS Proceedings AI*, pp 163–178
- Erdős P, Rényi A (1959) On random graphs. I. *Publ. Math. Debrecen* 6: 290–297
- Erdős P, Rényi A (1960) On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.* 5: 17–61
- Gerke S, McDiarmid C (2004) On the number of edges in random planar graphs. *Combin. Probab. Comput.* 13(2): 165–183
- Giménez O, Noy M (2009) Asymptotic enumeration and limit laws of planar graphs. *J. Amer. Math. Soc.* 22: 309–329
- Liskovets VA (1999) A pattern of asymptotic vertex valency distributions in planar maps. *J. Combin. Theory Ser. B* 75(1): 116–133
- McDiarmid C (2008) Random graphs on surfaces. *J. Combin. Theory Ser. B* 98(4): 778–797
- McDiarmid C, Steger A, Welsh DJA (2005) Random planar graphs. *J. Combin. Theory Ser. B* 93: 187–205
- Osthus D, Prömel HJ, Taraz A (2003) On random planar graphs, the number of planar graphs and their triangulations. *J. Combin. Theory Ser. B* 88: 119–134
- Tutte WT (1962) A census of planar triangulations. *Canad. J. Math.* 14: 21–38
- Tutte WT (1963) A census of planar maps. *Canad. J. Math.* 15: 249–271
- Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. *Nature* 393: 409–410
- Whitney H (1932) Congruent Graphs and the Connectivity of Graphs. *Amer. J. Math.* 54(1): 150–168

Geographical Information Engineering in the 21st Century

Gilberto Câmara¹, Lúbia Vinhas¹, Clodoveu Davis², Fred Fonseca³, Tiago Carneiro⁴

¹ National Institute for Space Research (INPE), Image Processing Division, São José dos Campos, Brazil

² Computer Science Department, Federal University of Minas Gerais, Belo Horizonte, Brazil

³ College of Information Sciences and Technology, Pennsylvania State University, State College, USA

⁴ Computer Science Institute, Federal University of Ouro Preto, Brazil
gilberto.camara@inpe.br, lubia@dpi.inpe.br, clodoveu@dcc.ufmg.br,
fred.fonseca@ist.psu.edu, tiago@icep.ufob.br

Abstract

This paper discusses the challenges facing GIS designers in the 21st century. We argue that GI engineers lack a sound theoretical basis that would allow them to make best use of new technologies that handle geospatial data. Considering three important topics for the new generations of GIS (change, semantics, and cognition) we show that GIS theory is in a state of flux. Thus, researchers and engineers need to cooperate more for the new generation of GIS to be built in the best possible way.

1 Introduction

Although the term ‘geographical information science’ (*GIScience*) is well-established in the scientific literature, the idea of geographical information engineering (*GIEngineering*) has received much less attention by both

researchers and practitioners. The idea of GIS (Geographical Information System) dates from Roger Tomlinson's pioneering work in the 1960s (Tomlinson 1972). The term 'Geographic Information Science' stems from the early 1990s (Goodchild 1992b), labelling a field that had developed in the 1970s and 1980s because of the need of the scientific foundation to further advance spatial information handling. The existence and evolution of GIS has motivated a significant part of the research agenda for GIScience. In the 1980s and early 1990s, there were papers describing the design of a GIS (Morehouse 1992; Herring 1992). As the discipline of GIScience evolved in the 1990s and early 21st century, there is a limited amount of published research on how GIScience has influenced the design and evolution of GIS technology. This is surprising, considering the widespread use of GIS technology that helped to promote GIScience as a scientific discipline.

During the 1980s and 1990s, the scientific results produced by researchers in this area helped to set up the current billion dollar industry of Geographical Information Systems (GIS). GIS is now regularly being used as a corporate tool to manage large geospatial databases, and as a research tool for understanding our environment. However, almost all current GIS applications use static data, which represent temporal information and information on change too simply, if at all. The new generation of GIS, called GIS-21 (or "GIS for the 21st century") will be different from GIS-20 (or "GIS for the 20th century"), thanks to scientific and technological advances. These advances include the distributed spatial processing on the Web and a new generation of mobile devices and remote sensors.

Ideally, there would be a stable corpus of scientific knowledge that would be the basis for the GI engineer's practice. Currently, such corpus exists only for GIS-20, mostly in the form of the OGC standards. *What about GIS-21, which will use new technologies like constellations of earth observation satellites, sensor networks, and mobile devices?* Based on the authors' experience on both sides of the trenches (research and technology), we consider GI engineers lack a sound theoretical basis that would allow them to make best use of these technologies. This paper aims to show why this happens, and how the GIScience and GIEngineering communities could cooperate to build reliable products that are also innovative.

In this light, this paper considers some questions: *"In what ways does GIS-21 differ from GIS-20? What would GI engineers need to know to build GIS-21? Is the relevant scientific knowledge organized and stable? How could GIScientists and GIEngineers cooperate?"* In what follows, we provide our views on these topics. We are aware that a full response would be hard. However, we consider that providing partial guidance and insights based on experience is useful for both communities.

2 From GIS-20 to GIS-21

We define geographical information engineering as “*The discipline of systematic construction of geographical information systems and associated technology, drawing on scientific principles. It also includes adapting existing technology to fit user and societal needs and the technical, legal and economic evaluation of GIS technology*”. This definition highlights the crucial role of the scientific principles as a basis for sound engineering. But there is a fundamental difference in the scientist and the engineers’ approach. Fred Brooks says that “*the scientist builds in order to study; the engineer studies in order to build*” (Brooks Jr. 1996). A good engineer studies the literature and chooses which scientific principles are relevant for his task. Following his advice, it is important the GI engineer gains a critical understanding of the science produced in his field.

How does Brooks’ view apply to geographical information systems? To answer this question, we need to consider how hard it is to set up the scientific basis for a GIS. To start, consider defining a “geographical information system”. In the 1980s and 1990s, a GIS was a stand-alone system that provided methods for input, storage, processing and display of geospatial data. In the 2000s, the technology was extended to corporative systems that support multiple users with a spatial database. Use of the Internet further broadened the technology, by allowing building of web-based visualisation and processing tools. The new generation of mobile devices allows geospatial data to be accessible almost anywhere. Thus, any information system that integrates, stores, edits, analyzes, shares, and displays geospatial data can be considered as a ‘GIS’.

Although the ways of using geospatial data are multiple, there is a common basis for all different types of GIS. It is here the centuries-old tradition of cartography that comes to rescue. We have grown familiar with the abstractions involved in map-making which include a two-dimensional projection of the earth’s surface and assigning boundaries. Thus, setting up the scientific principles for dealing with 2D static data was relatively straightforward. An early landmark was the *Harvard Papers on Geographical Information Systems* (Dutton 1978). Next, came Egenhofer’s work on topological spatial relations (Egenhofer and Franzosa 1991), Couclelis’ discussion of field and object models (Couclelis 1992), and Goodchild’s work on spatial data modelling (Goodchild 1992a). Frank and Egenhofer showed how object-oriented GIS would work (Egenhofer and Frank 1992). Their work had an immediate influence on the design of SPRING (Câmara et al. 1996), a free GIS that has a large user base. Later, other products such as ArcGIS adopted the object-oriented model.

This sound scientific basis on issues of 2D data structures, modelling and display enabled a generation of GIS technology to emerge, most of

them sharing similar design principles. This led to the establishment of standards on the field, an effort led by the Open Geospatial Consortium. GI engineers that develop GIS-20 products benefit from the substantial intellectual effort that went into setting up the OGC standards.

No such comfortable solution exists for GIS-21, where new technologies are a major force. Take the Internet. The abstractions encapsulated in OGC's Web standards (WMS, WCS, WFS, WPS) deal mostly with a non-cooperative environment. Using OGC's standards, users have access to information produced by others, mostly for visualization. The user is thus a passive consumer of information produced elsewhere. However, emerging Web applications emphasize cooperation and interaction. Using social networks in the Internet, GI engineers will build collaborative systems that go beyond the simple OGC abstractions.

Consider also geosensors, which provide a 'virtual' connection with the environment, and allow new approaches to the study of environmental processes. These new sources of information were not available earlier due to high cost of measurement or to inaccessibility for analysts. Current OGC standards associated with geosensors focus on low-level communication and issues such as fault tolerance, reliability, and scalability. These standards do not consider how to transform sensor data into information for monitoring the environment. This transformation will need the capacity to model the processes measured by sensor networks. GIS-21 systems need to move from low-level details to high-level domain conceptualizations about change.

Remote sensing images provide a further source of new data for understanding our environment. The new generation of remote sensing satellites already launched or planned for the next decade will provide much new data. Consider land imaging. Most images of the Earth's land surface come from a single source: the LANDSAT series of satellites. LANDSAT covers the Earth every 16 days with 30 meter resolution. From 2010 onwards, there will be a constellation of land imaging satellites, providing free moderate resolution (20-50 meter) images every two days for the whole planet. There will be many high-resolution satellites (2 meter resolution or better) that will provide frequent detailed information. This deluge of remote sensing data will allow new image analysis techniques. An environmental GIS-21 should be able to *search for changes* in a sequence of remote sensing images instead of the current *search for content* on a single image (Câmara et al. 2001). The emphasis should not be placed on simple *image classification* procedures, but on *capturing dynamics* over the landscape. Using multitemporal remote sensing data, GIS-21 tools should be able to describe the change trajectories at local and regional scales.

Thus, the relatively comfortable situation in the 1990s, where a shared conceptualization of GIS helped both designers and users to develop simi-

lar products, no longer holds. There is no longer a ‘typical GIS’. The new scientific and technological challenges created a new set of essential difficulties for the new generation of GIS. These challenges include modelling the semantics of communication of spatial concepts, understanding change in space and time, and developing information extraction methods for massive data sources. These problems are hard, and will remain so.

3 Change, Cognition, and Semantics: Three Critical Issues

As discussed above, GIS have evolved from automated mapping applications to a set of technologies concerned with information about processes in the human environment. To grasp the full extent of the difference between GIS-21 and GIS-20 we will consider three critical issues for GIS-21 applications that were mostly absent of GIS-20 designs. These are *change, semantics, and cognition*. In this section, we will give an outline of the main research challenges in these areas. In the next section, we will focus on the GIEngineering challenges for modelling change in more detail.

3.1 Change

Representing change in GIS-21 is not only an issue of handling time-varying data. It also concerns how objects acquire or lose their identity, how their properties change, what changes happen simultaneously, and what the laws of nature and the interactions among people that bring about change. Time can be viewed as an independent entity of the universe, a dimension in which events occur in sequence. That is the view subscribed by Newton and used in the tradition of experimental physics. A second view is to consider time as an intellectual structure within which humans sequence and compare events. This second view is the tradition of Leibniz and Kant. These two opposing views lead to the controversy in the philosophy of time over whether extension in time is analogous to extension in space, the so-called 3D/4D controversy. For a further philosophical discussion of spatio-temporal concepts, see Grenon and Smith (2003), Galton (2004) and Frank (2003).

Given the unsolved 3D/4D controversy, when a GI engineer has to design of a GIS that deals with change, he faces difficult choices. The first and most difficult question is: “*How can a GIS represent change?*” The engineer’s practical answer is “*it depends on the nature of the data*”. We consider the following broad choices.

For applications that involve moving objects, such as transportation, location-based, and or animal-tracking systems, there are some basic decisions about what details and constraints are to be represented. For autonomous objects on well-defined path (such as roads), we can use the ideas of *trajectory* and associated operations, along the lines proposed by Güting and Schneider (2005). In this case, change is stored implicitly in the objects' position. Applications whose concepts draw on Hägerstrand's "time geography" (1967) involve modelling personal choices (Miller 2003).

Cadastral applications need a different approach, as they undergo incremental change (as when a parcel is divided). Change is both a property of each object and the result of actions in these objects from external forces. A GIS-21 for cadastral applications should be able to capture both (a) the geographical entities subject to change and (b) the *goals* associated to the causes that cause these entities to change. A good starting point for the GI engineer of cadastral applications is the bitemporal spatial model of Worboys (1994) and Medak's model of lifestyles (2001). These models can be extended into a set of spatio-temporal types (Bittencourt et al. 2007). A more complete alternative is to use the event calculus proposed by Worboys (2005) to develop an application that would include both objects and events as primitives. Events (*occurrents*) correspond to the procedures that perform changes in objects (*perdurants*). Event modelling requires setting up the constraints, conditions, and operations that set off object evolution.

Environmental applications pose a different challenge for the GI engineer. Humanity is changing the rural and urban landscapes at an unprecedented pace, and human transformations of ecosystems and landscapes are the largest source of change in the natural systems on earth. GIS-21 should provide a computing environment for modelling human-environment interactions in ways that can be understood by practitioners from different disciplines. It should provide good information extraction tools from remote sensing images and from geosensors. For example, a remote sensing image is a measurement that captures snapshots of change trajectories. An environmental GIS-21 should be able to *search for changes* instead of the *search for content*. The emphasis should not be placed on simple *object matching and identification* procedures, but on *capturing dynamics* over the landscape (Silva et al. 2005).

In resume, finding a unique theory of spatio-temporal models and operators is an arguably unsolvable problem. This irremovable complexity is a direct result of the ambiguity when defining 'time'. The GI engineer who wants to represent change needs first to define the needs and constrains of his application and then choose a suitable approach, from the many available scientific proposals.

3.2 Semantics

We start to build a GIS by recognising objects in the real world and assigning geographical locations to them. This means that any GIS includes much semantics, a fact neglected until recently. Recognizing how important semantics is for interoperability and for intelligent GIS, some researchers proposed that GIS should be ontology-driven (Fonseca et al. 2002). Semantics also motivated institutions to build spatial ontologies. However, the applicability of such large ontologies remains limited and is mostly useful as means of documentation. Using ontologies for interoperability remains a difficult task, since the matching problem is hard to solve.

For a GI engineer, the most useful results in this area are insights into the problem of spatial semantics. These insights direct an engineer to build representations and interfaces that are more precise in their definition. A useful work is the distinction between continuants and occurrents on a spatio-temporal ontology, the so-called SNAP-SPAN ontology (Grenon and Smith 2003). Also useful is Frank's idea of 'tiers of ontology' (2001). He shows there are different levels of abstraction in a GIS. Frank's approach is relevant to GIE, since he takes a practical approach. Using this approach to build a GIS, the GI engineer would first select which tiers of ontology he will focus. For example, a remote sensing image processing software would transform between data on Frank's tier 1 (*observations of physical world*) to data on tier 2 (*objects with properties*). This is a possible way for building GIS that use semantic properties, even in a limited extent.

3.3 Cognition

Spatial cognition concerns the study of knowledge and beliefs about spatial properties of objects and events in the world (Montello 2001). The field is intensely multidisciplinary, with contributions from linguistics (Lakoff and Johnson 1980), psychology (Tversky 1993), and computer science (Freksa 1991; Krieg-Brückner and Shi 2006). GIScientists have been studying spatial cognition since the early 1990s, stressing issues such as navigation and wayfinding, spatial communication via language, and cognitive maps. Their research highlights how important cognition is for human use of space (Mark and Frank 1991).

From a GI engineer's viewpoint, new mobile devices with navigation possibilities have opened a big opportunity for GIS-21 applications. They range from map-based navigation systems in cars and mobile phone to intelligent transport applications. In the long-term view of transportation, different modalities (train, bus, car) would be linked. The user would be guided to the most efficient one based on his plans, route congestion, and

environmental preservation. The main drawback for the engineer's design is the absence of proven formal models for spatial cognition. Arguably, achieving a formal approach to cognition would be akin to solving the problem of consciousness (Searle 1997). The sheer complexity and variety of processes that interact in spatial cognition prevents a formal approach from being sufficient as a unique basis for sound GI engineering. In other words, engineers use good formal models plus a fair amount of hacking.

Early efforts on spatial cognition stressed image schemata and linguistic issues (Mark and Frank 1991; Kuhn and Frank 1991) and on human-centered views of space, described as "naïve geography" (Egenhofer and Mark 1995). Such research revealed many insights, but no comprehensive theory emerged. The main drawback for the GI engineer's planning to use results from spatial cognition in his tools is the scarcity of proven formal models.

Formal models exist only in a limited number of cases. Frank's papers on qualitative spatial relations (Frank 1996) show that it is possible to define cardinal directions with predicate calculus and relations. However, as Frank notes in a recent work (Frank 2007) one of the main challenges in spatial cognition is the intricacy of the formal models that describe even problems of limited scope. The sheer complexity of spatial cognition prevents a formal approach from being a basis for sound engineering. Nevertheless, the GI engineer can gather many interesting ideas for practical applications from works such as (Golledge 1999) and (Egenhofer and Golledge 1998). The discussions on "query-by-sketch" (Egenhofer 1997) are also noteworthy of this practical view.

4 Building New Tools to Model Change: An Engineering View

The previous section shows the theory on critical issues related to GIS-21 is still in flux. But we must move forward. Thus, in this section, we consider a concrete case: considering what we know today, how do we design a GIS-21 application to model environmental change?

This section describes briefly the major decisions on the design of TerraME, a tool for making models that combine society and nature (Carneiro 2006). Modelling the relations between the social and the natural environments is a hard task. It involves collecting data, building up a conceptual approach, implementing, simulating, calibrating, validating, and perhaps repeating one or more steps again. There is no proven scientific paradigm for human-environmental modelling. Different approaches exist in the literature, such as statistical modelling and agent-based modelling. After

considering what is there, TerraME designers decided to be as flexible as possible and to use sound advice whenever available. They made the following choices:

1. *Using a programming environment that supports higher-level functions:* As Andrew Frank has shown, generic higher-order functions are necessary for sound GIS type definitions (Frank 1999). Frank also argues that *functional programming* is a good basis for formal modelling of spatial data (Frank and Kuhn 1995). Following his advice, TerraME uses Lua, an open-source extensible scripting language that is simple and expressive (Ierusalimschy 1996). Lua's important advantage from other existing scripting languages (such as Python and Perl) is its support for functional programming and higher-order functions.
2. *Requiring Spatio-Temporal Database support:* TerraME has an interface to the TerraLib database environment. TerraLib provides many functions that are not part of the OGC standards, such as support for raster and spatio-temporal data (Câmara et al. 2008). TerraLib's spatio-temporal database design has been inspired by the ideas of Güting and co-authors (Güting and Schneider 2005; Güting et al. 2003).
3. *Designing a Nested-CA model:* TerraME uses a flexible, policy-free approach. Rather than choosing a single modelling technique (such as statistics or agent-based approach), TerraME provides a set of "building blocks" for model development. These "building blocks" include the ability to specify the spatial, temporal, and analytical components of the model separately. Thus, a large variety of approaches (and their combinations) can be expressed in TerraME.

5 A Problem and a Possible GIEngineering Solution: A Global Forest Information System

In this section, we briefly describe a challenging problem for GIEngineering and outline a possible solution that considers some of the issues raised in Section 3. We consider the problem of setting up a Global Forest Information System (GFIS), designed to enable knowledge sharing about forests. Its motivation is preserving the world's rain forests, one of the major environmental challenges of our generation. Rain forests are home to most the world's biodiversity, and play a major role in climate regulation and in the hydrological cycle.

Despite their richness and their ecological services, large areas of the world's rain forests are under increasing pressure of deforestation caused by human action. However, there is much doubt about the extent of

worldwide deforestation (Kintisch 2007). Ideally, all rainforest nations should produce detailed estimates and publish them on the Web, so there could be independent confirmations and concerted action. In practice, capacities differ substantially. Currently, Brazil is the only country that publishes detailed wall-to-wall maps of deforested areas in the Internet. Thus, a Global Forest Information System could help, by providing a web-based and cooperative approach that would allow countries, international organizations, NGOs and private companies to find, share, and produce information on the world's rain forests.

Designing a Global Forest Information System is a typical GIS-21 task. It needs a combination of tools that allow *reasoning about change*, provide *semantic information about the rain forests*, and *support cognitive navigation* over the world's tropical belt. The proposed GFIS design uses the Digital Earth metaphor, where geographical location is the common denominator. Diverse content such as satellite images, spatial data infrastructures, geobrowsers, research data, laws and policies, and citizen-provided information could be indexed, searched, discovered, and used by any interested parties. GFIS would enable people to interact based on their specific talents, interests, and experience. Thus, the GI engineer in charge of developing GFIS would need to adapt Web-based tools and techniques, such as social networking, content management, and mapping to the Digital Earth context. Fig. 1 presents this vision schematically.

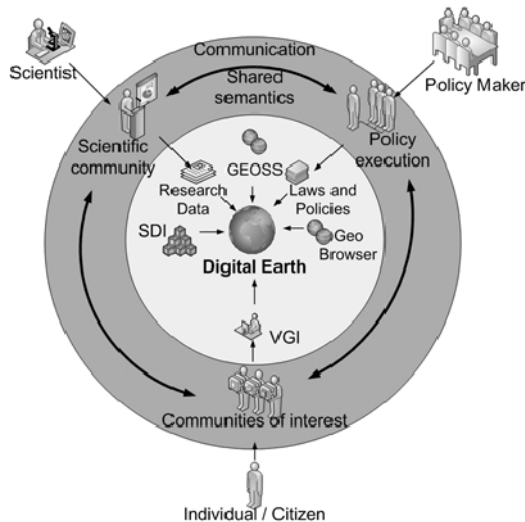


Fig. 1. The Global Forest Information System as a Digital Earth metaphor.

People and scientists from various disciplines often have multiple, and sometimes inconsistent, views on reality. This multiplicity brings a chal-

lenge for modelling. For instance, understanding deforestation requires a view of the problem from different perspectives: those of environmental experts, of policy makers, and of the common citizens. Each of them uses specific concepts, and treats problems and observations in different geographic scales, time granularities, and semantic categories. Thus, we propose the GFIS interface has three panels (see Fig. 2): a semantic representation view, a geographic information view and a document view. On the right-hand panel (documents), the application provides means to disseminate scientific data, laws and policies, and historical (baseline) data. On the central panel, we envisage geographic information (GI) as the glue among all other kinds of information. GI can be used to link scientific data and models to laws and policies, to blogs and independent reports. This way, GI might be able to connect information resources in unexpected and innovative ways. Navigation in the central panel should also consider multiple temporal and spatial scales. The user could have a global world view of a given year, or a local view of multi-temporal change.

On the left-hand panel (semantics), the application should also provide ways to improve understanding of rainforest conservation and monitoring. The user would be able to see the geographic information, browse documents related to it, and highlight main concepts. By navigating through these concepts, the user might ultimately learn about methods, expressed as workflows, which in turn link to executable models. Workflows are effective ways of communicating information about a data processing procedure. Using workflows would be a useful way for GFIS to show the differences between the different data processing tools it provides.



Fig. 2. Vision: GlobalForest would enable multiple perspectives.

By supporting multiple perspectives, GFIS would work a ‘learning space’ about the world’s rain forests. Navigation in the central panel

triggers change in the left panel (semantics) and right panel (documents). Consider that a user would navigate to the Brazilian Amazonia. The central panel would allow him to find different types of geospatial data about his region of interest. Using the semantic panel, he would select a topic of interest (e.g., manatee habitats) and the documents of the right panel would be automatically chosen to match the spatial region and the semantic topic. He could also select a model of manatee growth cycle from the semantic interface, run this model in the visualization interface, publish the information in the document interface and compare his results with those of other researchers.

Using GFIS, a developing nation could produce information about their rain forest. First, its specialists would look at the results and methods used by other countries that have similar characteristics. Data necessary for the inventory would be retrieved from the GFIS database. GFIS would offer the computer infrastructure to run the chosen methods remotely. Local researchers could interact with other experts using GFIS learning space, checking and improving the accuracy of their results. The results would be loaded back to GFIS' learning space. This would encourage neighbouring countries to use GFIS to do their national inventories. It will also increase the awareness of the problem in that region and support continuous monitoring.

6 Final Remarks: GI Engineers and GI Scientists Need to Cooperate

The preceding sections showed that much research is needed and, for GIS-21 tools, remains on a state of flux. Building the GIS-21 generation will be a tough job. The established paradigm of “mapping, spatial query and visualization” (used for 2D static data) does no longer capture the essence of the information. New technologies such as mobile sensors and new challenges such as modelling global environmental change need innovative solutions, directly tailored for the problem in question. GI engineers will not find references that provide a consistent and stable corpus of knowledge that allows them to concentrate on the technological challenges. This should be a cause for concern by both sides. GIScience will always be technologically motivated. Scientists use new tools as a source of inspiration for the next challenges. Should this innovation cycle slow down, then both sides stand to lose.

For the GI engineer, there is a lot to learn from the GIScience literature. For example, Andrew Frank's works are useful references for the GI engineer. His approach combines rigorous methods with a practical viewpoint,

which are the typical tools of good engineering. However, it is unrealistic to expect the GI engineer to find his way through the hundreds of papers of GIScience. The GI engineer will find no straightforward scientific solutions to tough problems in areas such as spatial cognition, semantics and change modelling. Addressing this challenge goes beyond the engineers' typical capabilities. It is up to the scientists to face the problem and to promote synthesis efforts that could help to build a stable basis for GIS-21.

Thus, the GIScience research agenda should consider the needs of the GI engineers of the 21st Century. Scientists need to develop GIEngineering into a field of research and teach it as a discipline. There should be a concerted effort to look at the current GIScience literature and identify those topics, which are relevant. By considering both directions of scientific-technological connection in spatial information, researchers and practitioners will both benefit from an increased dialogue.

Acknowledgments

The authors would like to thank Andrew Frank, whose research contributions have pointed to many areas, which are relevant to the GIE field. Frank has contributions in many areas, and this paper discusses only a limited part of his work. The authors also thank Werner Kuhn for many hours of inspired discussion, and the two anonymous reviewers for much useful guidance. Gilberto Camara's work is partially funded by CNPq (grant PQ 550250/2005-0) and FAPESP (grant 04/11012-0).

References

- Bittencourt O, Câmara G, Vinhas L, Mota J (2007) Rule-based Evolution of Typed Spatio-temporal Objects. In: Vinhas L, Costa AR (eds) IX Brazilian Symposium on Geoinformatics (GeoInfo 2007). INPE (ISBN 978-85-17-00036-2), Campos do Jordão, São Paulo, Brazil
- Brooks Jr. FP (1996) The computer scientist as toolsmith II. *Communications of the ACM* 39: 61–68
- Câmara G, Souza R, Freitas U, Garrido J (1996) SPRING: Integrating Remote Sensing and GIS with Object-Oriented Data Modelling. *Computers and Graphics* 15: 13–22
- Câmara G, Egenhofer M, Fonseca F, Monteiro AM (2001) What's In An Image? In: Montello D (ed) *Spatial Information Theory: Foundations of Geographic Information Science*. International Conference, COSIT 2001, Santa Barbara, CA. *Lecture Notes on Computer Science* 2205, Springer, Berlin Heidelberg New York, pp 474–487

- Câmara G, Vinhas L, Ferreira K, Queiroz G, Souza RCM, Monteiro AM, Carvalho MT, Casanova MA, Freitas UM (2008) TerraLib: An open-source GIS library for large-scale environmental and socio-economic applications. In: Hall B, Leahy M (eds) *Open Source Approaches to Spatial Data Handling*. Springer (ISBN 978-3-540-74830-4), Berlin Heidelberg New York, pp 247–270
- Carneiro T (2006) *Nested-CA: a foundation for multiscale modeling of land use and land change*. PhD Thesis in Computer Science (available at www.dpi.inpe.br/gilberto/teses/nested_ca.pdf). Computer Science Department, INPE, Sao Jose dos Campos
- Couclelis H (1992) People Manipulate Objects (but Cultivate Fields): Beyond the Raster-Vector Debate in GIS. In: Frank AU, Campari I, Formentini U (eds) *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*, Springer, Berlin Heidelberg New York, pp 65–77
- Dutton G (ed) (1978) *First International Advanced Study Symposium on Topological Data Structures for Geographic Information Systems*. Addison-Wesley, Reading, MA
- Egenhofer MJ (1997) Query Processing in Spatial-Query-by-Sketch. *Journal of Visual Languages and Computing* 8: 403–424
- Egenhofer M, Frank AU (1992) Object-Oriented Modeling for GIS. *Journal of the Urban and Regional Information Systems Association* 4: 3–19
- Egenhofer M, Franzosa R (1991) Point-Set Topological Spatial Relations. *International Journal of Geographical Information Systems (IJGIS)* 5: 161–174
- Egenhofer MJ, Mark DM (1995) Naive Geography. In: Frank AU, Kuhn W (eds) *Spatial Information Theory—A Theoretical Basis for GIS*, International Conference COSIT '95, Semmering, Austria. Springer, Berlin Heidelberg New York, pp 1–15
- Egenhofer MJ, Golledge RG (1998) *Spatial and temporal reasoning in geographic information systems*. Oxford University Press, New York
- Fonseca F, Egenhofer M, Agouris P, Camara G (2002) Using Ontologies for Integrated Geographic Information Systems. *Transactions in GIS* 6: 231–257
- Frank AU (1996) Qualitative Spatial Reasoning: Cardinal Directions as an Example. *International Journal of Geographical Information Science (IJGIS)* 10: 269–290
- Frank AU (1999) One Step up the Abstraction Ladder: Combining Algebras - From Functional Pieces to a Whole. In: Freksa C, Mark DM (eds) *COSIT - Conference on Spatial Information Theory. Lecture Notes in Computer Science* 1661. Springer, Berlin Heidelberg New York, pp 95–108
- Frank AU (2001) Tiers of ontology and consistency constraints in geographic information systems. *International Journal of Geographical Information Science (IJGIS)* 15: 667–678
- Frank AU (2003) Ontology for Spatio-temporal Databases. In: Koubarakis M, Sellis T, Frank AU, Grumbach S, Güting RH, Jensen CS, Lorentzos N, Manolopoulos Y, Nardelli E, Pernici B, Schek H-J, Scholl M, Theodoulidis B, Tryfona N (eds) *Spatio-Temporal Databases: The Chorochronos Approach*. Springer, Berlin Heidelberg New York: 9–78

- Frank AU (2007) Twenty years of reasoning with spatial relations. In: Fisher P (ed): *Classics from IJGIS*. CRC Press, Boca Raton, FL, pp 353–361
- Frank AU, Kuhn W (1995) Specifying Open GIS with Functional Languages. In: Egenhofer MJ, Herring J (eds) *Advances in Spatial Databases—4th International Symposium, SSD '95*, Portland, ME. Springer, Berlin Heidelberg New York, pp 184–195
- Freksa C (1991) Qualitative Spatial Reasoning. In: Mark DM, Frank AU (eds) *Cognitive and Linguistic Aspects of Geographic Space*. Kluwer Academic Press, Dordrecht, The Netherlands, 361–372
- Galton A (2004) Fields and Objects in Space, Time, and Space-time. *Spatial Cognition and Computation* 4: 39–68
- Golledge RG (ed) (1999) *Wayfinding Behavior: Cognitive Mapping and Other Spatial Processes*. Johns Hopkins University Press, Baltimore
- Goodchild M (1992a) Geographical Data Modeling. *Computers and Geosciences* 18: 401–408
- Goodchild M (1992b) Geographical Information Science. *International Journal of Geographical Information and Analysis* 6: 31–45
- Grenon P, Smith B (2003) SNAP and SPAN: Towards Dynamic Spatial Ontology. *Spatial Cognition & Computation* 4: 69–104
- Gütting RH, Schneider M (2005) *Moving Objects Databases*. Morgan Kaufmann, New York
- Gütting RH, Bohlen MH, Erwig M, Jensen CS, Lorentzos N, Nardelli E, Schneider M, Viqueira JRR (2003) Spatio-temporal Models and Languages: An Approach Based on Data Types. In: Koubarakis M (ed) *Spatio-Temporal Databases*, Springer, Berlin Heidelberg New York
- Hägerstrand T (1967) *Innovation Diffusion as a Spatial Process*. The University of Chicago Press, Chicago, IL
- Herring J (1992) TIGRIS: A Data Model for an Object-Oriented Geographic Information System. *Computers and Geosciences* 18: 443–452
- Ierusalimschy R, Figueiredo LH, Celes W (1996) Lua-an extensible extension language. *Software: Practice & Experience* 26: 635–652
- Kintisch E (2007) Carbon Emissions: Improved Monitoring of Rainforests Helps Pierce Haze of Deforestation. *Science* 316: 536–537
- Krieg-Brückner B, Shi H (2006) Orientation Calculi and Route Graphs: Towards Semantic Representations for Route Descriptions. In: Raubal M, Miller H, Frank AU, Goodchild MF (eds) *Fourth International Conference in Geographic Information Science (GIScience 2006)*, Münster, Germany, Lecture Notes in Computer Science 4197, Springer, Berlin Heidelberg New York
- Kuhn W, Frank AU (1991) A Formalization of Metaphors and Image-Schemas in User Interfaces. In: Mark DM, Frank AU (eds) *Cognitive and Linguistic Aspects of Geographic Space*. Kluwer Academic Publishers, Dordrecht, pp 419–434.
- Lakoff G, Johnson M (1980) *Metaphors We Live By*. University of Chicago Press, Chicago, IL
- Mark DM, Frank AU (1991) *Cognitive and Linguistic Aspects of Geographic Space*. Kluwer Academic Publishers, Dordrecht

- Medak D (2001) Lifestyles. In: Frank AU, Raper J, Cheylan J-P (eds) *Life and Motion of Socio-Economic Units*. ESF Series. Taylor & Francis, London
- Miller HJ (2003) What about people in geographic information science? *Computers, Environment and Urban Systems* 27: 447–453
- Montello D (2001) Spatial Cognition. In: Smelser NJ, Baltes PB (eds): *International Encyclopedia of the Social and Behavioral Sciences*. Pergamon Press, Oxford, 14771–14775
- Morehouse S (1992) The ARC/INFO Geographical Information System. *Computers & Geosciences* 18: 435–443
- Searle JR (1997) *The mystery of consciousness*. New York Review of Books, New York
- Silva MPS, Câmara G, Souza RCM, Valeriano DM, Escada MIS (2005) Mining Patterns of Change in Remote Sensing Image Databases. In: Han J, Wah B (eds) *The Fifth IEEE International Conference on Data Mining*. IEEE, Houston, USA
- Tomlinson R (ed) (1972) *Geographical Data Handling*. UNESCO/IGU, Ottawa, Canada
- Tversky B (1993) Cognitive Maps, Cognitive Collages, and Spatial Mental Models. In: Frank AU, Campari I (eds): *Spatial Information Theory: A Theoretical Basis for GIS, COSIT'93*, Elba, Italy
- Worboys M (1994) A Unified Model for Spatial and Temporal Information. *The Computer Journal* 37: 27–34
- Worboys M (2005) Event-oriented approaches to geographic phenomena. *International Journal of Geographical Information Science (IJGIS)* 19: 1–28

Towards Visual Summaries of Geographic Databases Based on Chorems

Robert Laurini

LIRIS, INSA de Lyon, F – 69621 Villeurbanne

Abstract

For many applications, it is very important to get an overview of database contents; and in the case of geographic database, a visual summary can be very helpful for a decision-maker or for any person in charge of a territory. To get such a visual summary, a two-step process is built, first a phase of spatial data mining process extracts geographic knowledge, and a second phase visualizes it by means of chorems – which can be defined as schematized representations of territory. In other words, semantic generalization must be followed by geographic generalization. The scope of this paper is to present the last results of an international project.

1 Introduction

For many decisions, visual tools are necessary, and especially for spatial decision making for which cartography is an essential tool. When it is the cartography of facts, usually decision-makers are satisfied, but when it deals with visualization of problems, conventional cartography is rather delusive: indeed it seems more interesting to locate problems and perhaps to help discover new problems or hidden problems. So the key-idea is to generate a global overview of the database contents, or more to summarize it.

A research program was launched between several research institutions (INSA-Lyon, France, University of Salerno, Italy, Tec de Monterrey,

Puebla, Mexico, etc.) in order to test whether cartographic solutions based on chorems¹ can be more satisfying. Invented by Brunet (1986, 1993), chorems can be defined as a schematized representation of territories. By schematized, one means that the more important is a sort of short global vision emphasizing salient aspects (Saint-Paul et al. 2005). This definition can be a good starting point to construct maps for spatial decision making. In other words, the goal of this research project is starting from existing databases, to analyze them so that to extract chorems by spatial data mining (Laurini et al. 2006) and visualize them.

This paper will be organized as follows. First chorems will be studied essentially as a new tool for visualizing and summarizing geographic information. Then the description of the architecture of a prototype system will be given.

2 What are Chorems?

2.1 From Conventional Cartography to Chorem Maps

As previously said, according to Brunet, chorems are a schematized representation of a territory. In the past, chorems were made manually by geographers, essentially because they had all the knowledge of the territory in their mind. This knowledge was essentially coming from their familiarity with the territory under study, its history, the climatic constraints and the main sociological and economic problems.

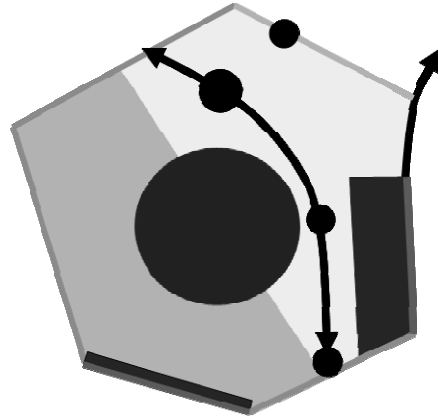
Fig. 1a shows a conventional map of France emphasizing administrative divisions whereas Figure 1b gives an example of a chorem map of France, in which the following aspects are stressed:

- the geometric shape is simplified,
- only big cities are mentioned (Paris, Lyon, Marseilles and Lille),
- only important mountains are shown, Alps as a frontier towards Italy, Pyrenees towards Spain, and the Massif Central forcing traffic to follow the Rhone river axis,

¹ Some people use the word “choremes” in English; but this is not correct taking into account usual rules for translating ancient Greek to English. Compare system, problem, etc. So, the word “chorem” is better and was selected.



(a) Conventional map of France



(b) A Chorem map of France

Fig. 1. Two maps of France.

- major traffic axes and seas are depicted,
- the blue lines show sea coastlines,
- and the French territory is divided in two parts, Eastern part the more developed, and Western part the less developed.

We claim that such a map is much more informative about the difficulties of France than a flat administrative map.

Another example refers to the water problem in Brazil. Indeed, a conventional map only showing main rivers as illustrated in Figure 2, does not lead to the solution of various problems such as:

- locations of places lacking water
- locations of the places with too much water
- locations of aquatic resources
- locations of humid zones
- locations of the water resources
- locations of the deserts,
- etc.

Bearing in mind all those examples, we claim that those chorem maps are much more informative and helpful to decision-makers. Those chorem maps can be seen both as the layout of geographic knowledge, and a kind of summary for geographic databases characterized by:

- a geographic generalization in order to simplify the shape of the territory under study,
- and a semantic generalization in order to select the more salient aspects of the non-spatial attributes of the geographic database.

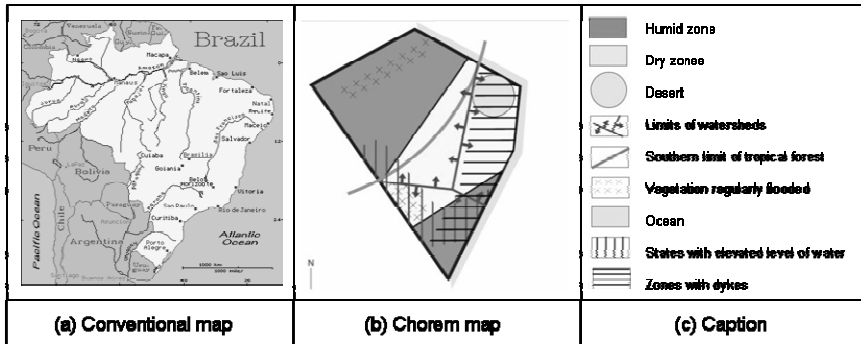


Fig. 2. The water problem in Brazil using a conventional river map (a) and a chorem map (b) issued from (Lafon et al. 2005).

However, some problems exist, especially due to their definitions:

1. For some people, chorems are too much simplified and do not reconstitute the complexity of a territory. By contrast, some chorematic maps can be very sophisticated (see for instance Peru's chorem in <http://flodemon.club.fr/choreme.htm>) which tries to represent several phenomena, but it is very difficult to understand or to explain.
2. When some boundaries are laid out, for instance between two zones; the reader must not forget that the lines corresponding to the boundaries are over-simplified or are approximated.
3. One of the major difficulties is to decide what the salient phenomena are and how to select them.

In order to overcome those limitations, our research program was based on the following assumptions:

1. The starting point will be an existing geographic database, not a so-called exhaustive knowledge of a territory under study;
2. The selection of important features will be based on spatial data mining;
3. Only a small subset of chorems will be used, not the entire Brunet's table.

Based on those assumptions, we do not want to re-make chorems of well-known territories, but rather explore some little-known database contents: in other words, we want to use our methodology only in domains which can bring some added value, such as:

- in geomarketing, when the CEO of an enterprise wants to have a global cartography of the places where his products are sold; and eventually to detect the anomalies in order to adapt a marketing strategy to sell more;

- in archeology, especially when spatial and spatio-temporal relationships must be exhibited and discovered;
- in sensor-based environmental monitoring and control, to rapidly discover anomalies, inconsistent sensor behavior and places where actions are need;
- in politics, especially after some elections to study the more salient aspects.

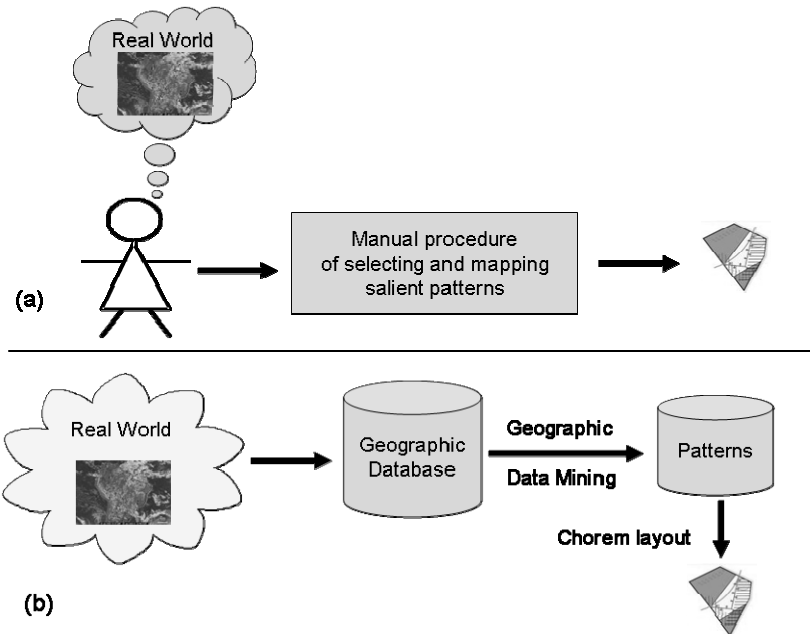


Fig. 3. Comparison of the conventional way of making chorems and our methodology. (a) the conventional way based on a partial vision of a territory; (b) based on a geographic database and geographic data mining for selecting the salient aspects.

Fig. 3 illustrates the difference between the conventional manual way of making chorems, and our methodology based on geographic databases and data mining.

Now the question is how to visually represent those salient aspects? Two possibilities exist:

- either to define a complete vocabulary (by means of icons) which can be used in any situation (this was Brunet's attitude when defining his chorems by means of a table),
- or to let the user define his own vocabulary by providing an ad-hoc caption.

2.2 Results of a Study of Existing Manually-Made Chorem Maps

A study was conducted about the chorems as they were used in several maps. Approximately 50 manually-made chorem maps were studied. The results are:

- even if the chorem concept is used by a lot of geographers, the Brunet's vocabulary is not often used;
- generally the users define their own chorem vocabulary,
- usually less than 10 chorems are used in a single chorematic map,
- the more used patterns can be regrouped into main categories such as (1) main cities (which can be retrieved by SQL SELECTs), (2) main regions which can be retrieved by clustering and (3) main flows which can be retrieved by both clustering and SELECTs.

2.3 Towards New Concepts for Geographic Databases

To conclude this section, it appears that chorems in addition to the initial definition (schematized representation of territories) can be potentially used for other goals such as:

- visually summarizing spatial database contents,
- global vision of a spatial database (Shneiderman 1997; Del Fatto et al. 2007),
- representing visual geographic knowledge,
- and new strategy for accessing spatial database.

As a chorem can be seen as a visual summary, some other layers of visual schematization can be defined from the database contents. So a sort of pyramid can be defined in which the apex is the chorem map, and the base is the database contents. At intermediate levels, several levels of geographic and semantic generalization can be defined. See Fig. 4 for such a pyramid.

To explore those new possibilities, some prototypes must be designed, implemented and tested. Let us examine a proposed architecture.

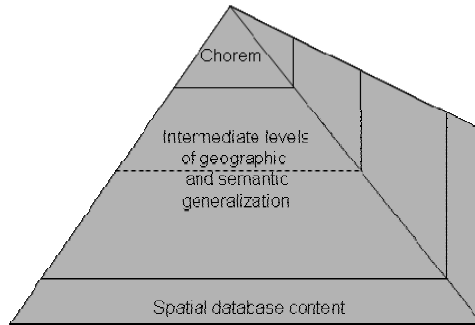


Fig. 4. A pyramid of contents.

3 Architecture of the System

An explorative system has been designed according to the main following specifications (Fig. 5):

1. chorem discovery based on spatial data mining, the result being a set of geographic patterns or geographic knowledge (upper part of Fig. 5),
2. chorem layout including geometric generalization, selection, algorithms for visualization (lower part of Fig. 5).

To facilitate spatial data mining and extract relevant semantics, a canonical database structure has to be defined. As an intermediary between chorem discovery and chorem layout, a language has been defined, named ChorML.

Another problem is the vicinity of the territory. Indeed, in several encountered manual chorem maps, external information must be added, such as the names of sea, of adjacent countries and so on. In order to provide this information, which is currently not in the initial database, a special table of the canonical database was defined. For instance, a canonical database (spatial and non-spatial) at country level will include:

- basic information such as cities, regions, main hydrology, main roads, mountains, etc.
- more elaborated information such as networks, flows, barriers
- external information such as boundary types, names of seas and of adjacent countries
- etc.

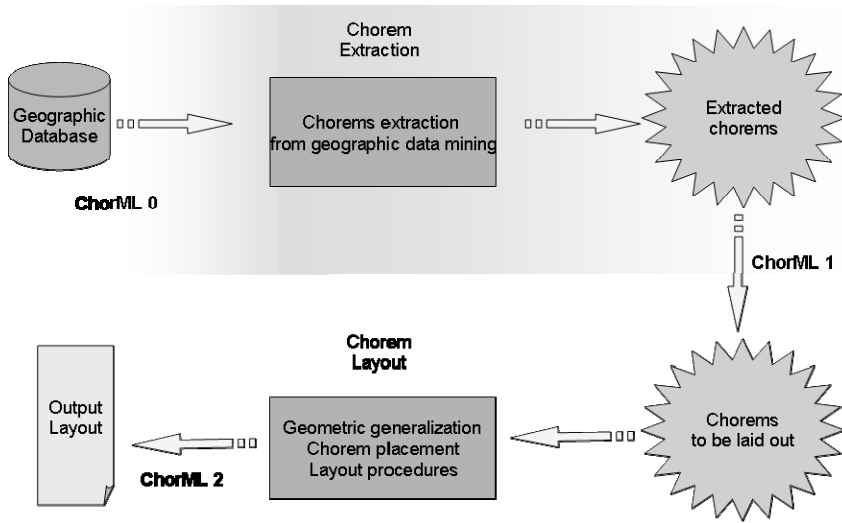


Fig. 5. Architecture of the system in two parts, the upper part corresponds to chorem extraction, and the lower part, chorem layout.

3.1 Spatial Pattern Discovery

As previously said, spatial patterns will be extracted from spatial data techniques. See (Ester et al. 1997) or (Pech et al. 2002) for details. However, in data mining it is well known that a lot of patterns can be retrieved. Two problems exist:

- setting of list of techniques to be used taking our context into account,
- selecting chorems from patterns.

So, among the relevant techniques, we have chosen to use first clustering and aggregation procedures together with SELECTs.

The next phase is how to identify chorems from spatial patterns, taking into consideration that a maximum of 10 chorems must be chosen. Those ten chorems must correspond to the more important spatial patterns. At this point, we have no clear-cut solution to reduce the number of patterns. In our first prototype we have decided not to implement an automatic solution: for that a visual interface will help the user choose the more important patterns (chorems) for the layout phase.

For the moment being, two kinds of spatial data mining techniques are used, those existing in ORACLE Spatial, and by using SUBDUE (Pech 2005).

3.2 Chorem Layout

Once the list of chorems and the set of constraints among them are obtained from the Chorem Extraction Subsystem, they are sent to the Visualization Subsystem in order to derive a visual representation of chorems and chorem maps, both in terms of layout and semantic content.

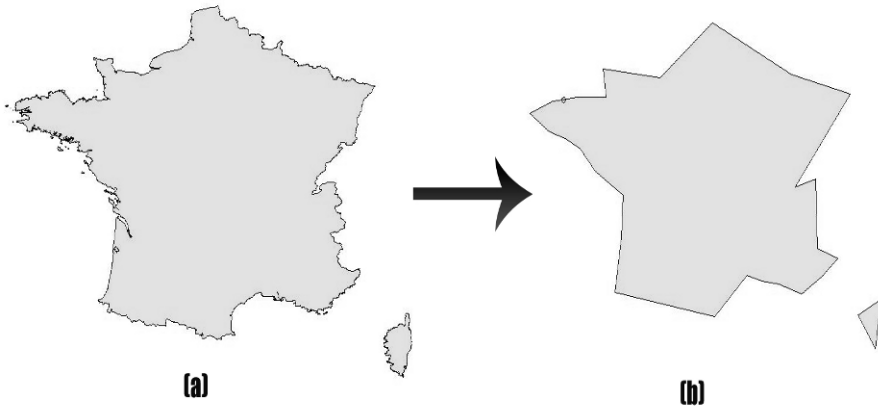


Fig. 6. An example the shape simplification process (generalization).

Five different tasks are performed by this subsystem, namely chorem drawing, coordinate translation, best-placement of chosen chorems, pre-layout computation and chorem editing. As for the chorem drawing, it is performed through three, not necessary interconnected, steps, named simplification, choremization and generalization, where some procedures and spatial operators are invoked. In Figure 6, such transformations are illustrated.

As for the generalization step, which is a well known set of techniques in cartography (Buttenfield and McMaster 1991), it may be invoked to group features that share some common properties, both geometric and descriptive, and generate a unique geometric representation of the involved elements. Fig. 9a and 9b depict such a transformation.

The choremization phase associates a regular shape with the possible simplified geometry of data.

One of the problems which may arise when simplifying and generalizing chorems, is related to the possible loss of crucial spatial constraints among elements of the original map. Thus, when the boundary is simplified, cities such as harbors which are located along the boundary must move with the boundary; otherwise, harbors would be positioned in the middle of the sea, or in the middle of the land. In order to preserve the spa-

tial consistency among geographic elements, topological constraints are checked and, if a violation occurs, the Visualization Subsystem modifies the city location, accordingly.

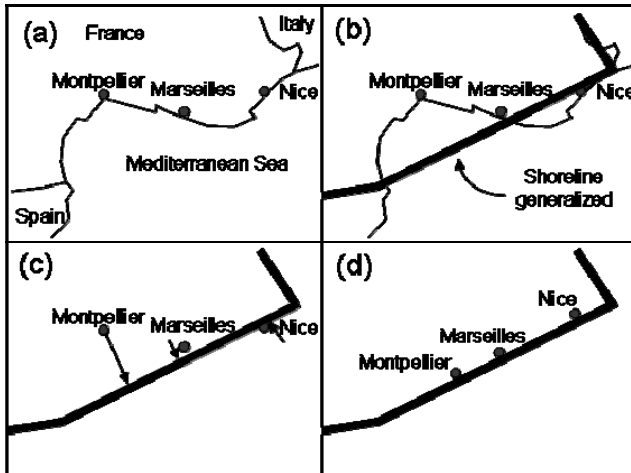


Fig. 7. Projecting harbors onto generalized shoreline. (a) situation before generalization. (b) generalized shoreline. (c) harbors must be moved. (d) final layout.

See Fig. 7 for an example along the French Mediterranean shoreline.

It is interesting to mention that as harbors must follow the topological relation “meet inside”, some places must follow “meet outside”; for instance consider the city of Geneva regarding France and generalized Eastern boundary.

It is worth to notice that in order to both preserve topological constraints and properly apply spatial operators, an underlying geographic reference system is maintained during the chorem drawing phase.

Once the drawing of the expected chorem is obtained, users are asked to specify details about the output map, such as the number of colours and the final layout format (for instance A4). The latter affects the number of choresms that can be introduced onto a map, since it is necessary to guarantee the readability requirement.

Based on the information provided by users, the next phase translates the chorem coordinates, acquired with respect to the original geographic reference system, into new coordinates defined with respect to a reference system local to the chosen visualization format.

At this stage, choresms extracted by the Chorem Extraction Subsystem are associated with a locally georeferenced visual representation. The goal of next step consists of aggregating choresms onto the output map. This is

accomplished by a multi-agent system whose aim is to spatially arrange chorems onto the chosen visualization format and determine their best placement (Jones 1989), preserving structural and topological constraints among them. It is worth to point out that in order to guarantee the best placement requirement, independent sets of interrelated chorems may be aggregated onto different maps, in order to provide users with more intuitive and readable chorem maps.

Anyway, some difficulties can occur regarding chorem placement and layout, as well as further refinements affecting semantic and graphic properties may be required by users. To this aim, users are provided with a tool for chorem editing which allows them to refine the expected output map.

In particular, the Chorem Editor performs the following tasks:

- import of a list of chorems positioned onto a chorem map;
- chorem display starting from the information derived from the previous steps;
- modification of both visual representation and semantic structure of chorems, without loss of consistency between them; in order to solve problems regarding chorem placement and layout the Chorem Editor can change chorem positions, colours and shape;
- generation of a graphical representation based on SVG (Scalable Vector Graphics) (SVG);
- export in graphical representation (SVG).

See Fig. 8 for an example of a visual summary after the study of an Italian database for population, and Fig. 9 for a ChorML excerpt.

3.3 ChorML

A language was designed to store chorems. Based on XML, ChorML presents several levels:

- level 0 corresponds to the initial database in GML (Geographic Markup Language) (GML);
- level 1 corresponds to the list of extracted patterns;
- level 2 is a subset of SVG (SVG).

For instance, at level 0, the feature coordinates can be longitude/latitude and feature attributes, whereas at level 1 the feature remains only if it belongs to a selected pattern, and finally at level 2, we deal with pixel coordinates, radius, line styles, colors and textures.

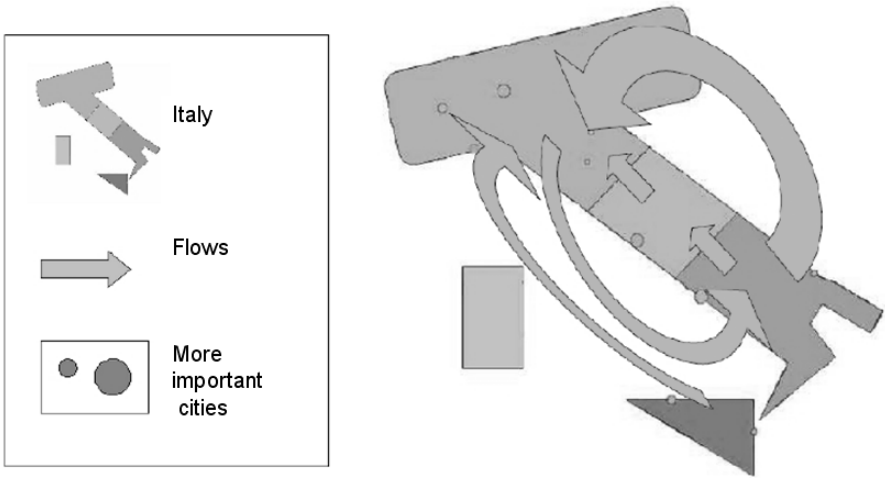


Fig. 8. Example of a visual summary from an Italian population database.

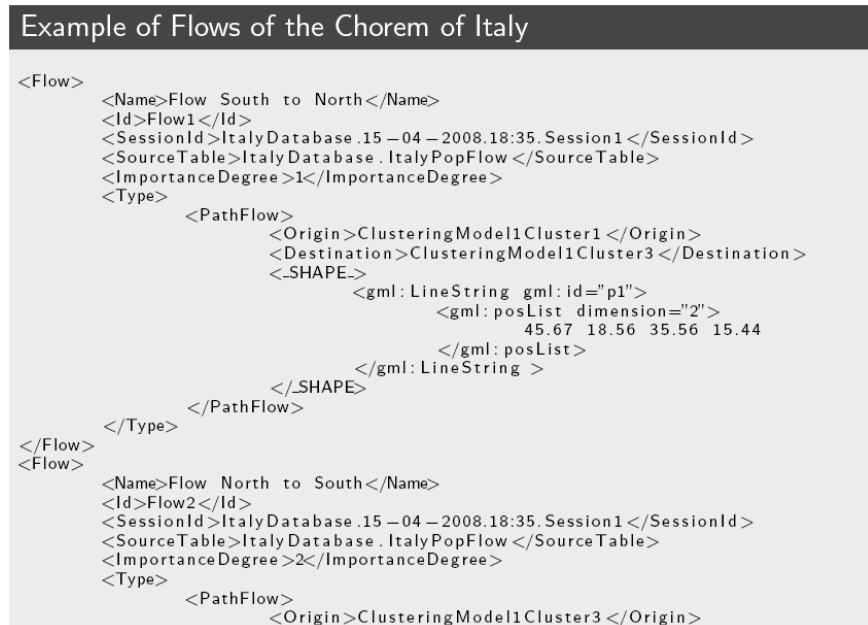


Fig. 9. Example of ChorML describing flows of Fig. 8.

3.3.1 Level 0

At level 0, the structure will include heading (database name, custodian, lineage, etc.) and database contents in GML.

3.3.2 Level 1

At level 1, the heading and additional information are practically not modified, but in place of the GML database contents, we have the list of patterns together with the way to obtain them (lineage). (Coimbra 2008) has shown that there are four kinds of patterns that result from data mining that appear to be the most interesting in chorem discovery:

- **facts**, for instance the name of a country capital,
- **clusters**, for instance any spatial regrouping of adjacent sub-territories,
- **flows** (one way or both ways)
- **co-location patterns**, especially to describe geographic knowledge; for instance “*when there is a lake and a road leading to that lake, there is a restaurant*”.

In addition to that, we need to include

- **topological constraints**, for instance that a harbor must be inside a territory, not in the middle of the sea,
- and **boundary description**, especially because outside information are usually not included in database, such as sea or neighboring country names.

Presently, this level is totally specified with an XML grammar.

3.3.3 Level 2

Once the patterns are selected, and the output format is known (for instance A4), we need to lay them out. At level 2, the selected patterns are now transformed into drawings encoded in SVG. This information is then sent to the chorem editor to finalize the result.

4 Final Remarks

The objective of this paper was to give some elements for the visual summarizing of spatial databases based on automatic discovery and layout of chorems. After a rapid analysis of existing manually-made chorems, some guidelines were exhibited, so that a prototype architecture can be proposed. This architecture can be split into two parts, the first one corres-

ponding to semantic simplification (chorem discovery) and the second to geometric simplification (chorem layout).

In the conventional way of designing chorems, the so-called “choremist” was supposed to have an exhaustive knowledge of the territory under study, to have clear-cut rules to decide what the salient phenomena are, and not to have problems to cartography them. Our assumption is that based on spatial data mining, the proposed methodology will both restrict the starting knowledge, and will provide a more rigorous approach to select the important features: if an important issue is missing in one chorematic map, it is because it is not included directly or indirectly within the database; not because the “choremist” wants to hide something important. Finally, gradually, we have changed the definition of chorems from “schematized visual representation of a territory” to “schematized representation of a geographic database” and to “visual summary of geographic databases”.

Regarding architecture, some modules have already been written and tested (for instance the chorem editor) whereas the specifications of the ChorML language and of the canonical database structure must be finalized. As shown in this paper, chorems look an interesting candidate to visualize geographic database summaries. Another interesting direction of research can be to use chorem as representation for geographic knowledge.

Even so our methodology can be applied to re-make very well known chorems, we claim that our methodology is much more interesting when studying little-known territories such as for geo-marketing, sensor-based GIS for environmental monitoring, archaeology, etc.

Present experimentations are processed based on ORACLE 11g using data from Italy. Next study will extract chorems from an historical database of the Mexican city of Puebla during the XVIIth and the XVIIIth centuries.

Among the difficult problems to solve, there are issues concerning spatial data mining, especially for getting the more important patterns. Presently, some difficulties appear to define mathematically “what is important” and to determine the good methodology to apply spatial data mining tools: so a new avenue for research must be outlined.

I have to thanks many persons for having collaborated to this project, namely, Dr. Françoise Raffort, Karla Lopez and André Coimbra from INSA-Lyon, Dr David Sol and Dr. Rosalva Loreto both from Puebla, Mexico, and Dr Giuliana Vitiello, Dr Monica Sebillio and Dr Vincenzo Del Fato from the University of Salerno, Italy.

References

- Brunet R (1986) La carte-modèle et les chorems. *Mappemonde* 86(4): 4–6
- Brunet R (1993) Les fondements scientifiques de la chorématique. In: La démarche chorématique, Centre d'Études Géographiques de l'Université de Picardie Jules Verne
- Buttenfield B, McMaster R, (1991) *Map Generalization: Making Rules for Knowledge Representation*, Longman, London
- Coimbra A (2008) *ChorML: XML Extension for Modeling Visual Summaries of Geographic Databases Based on Chorems*. Master Dissertation, INSA-Lyon, June 2008
- Del Fatto V (2009) *Visual Summaries of Geographic Databases by Chorems*. Ph Dissertation, April 2009, jointly-awarded by University of Salerno, Italy, and INSA-Lyon, France
- Del Fatto V, Laurini R, Lopez K, Loreto R, Milleret-Raffort F, Sebillio M, Sol-Martinez D, Vitiello G (2007) Potentialities of Chorems as Visual Summaries of Spatial Databases Contents. In: Qiu G, Leung C, Xue X-Y, Laurini R (eds) *VISUAL 2007, 9th Int'l Conference on Visual Information Systems*, Shanghai, China, 28-29 June 2007, Springer LNCS, Volume 4781 *Advances in Visual Information Systems*, pp 537–548
- Ester M, Kriegel H-P, Sander J (1997) *Spatial Data Mining: A Database Approach*. In: *Proc. of the Fifth Int. Symposium on Large Spatial Databases (SSD '97)*, Berlin, Germany, *Lecture Notes in Computer Science*, Vol. 1262, Springer, pp 47–66
- GML: <http://www.opengis.net/gml/>
- Holder LB, Cook DJ, (2005) *Graph-based Data Mining*. In: Wang J (ed) *Encyc. of Data Warehousing and Mining*, Idea Group Publishing
- Jones CB (1989) Cartographic Name Placement with Prolog. *IEEE Computer Graphics and Applications* 9(5): 36–47
- Lafon B, Codemard C, Lafon F (2005) *Essai de chorème sur la thématique de l'eau au Brésil*, <http://histoire-geographie.ac-bordeaux.fr/espaceleve/bresil/eau/eau.htm>
- Laurini R, Milleret-Raffort F, Lopez K (2006) *A Primer of Geographic Databases Based on Chorems*. In: *Proceedings of the SebGIS Conference*, Montpellier, Springer, LNCS 4278, pp 1693–1702
- Pech Palacio M (2005) *Spatial Data Modeling and Mining using a Graph-based Representation*. Ph Dissertation, December 2005, jointly-awarded by Universidad de las Américas, Mexico, and INSA-Lyon, France
- Pech Palacio M, Sol Martinez D, González J (2002) *Adaptation and Use of Spatial and Non-Spatial Data Mining*. In: *Proceeding of International Workshop Semantic Processing of Spatial Data (GEOPRO 2002)*, Instituto Politécnico Nacional, México, December
- Shneiderman B (1997) *Designing the User Interface*, Third edition, Addison-Wesley Publishing Company
- SVG: <http://www.w3.org/Graphics/SVG>.

Saint-Paul R, Raschia G, Mouaddib N (2005) General Purpose Database Summarization. In: Int. Conf. on Very Large Databases (VLDB 2005), Trondheim, Norway, Morgan Kaufmann Publishers, pp 733–744.

Intelligent Spatial Communication

Stephan Winter, Yunhui Wu

Department of Geomatics, The University of Melbourne, Victoria 3010, Australia, winter@unimelb.edu.au, y.wu21@pgrad.unimelb.edu.au

Abstract

People can give better route advice to a wayfinder than current navigation services due to many reasons, among them their more compatible spatial conceptualizations based on a common embodied experience of space, and their situatedness during communication, enabling them to make inferences to capture and adapt to context. So, how far are current navigation services from imitating humans in giving route directions? And what can we learn from this question, in terms of need for further research to build more intelligent services? This paper aims for a systematic framework to develop a research agenda for services to give better route directions. The framework is developed from various perspectives on human wayfinding communication.

1 Introduction

Consider the simple question of a person: “Can you tell me the way to ...”, which leads to typical everyday communication either with other persons or with a computing machine in form of a dedicated navigation service. However, current navigation services are frequently criticized for not adapting to the user’s needs and language (e.g., Timpf 2002; Pontikakis 2006).

Recent progress in technology has evolved along two directions. One is towards dynamic proliferation of more content, such as real-time traffic

data, points of interest, or find-and-recommend services. The other is towards more complex interfaces, such as perspective views, 3D, textures, or multi-modal information. These developments seem to counteract easing the cognitive workload of the user, and hence, the question arises whether more intelligent services can evolve at all from these directions of current development. With other words, does research and development need a correction of perspective? This chapter sets out to study what makes a truly intelligent navigation service.

Turing, laying the ground for what later became known as artificial intelligence, starts his landmark paper *Computing Machinery and Intelligence* with the words: “I propose to consider the question, ‘Can machines think?’” (Turing 1950, p 1). This question seems appropriate in the current context, where we are seeking what can make navigation services intelligent. In this sense we take the liberty to restrict Turing’s question by asking “Can machines think spatially?”.

Already Turing himself was aware of the problem of defining thinking or intelligence. He came up with an elegant suggestion: an anthropomorphic imitation game, which was later called the *Turing test*. In this game persons at a teletype interface are supposed to find out whether they are communicating with a machine or another person. Turing equaled anthropomorphic communication behavior with being intelligent. If the player cannot distinguish between machine and person the machine passes the Turing test.

We may borrow from this idea. Translated into our context, a machine can show intelligent spatial communication behavior if persons, requesting some route information from the machine, cannot find out whether they are communicating with a navigation service or with another person. Accepting modern forms of human computer interaction, and extending Turing’s rules of communication, we allow for graphical (Egenhofer 1997; Agrawala and Stolte 2001) and gestural (Kopp and Wachsmuth 2004; Cassell et al. 2007; Roth 2007) communication interfaces, since people may describe routes graphically and by gestures as well. We also allow for mobile devices that can be taken into the wayfinder’s decision situation, which enables, in principle, to catch the communication context.

Contrary to Turing’s own expectations the machine has not yet passed the Turing test. Even restricted versions are still a challenge. Others have limited the scope of the Turing test before. For instance, the Loebner Prize¹ is awarded annually to a program that passes a Turing test *of limited scope and tenor*. In our context, the *scope* would be restricted to the domain of orientation and wayfinding. We do not require a navigation service to be intelligent about other domains, let us say, football or food. Also, the *tenor*

¹ <http://www.loebner.net/Prizef/loebner-prize.html>

would be restricted to a natural discourse in this domain. We do not expect navigation services to cope with attempts to be outwitted or with comments out of context.

The communication behavior of a navigation service can be tested for the *information content*, either requested or conveyed, i.e., whether it is talking about appropriate routes, and for the *form of communication*, i.e., whether it is understanding the user's spatial descriptions and is responding in terms and references a person would choose (Allen 1997). But even a service that behaves reasonably well content-wise and language-wise on standard requests requires flexibility to be able to follow the course of a natural conversation on orientation and wayfinding. Persons may come with a variety of requests such as for more detail on a route, for alternative routes, for confirmation of their understanding of a route description, for comparisons or assessments, for clarifications of perceived inconsistencies, or for context-dependent additional information such as fares or kinds of tickets.

Spatial communication fails to be intelligent every time when an anthropomorphic quality in a service's communication behavior is detected missing. Since the total number of missing qualities cannot be determined in advance, it is impossible to prove that a machine can behave in their spatial communication like a person. But each closure of an identified gap forms a refutation of the hypothesis that machines cannot behave like a person – until the next gap is detected.

Looking at intelligent spatial communication this way, it provides a vision of an intelligent navigation service, something that current progress of technology is lacking. The aim of this paper is to establish a formal framework that will enable us to study the state of the art, and especially the gaps towards intelligent navigation services.

This paper starts with a review of the Turing test. It then lays out a framework to analyze intelligent spatial communication by studying the human wayfinding communication process from different perspectives. The first perspective looks at the phases of the communication. The second categorizes the elements of the spatiotemporal context of a wayfinding communication. The third perspective is taken by identifying characteristics of an intelligent agent for the communication partner of the wayfinder. The paper will conclude with a summary and outlook.

2 A Criterion for Intelligent Spatial Communication

When Turing (1950) suggested the anthropomorphic imitation game he was interested in finding a simple operational definition of intelligence. Nevertheless, his suggestion of the game (now called *Turing test*) sparked

an ongoing controversy in artificial intelligence and beyond². This controversy entwines around the notion of *thinking* or *intelligence*.

So, is it *intelligent* if the computer imitates a person successfully? Philosophers, for example Searle, insist that thinking requires a mind and consciousness. Searle's Chinese Room experiment (1980) basically says that a computer programmed to do a task (here: understanding Chinese) could also be replaced by a person running this computer program by hand. In his example the task is translating a Chinese text into English. As this person does not understand Chinese, nor does the computer. With other words, computers are mindless; they manipulate symbols in an order they were programmed. Already Lady Lovelace (1815–1852) realized that machines can only do what we have the skill to tell them to do (after Turing 1950). Even though a program adds abilities and programmers' knowledge to a computer, potentially including an ability to learn and hence to act in ways not predictable by their programmers, this teaching of abilities and knowledge only means a computer can be appropriately programmed to pass the Turing test, without a chance to claim having consciousness or a mind. Searle calls the former *weak* AI, and the latter *strong* AI, linking only strong AI to thinking and intelligence. Obviously the Turing test relies only on the communication behavior, i.e., the cognitive and linguistic performance capacities of a computer. It does not require to look like or to internally function as a human. Accordingly, we will abstain in the following from using the word thinking, and render our expectations more precisely as an imitation of a person's spatial communication behavior. This means we call a service intelligent if it appears to behave in its spatial communication like an intelligent agent: a person.

But the initial question can also be phrased slightly different. People may ask whether the computer is not more intelligent than a person anyway, so why bother with imitating? Where this question comes up, the objectivity of a computer and its large and accurate data sets seems to be able to generate more trust in spatial advice than a fellow citizen. However, this question shifts the focus from intelligent behavior to behavior superior to the human mind. The computer is superior to the human mind in at least two ways:

- A persistent and large memory enables a computer theoretically to access a complete and accurate travel network data set for route computation. This data set can be even kept up-to-date in real time by distri-

² French (2000) observed that Turing's paper became the most discussed paper in artificial intelligence, and Crockett (1994, p 1) notices: "Andersons's 1964 anthology, *Minds and Machines*, places Turing's paper first, perhaps following the ancient Semitic practice of placing the most important literature in a collection first".

buted sensors. A person is always bound to knowledge acquired by experience over time. This knowledge is subjective, selective, and historic.

- Algorithms to compute optimal routes can be shown to be theoretically correct, i.e., if an optimal solution exists at a particular time such an algorithm will find it, although it may take a long time. A person is bound to distorted cognitive spatial representations (Stevens and Coupe 1978), and human route selection is habitual and applies heuristics that potentially lead to suboptimal routes (Golledge 1999).

Now, superiority, once detected by a human communication partner, leads to failure in the Turing test. With other words, in Turing's sense it is not considered to be intelligent. Although this conclusion may surprise, there are arguments why an intelligent navigation service should not demonstrate its superiority. These arguments are based on cognitive costs, as will be discussed in the following.

First, in an inherently uncertain world it has an advantage why people do not (always) select the route optimal according to a cost function. Their selection is based on heuristics. Gigerenzer (2007) calls such heuristics convincingly the intelligence of the unconscious, referring to the delusion of finding an optimum in an uncertain world. Even a computer is limited in finding an optimum route facing the unpredictability of travel times in the future. People may favor simple paths or familiar ones and by this way ease their wayfinding process, including the communication of the route.

Secondly, in an inherently complex world it has another advantage to apply heuristics. A computer may take longer to solve some routing tasks exactly than it takes a person to find an acceptable suboptimal solution (Dry et al. 2006; Applegate et al. 2007).

Thirdly, short term memory is limited. Communicating by maps, metric information, perspective views or virtual reality animations – showing as much detail as possible, as many navigation services do – comes at costs. The wayfinder has to make special cognitive efforts to understand and realize these route descriptions. Map reading is known to be a complex task, and maps provide information about a whole area, i.e., far more information than required or expected for a route description. Metric information is difficult to realize for a human and has issues with granularity. Views and virtual reality animations are different from the embodied experience of the wayfinder in perspective, detail, light and street life. Thus it has an advantage when people sketch routes verbally or graphically, concentrating on essential and relevant route properties and relations to the environment. Grice's conversational maxim of relevance comes into play (Grice 1989), and Frank has shown that from the perspective of pragmatics longer route descriptions are not necessarily leading to better wayfinding processes (Frank 2003).

Last, but not least it is advantageous when people communicate routes by referring to cognitively salient features or properties, in contrast to references to travel network segments and nodes, the navigation services' primary data resource (Denis et al. 2007; Klippel et al. 2009; Tenbrink and Winter 2009). Route descriptions from people are more memorable and typically shorter.

In summary here is an argument that people are superior to computers in (in principle being able to) choosing more *appropriate* routes as well as choosing more appropriate route descriptions. Even more, they do this relatively effortlessly, and we are far from knowing how to tell the computer to replicate such skills. While we argue here by principle, it is not claimed that any human route description is per se more appropriate. People can choose routes or route descriptions that fail, that are far from any optimum, or that are ambiguous.

If we agree that human spatial communication behavior is intelligent, then it is desirable to design navigation services capable of such intelligent spatial communication behavior. Accordingly, one final question has to be answered: Can a computer ever successfully imitate a person?

Crockett (1994) approaches this problem referring to the frame problem. The frame problem is the problem, given a dynamic world, of how to limit axiom revisions in logical systems (McCarthy and Hayes 1969), or more generally of how to limit the updates of beliefs about the world given that the world changes or we interact with the world (Pylyshyn 1987). Assuming that a system knows the states of the world at a time t_0 , then at time t_1 changes have occurred. Some of them may be known to the system and can be introduced by axiom revision, but what about the other states? The frame problem is especially relevant in a dynamic domain such as way-finding. For example, when agents have moved, their location can be updated. But at the same time the state of mind of the agents may have changed, the state of the traveling network may have changed, and costs of traveling may have changed.

Crockett points out that a computer, to pass the Turing test, has first to solve the frame problem. He argues that, since a solution of the frame problem is not in sight, the answer to this final question is negative: a computer most likely will never successfully imitate a person. All this can only mean that a requirements analysis for an intelligent navigation service is always tentative. An intelligent navigation service can only be approximated. Ideas to design such a spatial Turing test are laid out elsewhere (Winter 2009).

3 A Framework for the Requirements Analysis

In this section a systematic framework to study the desired characteristics of an intelligent navigation service is developed. We will concentrate on wayfinding scenarios; accordingly the subsections will combine three independent approaches to wayfinding communication. One is about the phases of wayfinding communication and the individual tasks of the communication partners during these phases. The second is about the spatio-temporal context of wayfinding communication, and the extent to which it is considered by the communication partners. The third approach is about the characteristics of an intelligent autonomous agent to be able to imitate their communication behavior.

3.1 The Phases of Wayfinding Communication and Their Tasks

Klein (1982) and Wunderlich and Reinelt (1982) have identified four phases in the human wayfinding communication process:

1. the *initial phase*: a wayfinder asks an informant for directions,
2. the *center phase*: the informant provides route directions,
3. the *securing phase*: either the wayfinder or the informant want to make sure that the wayfinder has understood the given route directions, and
4. the *closing phase* of closure and separation.

Nearly all research so far focuses on the center phase, studying either route directions provided by people (e.g., Klein 1979; Denis 1997), or studying how route directions can be generated automatically (e.g., Dale et al. 2005; Richter 2008). This is the only necessary phase; the other ones are optional. An extreme example might be the printed travel guide, providing a route description to the reader without a specification of a request by the reader, without a securing phase, and without a closing phase other than that the reader closes the guide book or turns the page, i.e., averts his attention.

At each stage of the communication Klein (1982) as well as Wunderlich and Reinelt (1982, p 183) identify three subtasks present:

1. a *cognitive* task (e.g., activating a spatial cognitive representation);
2. an *interactional* task (e.g., initiating and terminating the verbal exchange, or providing a route description);
3. a *linguistic* task (e.g., expressing a comprehensible route description).

Interaction between the cognitive and the interactional task, in the context of a machine as informant, includes not only the cognitive abilities of the

wayfinder, but also the internal data models and algorithms of the navigation service. Between these tasks the focus is on identifying and modeling the references that have to be conveyed and understood (the content). The third task focuses on their actual representation in a specific sign system (the language). Wayfinding communication can use multiple sign systems (verbal, gesture, graphics), which may all vary between different cultures.

3.1.1 The Initial Phase

In the initial phase the wayfinder has the lead role and talks to the route service. According to Klein (1982, p 168), the initial phase consists of three subtasks for the wayfinder:

- getting into contact with the informant;
- making clear what he wants;
- succeeding in getting the informant to take over the task of giving him route directions.

Neither Klein nor Wunderlich and Reinelt (p 183) went on to study the initial phase in detail. However, we identify the three subtasks:

- A cognitive task. The wayfinder has to find a proper specification for his route request, which means a specification based on their own spatial cognitive representation that is sufficient for the informant in the given communication context. Vice versa, informants have to activate their spatial cognitive representation to identify the specification of the wayfinder.
- An interactional task. The wayfinder has to manage to get into contact, to specify a route, and to convey his request. The informant has to pay attention, listen, and respond by confirming that the specification of the route was received and sufficient.
- A linguistic task. Wayfinder and informant interact via sign systems, and all three subtasks of the initial phase have to be (a) expressed in one of these sign systems and (b) understood by the recipient. The wayfinder has to be ensured via signs that the informant took over.

An example for the initial communication phase with a navigation system, a web-based public transport planner (Einert 2006), is shown in Fig. 1. A more thorough discussion of this example can be found elsewhere (Winter and Wu 2008).

3.1.2 The Center Phase

The phase of giving route directions is initiated and terminated by the informant (Wunderlich and Reinelt, 1982, p 187). In this phase again we can identify:

- A cognitive task. The informant – e.g., the navigation service – has to interpret correctly the specification of a route request by the wayfinder.
- An interactional task. The informant has to plan a route according to the specification.
- A linguistic task. The informant has to express the route in a comprehensible manner, verbally and/or graphically.

Fig. 1. The initial communication phase with Metlink’s Journey Planner (snapshot from June 2008, © Metlink).

Fig. 2 shows a route description of the web-based public transport planner of Fig. 1. A more thorough discussion of the cognitive, interactional and linguistic tasks of this kind of route descriptions can be found elsewhere (Tenbrink and Winter 2009).

Travel by	Time	Details	Map	Information
	DEP: Sun, 10:58 am	From Stop 11-University of Melbourne/Royal Pde (Parkville)		
		Take the Route 19 tram towards City (Elizabeth/Flinders St)		Time 13 min Frequency: 6 min Zone(s): CitySaver Operator: Yarra Trams Leo Timetable
	ARR: Sun, 11:11 am	Get off at stop 1-Flinders Street Railway Station/Elizabeth St (Melbourne City)		

Fig. 2. Route directions given by Metlink’s Journey Planner, upon a request for a trip from *University / Royal Parade* (a stop name) to *Flinders Street Railway Station* (a stop name) departing earliest at 10:50am on 27 July 2008 (© Metlink).

3.1.3 The Securing Phase

Wunderlich and Reinelt (1982) report a large variety of communication patterns in the securing phase between people. They can consist for example of summaries, repeats, paraphrases, more detailed descriptions of crucial parts, additional information for the decisions points along the route, or a discussion of alternatives. Corresponding to this diversity we identify a variety of cognitive, interactional and linguistic tasks, some of them assigned to the wayfinder, some to the informant. However, they basically repeat the initial and center phase: expression of a question, understanding, acting on a response (e.g., modifying, generalizing or precisifying the plan), and conveying the response.

Aspects of a securing phase are present in Fig. 2. A wayfinder can click on the hyperlinks in the verbal route descriptions to get more information on the stops, and also stop maps and leg timetables can be requested. Further buttons provide options for re-enquiry (*Modify*, *Search again*, *Return journey* and *Onward journey*). The securing phase is terminated as soon as the wayfinder initiates the closing phase.

3.1.4 The Closing Phase

As Wunderlich and Reinelt (1982) remark, “only . can state that the request has been satisfactorily fulfilled” (p 188). A typical initiation of the closing phase is an expression of gratitude, and termination is made by turning away. In this phase we can identify:

- A cognitive task. The wayfinder determines that he is satisfied with the given information.
- An interactional task. While the wayfinder’s attention moves to realize the given information, the informant can deactivate his cognitive map and return to conventional communication or other tasks.
- A linguistic task. The wayfinder should indicate that he is satisfied.

Our example of the web-based public transport planner gives the wayfinder the opportunity for giving feedback, as an expression of gratitude, and for printing the directions (also indicating satisfaction). Further a wayfinder can follow some links to external webpages, or they can simply close the web client and turn away from the machine, be it a mobile device, a terminal, or a desktop computer.

3.2 The Spatiotemporal Context of Wayfinding Communication

Communication with a navigation service takes place in a spatiotemporal context. To capture and categorize this context let us refer to Janelle

(2004), who studied spatial and temporal communication constraints between communication partners given the diverse range of communication channels. His categories concern (see also Table 1):

- location of the communicators: *physical co-presence* or *telepresence*
- time of the communication: *synchronous* or *asynchronous*

Compared to Janelle's two dimensions for a general communication context, a wayfinding communication is coming with two other context dimensions (called indexes or deixis in pragmatics, see Suchman 1987):

- location of departure: *from here* or *from elsewhere*
- time of departure: *now* or *in future*

Table 1. Janelle's spatial and temporal communication constraints (2004) applied on seeking route advice.

	synchronous	asynchronous
physical co-presence	e.g., face-to-face, or from mobile location-aware device	e.g., from you-are-here maps, or departure plans at bus stops
telepresence	e.g., via telephone or from web service	e.g., departure plan from a web page

With this categorization at hand, one can distinguish the communication context for different navigation services. For example, services on mobile devices – such as location-based services, car navigation services, or tourist guides – establish a context characterized by the quadruple $\{physical\ co\text{-}presence, synchronous\ communication, from\ here, now\}$. They can infer the meaning of *from here* by mobile positioning. In comparison, services provided on the web for in-advance trip planning establish a context characterized by the quadruple $\{telepresence, (quasi\text{-})\ synchronous\ communication, from\ anywhere, anytime\}$. Especially web-based navigation services have no clue to distinguish between wayfinders seeking advice for immediate departure, i.e., from their current location, and wayfinders seeking advice for any time in the future, i.e., from possibly another than their current location. Nevertheless, web-based navigation services typically pre-fill the departure time with the actual time as a default (Fig. 1). They do not yet use positioning technologies to pre-fill the departure location.

3.3 Representing an Intelligent Agent in Wayfinding Communication

A service has to understand a person's wayfinding request and has to respond as another person would do. This was called intelligent communication behavior. For artificial intelligence Brooks (1991) has identified three

characteristics of an intelligent agent: being able to cope with situatedness, embodiment and emergence. To be precise, Brooks lists a fourth property, intelligence. For him, intelligence shows in the complexity of behavior “determined by the dynamics of interaction with the world” (p 584). In the present paper, however, the intelligent agent – the navigation service – is not itself physically autonomous in the world, but communicates to an agent that is situated, embodied and capable to cope with emergence: the wayfinder. Furthermore, the service is supposed to communicate like a person, who has all these abilities. Accordingly, in our case intelligence does not appear in the complexity of any physical behavior of the service, but in the complexity of its communication behavior. This means the service requires an awareness of situatedness, embodiment and emergence to be able to give route advice like a person. This argument is still in line with Brooks’s (1991) argument for a bottom-up emergence of intelligent behavior. It is also in line with current human computer interaction paradigms. For example, Dourish (2001) – “dialog is central to our notion of interaction with the computer” (p 10) – identifies embodiment as the common ground and challenge for human computer interaction (p 22).

Hence, an intelligent navigation service’s communication behavior should be:

- **Situated.** It should be aware of the context of the communication situation. Beyond the spatiotemporal context discussed above (section 4), this includes perception and an awareness of the environment of the current location of the wayfinder, and what it offers and affords. Our example (Fig. 1 and 2) is poor of situatedness except the pre-fill of departure time.
- **Embodied.** It should be aware of the human capabilities to move in an environment, and their commonsense, or naive understanding of the world (Egenhofer and Mark 1995). The particular person and its abilities and preferences can be taken into account. With respect to content of advice, this concerns concepts of mobility such as comfort or convenience, costs, risk or trust. And with respect to language, this concerns proficiency in relative, qualitative and egocentric spatial concepts, which also enable to uncertainty. Our example (Fig. 1 and 2) is tuned for public transport users, but not flexible enough to adapt to other means of transport or more specialized individual requirements.
- **Aware of emergence.** It should be aware of the coherent cognitive structures of the wayfinder that have evolved during the process of learning spatial environments. This concerns their procedural and declarative spatial knowledge, in particular the hierarchic organization of spatial cognitive representations. Our example (Fig. 1 and 2) does not show any

consideration of previous knowledge of the wayfinder, but it has a hierarchic approach of releasing more details on request.

4 Conclusions

This paper suggests a criterion for intelligent navigation services: a Turing test of limited scope and tenor. The test is limited to the domain of giving route advice, and limited to the tenor of a natural wayfinding discourse. In accordance with the original Turing test, the communication channel may still be teletype (text, such as in web forms for the wayfinder, and in verbal descriptions by the web service), but a variety of other channels exists as well, such as speech, graphical interfaces, and visual interfaces to cope with facial expressions and gestures. A navigation service will pass this test if it behaves in its spatial communication like a person. This criterion forms a vision and also a benchmark for the directions of further technological developments in this area (Winter 2009).

Since this criterion requires imitating human communication behavior, it is then used as a motivation to study the desired characteristics of an intelligent spatial communication. A framework is presented of characteristics derived from combining three independent approaches to study the human wayfinding communication process: the phases of wayfinding communication, the spatiotemporal context of wayfinding communication, and the characteristics of an intelligent autonomous agent. Since these approaches are sufficiently orthogonal they can be intersected. Fig. 3 illustrates this for two of the three approaches, communication phases and tasks in each phase (section 3.1) and situation awareness (section 3.3); spatiotemporal context (section 3.2) forms the fourth dimension.

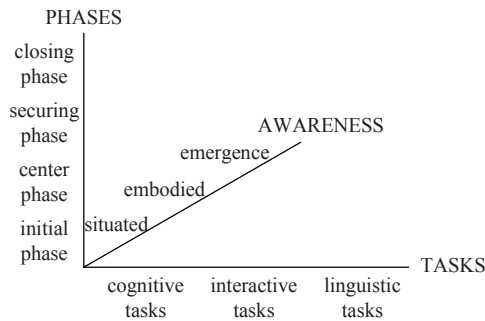


Fig. 3. The framework built by the phases of wayfinding communication with their individual subtasks, crossed with the properties of an intelligent agent. Missing in this graph: the spatiotemporal context, forming a fourth dimension.

From the discussion in section 2 we can conclude that any framework is insufficient to facilitate designing a service guaranteed to pass the test to be called intelligent. Nevertheless, in our framework we could find reasons for each approach to be included in a benchmark, among them their orthogonality.

The four dimensions of this framework have already shown to be useful in a first investigation of the initial phase of the communication process identifying requirements and needs for further research (Winter and Wu 2008). It is expected that in the future the framework can be used to study systematically the structure of existing knowledge, gaps of knowledge, and requirements for research of the other phases as well. Also existing navigation services can be investigated by this way.

Acknowledgements

This work is supported under the Australian Research Council's Discovery Projects funding scheme (project number 0878119). Anonymous reviewers gave valuable comments on an earlier version of this paper.

References

- Agrawala M, Stolte C (2001) Rendering Effective Route Maps: Improving Usability Through Generalization. In: Proceedings of the SIGGRAPH 2001, ACM, Los Angeles, CA, pp 241–250
- Allen GL (1997) From Knowledge to Words to Wayfinding: Issues in the Production and Comprehension of Route Directions. In: Hirtle SC, Frank AU (eds) Spatial Information Theory: A Theoretical Basis for GIS, Lecture Notes in Computer Science, 1329, Springer, Berlin Heidelberg New York, pp 363–372
- Applegate DL, Bixby RE, Chvatal V, Cook WJ (2007) The Traveling Salesman Problem: A Computational Study, Princeton University Press, Princeton, NJ
- Brooks RA (1991) Intelligence Without Reason. In: Myopoulos J, Reiter R (eds) 12th International Joint Conference on Artificial Intelligence IJCAI-91, Morgan Kaufmann Publishers, San Mateo, CA, pp 569–595
- Cassell J, Kopp S, Tepper PA, Ferriman K, Striegnitz K (2007) Trading Spaces: How Humans and Humanoids Use Speech and Gesture to Give Directions. In: Nishida T (ed) Conversational Informatics. Wiley Series in Agent Technology, John Wiley & Sons Ltd., Chichester, UK, pp 133–160
- Crockett LJ (1994) The Turing Test and the Frame Problem: AI's Mistaken Understanding of Intelligence, Ablex Series in Artificial Intelligence, Ablex Publishing Corporation, Norwood, NJ

- Dale R, Geldof S, Prost J-P (2005) Using Natural Language Generation in Automatic Route Description. *Journal of Research and Practice in Information Technology* 37 (1): 89–105
- Denis M (1997) The Description of Routes: A Cognitive Approach to the Production of Spatial Discourse. *Current Psychology of Cognition* 16 (4): 409–458
- Denis M, Michon P-E, Tom A (2007) Assisting Pedestrian Wayfinding in Urban Settings: Why References to Landmarks are Crucial in Direction-Giving. In: Allen GL (ed) *Applied Spatial Cognition: From Research to Cognitive Technology*, Lawrence Erlbaum Associates, Mahwah, New Jersey, pp 25–51
- Dourish P (2001) *Where the Action Is: The Foundations of Embodied Interaction*, The MIT Press, Cambridge, Mass.
- Dry M, Lee MD, Vickers D, Hughes P (2006) Human Performance on Visually Presented Traveling Salesperson Problems with Varying Numbers of Nodes. *The Journal of Problem Solving* 2006(1): 20–32
- Egenhofer MJ (1997) Query Processing in Spatial-Query-by-Sketch. *Journal of Visual Languages and Computing* 8(4): 403–424
- Egenhofer MJ, Mark DM (1995) Naive Geography. In: Frank AU, Kuhn W (eds) *Spatial Information Theory, Lecture Notes in Computer Science*, 988, Springer, Berlin Heidelberg New York, pp 1–15
- Einert G (2006) EFA goes Down-Under, *mdv news*, 2006 (II), pp 13–15
- Frank AU (2003) Pragmatic Information Content - How to Measure the Information in a Route Description. In: Duckham M, Goodchild MF, Worboys M (eds) *Foundations in Geographic Information Science*, Taylor & Francis, London, pp 47–68
- French RM (2000) The Turing Test: The First Fifty Years. *Trends in Cognitive Sciences* 4(3): 115–121
- Gigerenzer G (2007) *Gut Feelings: The Intelligence of the Unconscious*, Viking Penguin, New York, NY
- Golledge RG (1999) Human Wayfinding and Cognitive Maps. In: Golledge RG (ed) *Wayfinding Behavior*, The Johns Hopkins University Press, Baltimore, MA, pp 5–45
- Grice P (1989) *Studies in the Way of Words*, Harvard University Press, Cambridge, Massachusetts
- Janelle DG (2004) Impact of Information Technologies. In: Hanson S, Giuliano G (eds) *The Geography of Urban Transportation*, Guilford Press, New York, pp 86–112
- Klein W (1979) Wegauskünfte. *Zeitschrift für Literaturwissenschaft und Linguistik* 33: 9–57
- Klein W (1982) Local Deixis in Route Directions. In: Jarvella RJ, Klein W (eds) *Speech, Place, and Action*, John Wiley & Sons, Chichester, pp 161–182
- Klippel A, Hansen S, Richter K-F, Winter S (2009) Urban Granularities - A Data Structure for Cognitively Ergonomic Route Directions. *GeoInformatica* 13(2): 223–247
- Kopp S, Wachsmuth I (2004) Synthesizing Multimodal Utterances for Conversational Agents. *Computer Animation and Virtual Worlds* 15(1): 39–52

- McCarthy J, Hayes PJ (1969) Some Philosophical Problems from the Standpoint of Artificial Intelligence. In: Melzer B, Michie D (eds) *Machine Intelligence*, Edinburgh University Press, Edinburgh, pp 463–502
- Pontikakis E (2006) *Wayfinding in GIS: Formalizing the Basic Needs of a Passenger When Using Public Transportation*. PhD thesis, Institute for Geoinformation and Cartography, Technical University Vienna, Austria
- Pylyshyn ZW (ed) (1987) *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*, Theoretical Issues in Cognitive Science, Ablex Publishing, Norwood, NJ
- Richter K-F (2008) *Context-Specific Route Directions*, Monograph Series of the Transregional Collaborative Research Center SFB/TR8, 3, Akademische Verlagsgesellschaft, Berlin
- Roth W-M (2007) From Action to Discourse: The Bridging Function of Gestures. *Cognitive Systems Research* 3(3): 535–554
- Searle JR (1980) Minds, Brains, and Programs. *The Behavioral and Brain Sciences* 3: 417–424
- Stevens A, Coupe P (1978) Distortions in Judged Spatial Relations. *Cognitive Psychology* 10(4): 422–437
- Suchman LA (1987) *Plans and Situated Actions: The Problem of Human Machine Communication*, Cambridge University Press, Cambridge, UK
- Tenbrink T, Winter S (2009) Granularity in Route Directions. *Spatial Cognition and Computation* 9(1): 64–93
- Timpf S (2002) Ontologies of Wayfinding: A Traveler's Perspective. *Networks and Spatial Economics* 2(1): 9–33
- Turing AM (1950) Computing Machinery and Intelligence. *Mind* 59(236): 433–460
- Winter S (2009) *Spatial Intelligence: Ready for a Challenge?* *Spatial Cognition and Computation* 9(2), accepted 12 March 2009
- Winter S, Wu Y (2008) Towards a Conceptual Model of Talking to a Route Planner. In: Bertolotto M, Ray C, Li X (eds), *W2GIS 2008, Lecture Notes in Computer Science*, 5373, Springer, Berlin Heidelberg New York, pp 107–123
- Wunderlich D, Reinelt R (1982) How to Get There From Here. In: Jarvella RJ, Klein W (eds), *Speech, Place, and Action*, John Wiley & Sons, Chichester, pp 183–201

Training Games and GIS

Marcelo G. Metello, Marco A. Casanova

Department of Informatics
Pontifical Catholic University of Rio de Janeiro

1 Introduction

By the end of the 20th century, the computer gaming industry already was one of the largest entertainment industries. Together with the movie industry, it attracted large investments in research in computer graphics, among other areas of computer science. As a parallel line of development, military and flight simulators also received huge investments. These simulators are akin to entertainment games in terms of their technical requirements.

More recently, the so-called *serious games* (Susi et al. 2007) also started to gain attention. The basic motivation lies in that the technology developed for games and simulators can also be applied to other areas, such as medicine, architecture, education, urban planning, and government (Smith 2007). The development of serious games poses specific challenges, however, since their requirements differ from those of entertainment games. In particular, serious games require simulation models that reproduce certain aspects of the real world. By contrast, entertainment games have much more freedom to simplify their real world model, which is convenient when developers face technical limitations.

We use the term *geospatial training games* to denote a subtype of serious games designed for computer-based training that require the use of spatio-temporal data representing geographic features. Games designed for emergency response training are a typical example of this class of games (Metello et al. 2008).

This paper is organized as follows. Section 2 discusses four requirements that heavily influence the design of geospatial training games. Section 3

proposes a general architecture for this class of games. Section 4 focuses on how to model processes in this context. Section 5 provides an example of a geospatial training game. Finally, Section 6 contains the conclusions.

2 Requirements for Geospatial Training Games

In this section, we discuss four requirements that heavily influence the design of geospatial training games: realistic user experience; interoperability with existing GIS; time flow control; player performance evaluation.

2.1 Realistic User Experience

Geospatial training games must offer players a realistic experience. This implies that the game must simulate real-world situations and must offer a multimedia user interface with a minimum degree of realism.

Indeed, although geospatial training games and entertainment games may be built over the same technologies, they are essentially different with respect to the situation they simulate. Geospatial training games must simulate real-world situations, while entertainment games do not.

As for the user interface of geospatial training games, the rendering of spatial data on the computer screen should have enough similarity with the geographical features depicted so as to help users reason spatially about the real world. Likewise, computer animation of spatio-temporal data should help users reason about dynamic geographic phenomena. Furthermore, animations may adopt different time scales without impairing how users assimilate the information. For example, if an animation shows in seconds the evolution in land use of a particular region during a period of some years, the animation must still give the user a sense of what happened with occupation of the region.

2.2 Interoperability with Existing GIS

Because of their need for realism, geospatial training games require the ability to use real-world data. This means accessing spatio-temporal data stored in geographic information systems (GIS). However, such systems have been traditionally more concerned with managing static spatial data than spatio-temporal data.

In particular, among other problems, the technology developed for computer games focuses on displaying data in real time. A continuous attention on performance is needed to keep an acceptable frame rate for a better user

experience. Furthermore, the game may potentially be multi-user, which makes the problem even more complicated. However, geographic information systems are not prepared to sustain the throughput geospatial training games require to refresh the data displayed. Therefore, specific indexing and caching mechanisms may be necessary to provide game-compatible frame rates (Metello et al. 2007).

2.3 Time Flow Control

Geospatial training games must offer control over the time flow. That is, they must offer the ability to stop the game, accelerate the pace of the game, and return the game to an earlier point in time.

Indeed, this requirement is important for the usability of training games. For example, even if an emergency situation may last for days, the simulation should obviously not take the same amount of time. Periods requiring no decision making should be fast-forwarded. Likewise, the ability to go back in time is also highly desirable to test alternative decisions.

To better support time flow control, the game may record simulation data. However, this requires investigating techniques to avoid two problems: an undue growth of the database; and slowing down the simulation.

Indeed, one should investigate alternative ways of storing spatio-temporal data other than storing a new version of a data item every time it is updated. As for the second problem, first note that storing simulation data may generate a massive number of database update requests. This poses an interesting problem on how much indexing should be used. Indexes can speed up queries, but they necessarily slow down update requests. In particular, spatial feature classification can play an important role with respect to optimizing the amount of indexing, if it can help differentiate features with high update frequency from nearly static ones.

2.4 Player Performance Evaluation

Geospatial training games must offer tools to evaluate the performance of the players.

Indeed, in the case of training games, player performance evaluation is often an important part of the learning cycle (Borodzicz and van Haperen 2002). A successful training system should improve the players' decision making abilities. Therefore, the system must allow players to trace the results of the game back to their decisions (Zagal et al. 2006).

If the evaluation process is totally manual, this requirement reduces to the ability of playing back a simulation, or at least the ability of providing

a timeline registering the main events that occurred, including the players' decisions.

More sophisticated training games are aware of the players' action model. For example, in organizations with predefined procedures, the game may match the sequence of players' actions to the predefined procedures to check whether the players acted as expected. Going one step further, the game may be used to evaluate the effectiveness of the predefined procedures, detect their flaws and help with their evolution (Smith 2004). In this case, learning is not limited to the individual level, since we may consider that the evaluation of the effectiveness of the predefined procedures represents some kind of collective or institutional learning.

3 An Architecture for Geospatial Training Games

In this section we introduce a high level, modularized architecture for geospatial training games (see Fig. 1). This architecture is general enough to represent the way most computer games are implemented. Throughout the rest of the paper, we will further specialize this abstract architecture into one that better matches the requirements of geospatial training games.

Indeed, serious games tend to rely on modular architectures (Westera et al. 2008). However, many entertainment games do not follow a modular architecture to achieve better performance. This is particularly true for action games, classified as *intensively performative* in Apperley (2006).

Referring to Fig. 1, the *world representation* component is basically a data structure, stored in main memory or in any persistence device, which models a state of the simulated situation. It is a passive component since it does not act on anything and remains unchanged while not receiving external influence. The *event manager* is the component responsible for all activity in the game simulation. As the game flows, the event manager continuously updates the world representation. Finally, the *renderer* component is responsible for materializing the player view of the world. Even though the rendering of dynamic geospatial data poses interesting challenges, this topic is out of the scope of this work due to its complexity. Instead, the focus here will be in modeling the dynamics of geospatial training games.

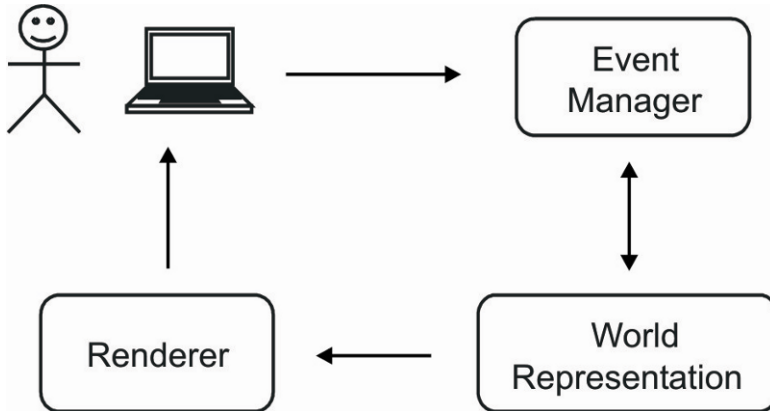


Fig. 1. The main components of an architecture for geospatial training games

All game activity is represented by *event streams*, as discussed in Section 4, in the sense that every change made to the world representation is carried out by some event. Both human player actions and simulation models generate event streams. The event manager is responsible for synchronizing all events and for leaving the world representation always in a consistent state.

Adding a human player as an interacting element in the simulation raises some interesting modeling challenges, as compared to fully automated methods traditionally used in GIS for simulations of dynamic phenomena. The human element is essentially different from all other simulated elements since his behavior is not deterministic. This fact makes it impossible to pre-compute the behavior of elements possibly affected by the actions of the player. For example, consider an emergency situation where oil leaked into the ocean. The dispersion of the oil on the water may involve complex calculations, which may be pre-computed. However, player actions, such as the placement of containment barriers, may affect oil dispersion, making the pre-computation useless in this case.

The non-deterministic behavior of human players requires more than simply modeling spatio-temporal data: it requires modeling processes that may run in parallel and that may interfere with each other. In the previous example, the parallel processes are the dispersion of the oil and the execution of some contingency plan by the human player.

Another difficulty is that human player processes and computer-simulated processes are inherently asynchronous, as illustrated in Fig. 2. Here two different processes generate event streams, one representing the actions of a human player and another representing a simulation made by some traditional dynamic modeling technique. This situation is similar to that of database management systems, where different asynchronous trans-

actions act on the same data items. In both cases, a synchronization mechanism is required to keep the data always in a consistent state. The event manager is responsible for implementing the synchronization mechanism in such a way as to facilitate time flow control, in special, rollback operations.

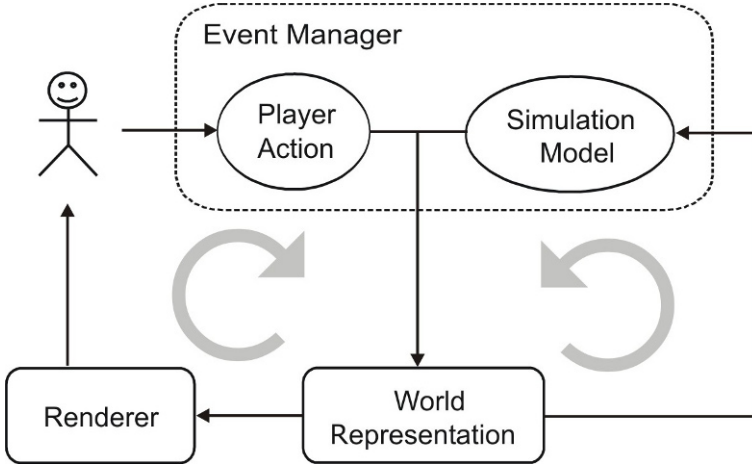


Fig. 2. Two inherently asynchronous processes: human player decision making (left) and automated simulation (right)

4 Process Modeling

We address in this section the central question of how to model processes. The main goal here is to model different kinds of processes in such a way that the event manager is able to properly synchronize their execution. As usual, we classify simulation models with respect to their time representation as *discrete* and *continuous*, and we model player processes as workflows.

4.1 Discrete Process Modeling

In GIS, some of the most popular dynamic modeling techniques, such as cellular automata (von Neumann 1966), are based on a discrete time scale. For each time t_i , the state S_i of the world representation is well defined. In this case, the perception of the process is represented as a sequence of states $S = (S_0, S_1, S_2, \dots, S_n)$.

Consider a continuous time interval $[a,b]$ such that $a \leq t_i \leq b$, for all $i \in [1,n]$. If we map the previously defined discrete time scale into this interval, the state of the simulation will be defined only for a few time values in the interval. Hence, this representation allows other processes to observe the state only at these few time values.

A first solution to the above problem is to force all processes that may possibly interfere to use compatible time scales. Therefore, every time one process needs to observe the state S_i of another process, S_i will be defined. By doing so, we are actually forcing the process models to be aware of the synchronization problem. The first drawback of this solution therefore is that it introduces dependencies between the process models, which reduce the reusability of the models. A second drawback is that the solution is not compatible with other types of process models, such as continuous models.

An alternative solution is to interpolate the state between any two consecutive time values, which allows the state to be observed at any point in time during the execution of the process. The simplest form of interpolation is to define a step function f , in the sense that the state S_i remains unchanged during the interval $[t_p, t_{i+1})$. Fig. 3 illustrates a single-variable step function whose domain is the time interval in question. The advantage of this simple interpolation method is that, for $t \in [t_p, t_{i+1})$, $f(t)$ does not depend on S_{i+1} , which is essential, since S_{i+1} cannot be computed before t_{i+1} because other processes may interfere with the process in question during the interval $[t_p, t_{i+1})$.

However, the use of a step function f may introduce errors because other processes may produce new data that f does not take into account. If this error is not acceptable, the discrete model should adopt a finer time granularity. The hypothesis here is that the error can be reduced to any extent by increasing the granularity of the discrete model.

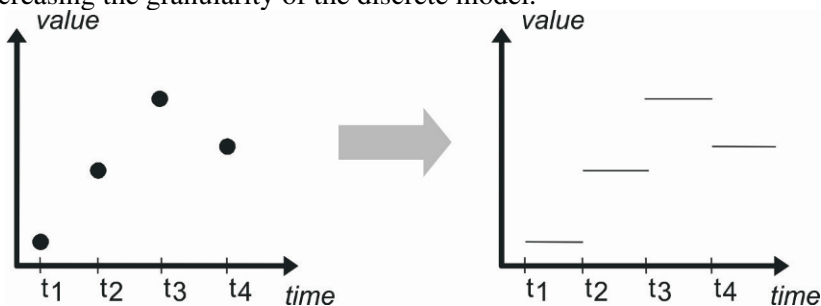


Fig. 3. Defining the state for all time values through a step function

In summary, the approximation by step functions may also bring to the modeling stage certain questions related to process interaction. However, the impact should be less than in the first solution. Step functions may re-

quire processes models to be more precise, but they are never restricted with respect to their time scales.

4.2 Continuous Process Modeling

The most widely used continuous dynamic models are based on systems of differential equations (Lee and Zheng 2005). Such models are able to define the state at any point in a continuous time interval. If the state needs to be observed at time t , the system of equations is solved for t . Such systems provide potentially unbounded precision with respect to time.

This approach poses no problems for other asynchronous processes to observe the state at any point in time. However, it is somewhat limited because it only works with numeric floating point values. It is not appropriate, for example, to deal with enumerable types where there is no notion of ordering of possible values. Besides, it is not clear how these models may be adapted to interact with other asynchronous processes.

In the scope of this paper, continuous process models will be restricted to simulated elements whose behavior are known in advance and which are not subject to change. In our oil leak example, the oil dispersion cannot be modeled as a continuous process simply because it is not known in advance where the containment barriers will be dropped.

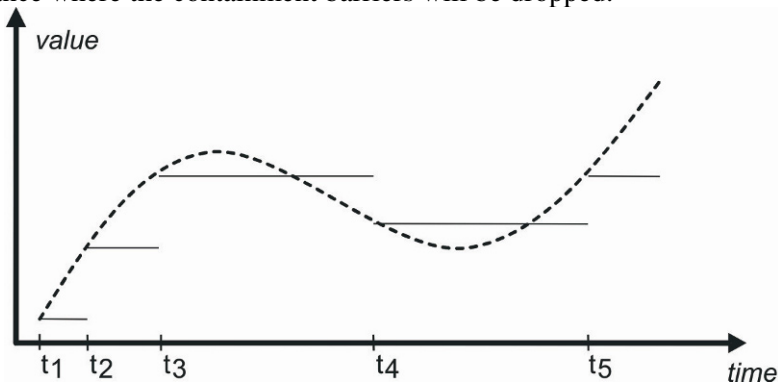


Fig. 4. A continuous function (dashed line) approximated by a step function (continuous lines)

It is also possible to map a continuous process model into a discrete one by using step functions to approximate the functions that the equations define, as illustrated in Fig. 4. Although making process interaction simpler, the infinite precision of continuous models is lost. Again, the granularity should be controlled to match the required precision, as discussed in Section 4.1.

4.3 Workflows

Workflows are widely used for business process modeling. For example, emergency plans may be modeled as workflows (Carvalho et al. 2001). Therefore, for player performance evaluation, it seems natural to model player activity as workflows with the purpose of comparing their actions with the predefined plans.

A workflow is essentially a set of actions and a control structure, which defines in what order actions should be executed in a given situation (van der Aalst 2003). Most workflow representations do not define the time at which the actions should be executed, and how long they will take to finish. In fact, many representations assume that actions are atomic.

Workflows do not necessarily define how actions affect the world. One approach is to define actions through their pre- and post-conditions, following the tradition of AI planning systems (Fikes and Nilsson 1971). But note that this representation still assumes actions to be atomic.

However, the assumption that actions are atomic may be too restrictive. For example, consider the action of walking. It may not be realistic to change the position of the character from the origin to its destination in one single instantaneous step. Instead, it is more realistic to simulate the trajectory of the character to the destination point through multiple state changes, so that his trajectory may be observed. If the world model requires that actions have duration and make changes to the world during their executions, we will have to model actions as processes, as described in Sections 4.1 and 4.2.

4.4 The Event Framework

In this section, we outline how to reduce discrete process models (or discretized continuous processes) and workflow models to a unifying *event framework*.

The situation simulated in a game is entirely captured by the world representation and an event stream. From another perspective, the world representation and the event stream represent a basic separation between the static and the dynamic parts of the reality, an approach used in (Sowa 2000) to model knowledge about processes.

Intuitively, an *event* represents an activity carried out in a game. An event has a *start time* and an *end time*, which define the time interval during which the activity takes place. An event that has the start time equal to the end time is called an *instantaneous event*. Only instantaneous events are allowed to change the world representation. This restriction is necessary to keep a well-defined and observable state at all times.

All events must use the same *global time scale*. If some process is modeled in a *local time scale*, there should be a way to translate between the local and the global time scales.

To introduce abstraction levels, an event E , called a *parent event*, may have *child events*. Child events are restricted with respect to time so that no child event starts before or finishes after its parent.

In this event framework, a discrete process (or a discretized continuous process) is represented as a stream of instantaneous events, whose start times are the points of discontinuity and which change the state of the process. A workflow is represented as a single parent event and a stream of child events corresponding to its actions.

A detailed discussion on the synchronization model is beyond the scope of this paper. Basically, synchronization is achieved by serializing the event streams from all processes, respecting their start times. If two instantaneous events start at the same time, a tiebreak rule is used.

5 An Example of a Geospatial Training Game

In order to illustrate the use of the proposed architecture, we implemented an emergency training game that simulates emergency situations in the context of oil terminals. The InfoPAE system (Carvalho et al. 2001) served as a basis for the implementation, since it includes a GIS module and is specifically designed to support emergency response.

5.1 Emergency Plans

To effectively handle emergency situations, teams must be trained to respond quickly and in a well organized manner. Practice has shown that this will happen only if the response strategies have been planned ahead of time, based on a risk analysis that considers the possible accidental scenarios. The strategies and procedures are usually documented as *emergency plans*, which comprise a set of instructions that outlines the steps that should be taken before, during, and after an emergency. A plan should also be backed up by a database of human and material resources, and a document repository.

Traditionally, emergency plans take a representational form. In more detail, to model knowledge about dynamic phenomena, Frasca (2003) discusses two different approaches: representation and simulation. According to the author, the main difference between both forms is that simulation attempts to model the behavior of the elements involved in the phenomenon, while representation is limited to retaining the perceptual characteristics of

it. To make it clear, the author gives the example of an aircraft landing procedure. A representation of a specific landing could take the form of a video where an observer would be incapable of interfering. By contrast, a flight simulator would allow the user to interfere with the phenomenon in a way that simulates the real aircraft. This flexibility is only possible due to the simulation characteristic of modeling the behavior of the elements, independently of any specific scenario.

In the context of the discussion in Section 4.3, the representational form of an emergency plan takes the form of a workflow. However, typically, the workflow does not model the dynamics of the actions in a realistic way, which would otherwise help emergency managers create more reliable plans. In other words, an emergency plan should take a hybrid form, where certain actions are coupled with simulations. For example, a plan may include an action that instructs the user to send two boats to place a barrier to contain an oil spill. If the actions were backed up by a simulation of the sailing conditions of the boats (and by a simulation of the oil dispersion), it would be easier to determine where the boats should drop the barrier.

After the plan is developed, it should be tested to verify its effectiveness and to train and evaluate the operational readiness of the emergency teams. Testing usually takes the form of field exercises or drills, which can be very time consuming and expensive. Another point to consider is the difficulty of representing detailed and realistic situations, as required to effectively test the emergency plan. In such cases, the use of emergency training games can be very helpful.

5.2 Example of Running the Emergency Training Game

To exemplify the use of the emergency training game developed, consider an accidental scenario where a considerable volume of oil spills into a bay. This scenario was chosen because: (1) it is typical of marine oil terminals; and (2) it must be accounted for by a combination of a simulation process (of the dispersion of the oil on the water) and player actions (of placing containment barriers) that interfere with each other.

In a typical oil spill scenario, the initial goals are to attempt to control the oil leak at the source of the spill and to limit the propagation of the floating oil as much as possible. These goals are typically achieved by using containment, recovery and clean-up strategies, which are implemented through specific operational procedures. These procedures depend on the accidental scenario, whose description includes: the oil type and its characteristics; the location of the source of the leak and its nearest landmarks; an

estimation of the amount of oil spilled; weather and sea conditions; and characteristics of the coastal area that might be affected.

We run a game that considered the oil spill scenario after the leak has stopped and that focused on oil containment. Clean-up operations for the oil that reached the coastal areas were not considered. The goal of the game was to test the emergency plans for leaked oil containment to uncover possible flaws, as well as to help planning equipment installation and location.

The initial conditions of the game were specified in a document, and included the location, amount and type of leaked oil, maps of the nearby coastal areas, the location of all available equipment, and weather conditions.

The dynamic elements of the game were modeled as follows. The dispersion of the oil on the water was modeled by a cellular automaton, which considered: the current location of the oil spot; environmental conditions, such as wind direction and speed, defined globally; obstacles, such as coastal lines and containment barriers; the type of the containment barriers, which differ on their absorption and containment capabilities; the type of the coastal areas, which also differ on how they absorb oil (for example, beaches absorb much more oil than rocky coasts).

The state of each cell describes: the amount of oil in the cell; if the cell intersects a barrier or the coastal line, in which case the cell is considered an obstacle cell; the type of the obstacle, if it is the case. Note that the state of a cell varies according to amount of oil in it, and if a barrier happens to be dropped in the area the cell represents (coastal lines are supposed to be fixed, but this assumption might be revised to take into account the tide).

Boats were also considered dynamic elements of the game. Boats are responsible for dropping containment barriers and for carrying oil recovery equipment, such as pumps and skimmers. The initial location of the boats was defined in the document that described the initial conditions of the game.

The movement of the boats was simulated by taking into account their speed and cargo capacities, as well as environmental conditions, such as wind, sea currents and tide. During the simulation, players guided the boats through way-points. They could place way-points wherever they want and send the boats to any of them. Of course, boats might also encounter obstacles such as islands. In this case, they just stopped and waited for further instructions.

The placement of a containment barrier is carried out by two boats, each one holding one end of the barrier. The command to place a barrier has parameters – the type of the barrier, the angle at which it should be put, the distance the boats should keep from each other, and the curvature that should be kept – and preconditions – the maximum distance between the

two boats, the length of the barrier, and favorable environmental conditions, among others.

5.3 Benefits of Using the Emergency Training Game

Some of the main benefits of using the emergency training game described in Section 5.2 include:

- to help finding flaws in the emergency plans
- to help testing whether available resources are sufficient to handle all accidental scenarios
- to help training emergency personnel and thereby improve their responsiveness
- to reduce costs, since simulation is obviously significantly cheaper than full scale exercises

6 Conclusions

We first listed four basic requirements that heavily influence the design of geospatial training games. Then, we introduced a very high-level architecture for geospatial training games that calls attention to the importance of simulating real-world phenomena and players' actions.

We moved on to discuss how to model processes, including players' actions. We introduced an event framework that integrates discrete modeling techniques used in GIS and workflows.

Finally, we described a prototype implementation of an emergency training game in the context of the InfoPAE emergency response system. Early results with the implementation indicated that the error introduced by step function approximations can be controlled by increasing the granularity of the event streams generated by the processes involved. The assumption proved valid for the case of oil dispersion and containment barrier placement. However, further work is required to define a generic process synchronization model that covers discrete processes and workflows.

References

Apperley T (2006) Genre and game studies: Toward a critical approach to video game genres. *Simulation & Gaming* 37(1): 6–23

- Borodzicz E, van Haperen K (2002) Individual and Group Learning in Crisis Simulations. *Journal of Contingencies and Crisis Management* 10(3): 139–147 doi:10.1111/1468-5973.00190
- Carvalho MT, Freire J, Casanova MA (2001) The Architecture of an Emergency Plan Deployment System. In: Proc. III Brazilian Symposium on Geoinformatics, Rio de Janeiro, Brazil
- Fikes RE, Nilsson NJ (1971) STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence* 2: 3–4
- Frasca G (2003) Simulation versus Narrative: Introduction to Ludology. In: Wolf MJP, Perron B (eds) *The Video Game Theory Reader*, Routledge
- Lee EA, Zheng H (2005) Operational Semantics of Hybrid Systems. Invited paper in: Proc. of Hybrid Systems: Computation and Control (HSCC), Zurich, Switzerland, Springer LNCS, 3414, pp 25–53
- Metello M, Vera M, Lemos M, Masiero L, Carvalho MTM (2007) Continuous Interaction with TDK Improving the User Experience in Terralib. In: Proc. IX Brazilian Symp. on GeoInformatics, Campos dos Jordão, Brazil
- Metello M, Casanova MA, Carvalho MTM (2008) Using Serious Game Techniques to Simulate Emergency Situations. In: Proc. X Brazilian Symposium on GeoInformatics, Rio de Janeiro, Brazil
- Smith D (2004) For Whom the Bell Tolls: Imagining Accidents and the Development of Crisis Simulation in Organizations. *Simulation & Gaming* 35(3): 347–362
- Smith R (2007) Game Impact Theory: Five Forces That Are Driving the Adoption of Game Technologies within Multiple Established Industries, *Games and Society Yearbook*
- Sowa J (2000) *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brook/Cole, a division of Thomsom Learning: Pacific Grove, CA
- Susi T, Johannesson M, Backlund P (2007) *Serious Games - An Overview*. Technical Report HS-IKI-TR-07-001, School of Humanities and Informatics, University of Skövde, Sweden
- van der Aalst WMP, der Hofstede AHM, Kiepuszewski B, Barros AP (2003) Workflow Patterns. *Distributed and Parallel Databases* 14(1): 5–51(47)
- von Neumann J (1966) *Theory of self-reproducing automata*, A.W. Burks, Illinois
- Westera W, Nadolski RJ, Hummel HGK, Wopereis IGJH (2008) Serious games for higher education: a framework for reducing design complexity. *Journal of Computer Assisted Learning* 24: 420–432
- Zagal JP, Rick J, Hsi I (2006) Collaborative games: Lessons learned from board games. *Simulation & Gaming* 37(1): 24–40

Cadastre and Economic Development

Erik Stubkjær

Department of Development and Planning, Aalborg University, Denmark

Abstract

Why are some countries economically much more successful than others? In search for empirically founded explanations, recent research investigates institutions in the sense of institutional economics, for example in terms of level of democracy, and of protection of property rights. The role of rights in land is so far only partly understood, due to the multifarious land tenure systems. The chapter provides a review and discussion of recent research on this issue.

Economic development has been supported by the World Bank Group and national aid agencies, among others. Interestingly, in recent years an alternative to this development incentive evolved. This is due to the fact that investors in emerging markets have looked for a market based alternative to government bonds, the security of which depends on the country's ability to collect taxes. An alternative is covered bonds, the security of which is backed by mortgages in title of real property. However, the success of this alternative depends on the availability of a well functioning land registration with cadastre, which motivates the title.

1 Introduction

1.1 A Research Brief on Cadastre and Economic Development

Across cultures, land and natural resources are managed in a variety of ways. The economic success of Western countries motivated the design of

land administration projects, aiming at extending individual property rights in developing countries. However, the limited success of these development projects during the last decades calls for more basic reflections on the causes of economic growth. This section outlines the political and theoretical context of the development projects, as well as selected research, as a backdrop for more comprehensive reporting on research findings in the following sections.

Cadastré is the term applied for 'land record[s] for tax purposes', according to S. Rowton Simpson (1976), who refers to French usage. In the English speaking world, the preferred term is Cadastral Surveys (Dale 1976; Simpson 1976; Bruce 1998). The term apparently reflects ongoing efforts to map and record natural resources. Simpson distinguishes two functions of land registration: In Continental Europe, the state wanted to make 'an inventory of the national land resources for fiscal purposes or in order to ensure proper development' (p 3). In England and in her former colonies, land registration refers to a means 'solely to simplify conveyancing (as the business of creating and transferring rights in land is called)' (p 3). The different focus on state and individual interests, respectively, to be illustrated below in section 3, without doubt contributes to the wide variety of terms for the domain concerned, including land registration, cadastral systems, land information management, land administration, and recently (part of) spatial data infrastructures.

Research within the domain intensified in the 1990s, as '[s]everal studies show a linkage between economic development and cadastral systems' (Frank 1996). Economic development and its driving factors may be conceived in a number of ways. However, the collapse of Soviet Union in 1989 gave wide political legitimacy to the neo-liberal paradigm, which was introduced by President Reagan in the United States and Margaret Thatcher in the United Kingdom. As recorded by FAO staff: 'Almost all the major donors are now supporting land reform programmes in conceptual terms that are compatible with the Washington Consensus on the role of the market, property rights and institutional reform.' (FAO 1998)

The theoretical foundation of much research and development regarding property rights and land registration was New Institutional Economics, which may be applied both to mundane business transactions as well as to centennial change. For example, the concept of 'transaction costs' was applied to the domain of real estate in a European joint research action 'Modeling Real Property Transactions' (Zevenbergen et al. 2007). This effort aimed at modelling and comparing processes regarding real estate, e.g. conveyancing and subdivision (Navratil and Frank 2004, 2007).

Nobel laureate Douglass C North contributed to New Institutional Economics among others through analysis of US land law through the 19th century in terms of institutional change and path dependency (North 1990).

Linking similar basic issues with business transactions, Hernando de Soto in a number of developing countries charted the processes needed to buy a piece of land. For example it would take 111 steps and 4112 days in Haiti to obtain the needed sales contract (De Soto 2000). His question, ‘Why Capitalism triumphs in the West and fails everywhere else’, very much in line with the Washing Consensus mentioned above, was addressed among others with reference to the fact that ‘[t]he European and U.S. tradition points to self-help groups that initially organized mutual credit in Savings and Loan Associations, where reliance was not exclusively on the legal rules, but on simple rules laid down by the associations’ (Frank 2008, p 275).

Despite substantial efforts in terms of land administration development projects during the last decades, the Washington Consensus focus on individual property rights did not pay of in terms of streamlining processes regarding real estate, and especially raising capital through mortgaging of secure title (Mitchell et al. 2008). Reflecting the past experiences, a World Bank Policy Research Report among others admitted more concern to land rentals (Deininger 2003).

This brief survey may suffice to give way to the following introduction of the remainder of the chapter.

1.2 “Why Isn’t the Whole World Developed?”

The quotation is taken from a frequently cited Presidential address to the Economic History Association in 1980. Richard A. Easterlin (1981) sets out from the diffusion of new production techniques, but rather than focusing on patterns in trade, he investigates in the spread of knowledge. He contends that the diffusion of knowledge concerning production techniques by different countries has been governed largely by whether their populations have had formal schooling. His approach gave impetus to a number of statistical analyses, seeking to explain the causes of economic growth. One of these analyses regards the different effect of British and French colonial policies. It is reported in some detail, because it reveals that the creditable objective of including all citizens of the realm into the mainland culture is not easily implemented, as we shall see developed in section 2.

So far, the statistical analyses have not conclusively answered the headline question. There is a general agreement that institutions (in the sense of institutional economics) matter as a factor of growth. Therefore, also the institution of property rights is related to growth, although many opinions wrestle on the identification of adequate institutions and the level of quality of these institutions, cf. section 3.1. One of the well known positions is

stated by Hernando de Soto, who as indicated above rephrased Easterlin's question on economic growth and pondered on 'The Mystery of Capital' (2000). In addition to the mentioned careful recording of troublesome processes, he invites to research in the history of evolving institutions, in order to clear up the mystery of diverse development. His research thus points to the fact that economic growth presupposes a recurrent moulding of institutions in order to catch new options and respond to varying risks. This line of investigation has been pursued since the 1990s by a group of researchers, who have studied the origins of development economics in the early modern states of Continental Europe (Jomo and Reinert 2005). Section 3.2 provides an overview of these dynamic, institutional issues.

The development projects of aid agencies and the World Bank Group, mentioned in the introductory section, generally include an effort to strengthen the organisations which identify units of real estate and record rights in land and other attributes. Recently, an impetus to support such development has come from quite another kind of actors, namely the investors who look for the potential of emerging markets. Investors look for an asset portfolio which maximize yield on condition of investment period, risk tolerance, and similar profiles. For investors with long horizons and low risk tolerance, for example pension funds, government bonds are a likely option. However, taking a global perspective, the government bonds of a country with weak public administration and frailty in collection of taxes demands attentive risk assessment.

In recent years, alternative type of bonds have been investigated and recommended, including mortgage or covered bonds. Mortgage bonds are issued by an agency which may be related to government only through statutory law. They are backed by a mortgage deed in a unit of real estate. 'Covered bonds' is the term mostly used in Anglo-American language on European varieties of mortgage bonds, especially the German Pfandbriefe and the Danish Realkredit-obligationer. Section 4 characterizes this variety, quoting recent comparisons. In the above context of investigations into economic growth, the question arises why Continental European institutional variants did not spread to the otherwise leading centres of growth. The question is related to different views on the state as an agent of growth. The review of causes of growth is summarized in the conclusion.

2 Statistical Analyses of the Causes of Economic Growth

Economic performance has varied substantially over time and space, as evidenced by the estimates given in Table 1 (Berger 2007, quoting Maddison 2001). As of the year 1000, the inequalities amounted to about 10 per cent. Around 1910, the per capita gross domestic product of the 'Non-

European West' was about ten times that of Africa. Around 2000, 'the West' and Japan scored three times the value of South America, and 4–5 times that of Eastern Europe and former USSR.

Table 1. Per capita gross domestic product and its interregional spread 1000–1998 (Berger 2007, p 10, source Maddison 2001, p 126)

	1000	1500	1820	1870	1913	1950	1973	1998
Westeuropa	400	774	1232	1974	3473	4594	11534	17921
Westliche Ableger	400	400	1201	2431	5257	9288	16172	26146
Japan	425	500	669	737	1387	1926	11439	20413
Asien (ohne Japan)	450	527	575	543	640	635	1231	2936
Lateinamerika	400	416	665	698	1511	2554	4531	5795
Osteuropa und frühere UDSSR	400	483	667	917	1501	2601	5729	4354
Afrik	416	400	418	444	585	852	1365	1368
Die gesamte Welt	435	565	667	867	1510	2114	4101	5709
Interregionale Spreizung	1.1:1	2:1	3:1	5:1	9:1	15:1	13:1	19:1

2.1 Technology and Belief

It seems generally accepted that the prime cause of economic growth has been the sharp acceleration in the rate of technological change in a relatively small number of nations. Moreover, several studies support the view that a common technology diffused from one country to the next. It is evidenced as well by the striking likeness of modern industrial technology among the various high productivity nations themselves (Easterlin 1981, p 3). Easterlin notes that technology diffused together with a belief system, a conception of natural reality as law-like, causally ordered, and manipulable (1981, p 16) and Berger agrees to this, pointing to the impact of an intelligible world view on technological development (2007, pp 17, 21).

2.2 Formal Education

The transfer of technology depends on a personal element, namely the comprehension of and training in actual operations. This leads Easterlin to an investigation into the formal schooling of a population, because 'the more schooling of appropriate content that a nation's population had, the easier it was to master the new technological knowledge becoming available.' Also, 'substantial increases in formal schooling tend to be accompanied by significant improvement in the incentive structure.' (Easterlin

1981, p 6). Data for twenty-five of the largest countries of the world during 1830–1975 are analyzed, taking primary school enrolment rate as an indicator of educational development. Within Europe the most advanced nations educationally, those in Northern and Western Europe, were the ones that developed first. Not until the end of the nineteenth century did most of Southern and Eastern Europe start to approach educational levels comparable to the initial levels in the north and west, and it was around this time that these nations began to develop.

With regard to the overseas descendants of Europe the picture is the same: The leader in schooling is the leader in development, the United States. In Asia, Japan's nineteenth-century educational attainment is clearly distinctive, even before the Meiji Restoration from 1870s onwards. Contrary, Turkey, a nation subject in many ways to external economic and political pressures similar to those experienced by Japan, had still by 1940s low educational levels and also failed to show substantial technological modernization (Easterlin 1981, p 7).

The role of education is illustrated also through an investigation into the effect of the identity of the colonial power for the country's subsequent growth and development (Grier 1999). Grier pooled data from 63 ex-colonial states over the period 1961–1990 and found that former British colonies performed significantly better on average than their French and Spanish counterparts. Moreover, the longer the country was held by the mother country, the better it did economically in the postcolonial era. Based on a reduced sample of 24 countries in Africa, the length of colonization is still positively and significantly related to economic growth and former British colonies still outperform their French counterparts. By adding two variables to the regression, representing human and physical capital at the time of independence, Grier found that the newly independent British colonies were significantly more educated than the French ones and that the inclusion of education at independence can explain the development gap between the former British and French colonies and the positive relationship between length of colonization and growth (Grier 1999, p 318).

The specific character of the colonial powers suggests causes explaining these findings. The republican principles of France implied that the republic was one and indivisible: colonies were an intrinsic part of it, and full cultural assimilation should be aimed at. As a consequence, students were required to speak French, and all vernacular languages were forbidden. This resulted in large numbers of the population failing to achieve any kind of literacy. In fact, in the late 1960s, up to 95% of the population in France's former Black African territories were illiterate. Contrary to that, the British were more decentralized in their colonial approach. The British government kept control over constitution and foreign relations, while do-

mestic policies and budgetary matters, were resolved by the colonial legislatures. Local governments were allowed to keep all revenue surpluses. Moreover, British colonial education policies made a conscious effort to avoid alienating the native culture, by teaching in the vernacular languages and training teachers from the indigenous tribes (Grier 1999, pp 319–320).

While the British trade policy is also mentioned to explain why her former colonies performed significantly better in the post colonial period, the main message of Grier's investigation is the impact of general education on economic development: the British left a more educated populace than the French in Africa and there is empirical support for the claim that the amount of human capital has significant and permanent effect on subsequent growth and development.

2.3 The Spread of News

Trade routes constituted channels for spread of news before the age of the telegraph, radio transmissions, and the Internet. Grier describes how the Seville merchants, as the most powerful guild in Spain, controlled all incoming and outgoing trade with the Indies. Similar, in Mexico, Veracruz was the only legal port for international trade. Moreover, Spain secured its trade monopoly with the colonies through a monopoly convoy system where fleets were sent periodically throughout the year from Seville or Cadiz to established colonial ports (1999, p 320). Although the convoy system was the easiest way for Spain to protect the trading ships from possible seizure by the English or Dutch, the monopoly system restricted the spread of news. The British trade with her colonies was restricted, too, at the outset, but in 1846, the colonies were no longer forced to give British goods preferential treatment.

The importance of access to news appears from an investigation into measures of the 'social capability' of a country. Data on 41 social, political, and economic indicators were collected for 74 developing countries, generally for the period 1957–1962. The study established a strong correlation between growth and the extent of mass communications and newspaper circulation in 1960 (Temple and Johnson 1998). The effect of mass communications is particularly strong. Combined with initial income, these two variables explain about 30 percent of the variation in growth rates. If three influential outliers are omitted (Ghana, Japan, and Syria) then this figure rises to 40 percent (Temple and Johnson 1998, p 979).

The above-mentioned index of socioeconomic development was in fact constructed in the early 1960s by the development economists Irma Adelman and Cynthia Taft Morris and data were collected and published in 1967. Temple and Johnson state that if observers in the early 1960s had

given more emphasis to these indexes of social capability, they might have been rather more successful in predicting the fast growth of East Asia, and the underperformance of sub-Saharan Africa (1998, p 987).

2.4 Limits to Statistical Analyses of Causes of Economic Growth

Statistical analyses of the causes of economic growth include also the impact of the protection of personal and property rights. The review of these analyses is deferred to next section for two reasons. First, the above mentioned analysis combined initial income with mass communication to explain the growth rate. This raises concern, whether growth depends on independent factors or more or less depends on (level of) growth (rates) also.

Paldam and Gundelach contrast two alternatives: one assuming that institutions are chosen, while the growth rate is an outcome. According to this view, dubbed 'the Primacy of Institutions' view, cross-country data will not necessarily reveal a systematic pattern of development since the selection of certain institutions will not be ubiquitous and will not follow a deterministic ordering. The alternative view, 'Grand Transition', assumes a lucky spark to set development into motion, but then things gradually change in much the same way and a steady economic growth causes transitions of all institutions. This view would interpret cross-country data as representing an underlying systematic pattern overlaid with country heterogeneity and noise. Investigating degree of democracy and degree of corruption for typically 100 to 170 countries, their conclusion is that 'that both views tell some important part of the complex story of the great process of development'. Through arguments and analyses, too complex to address here, they warn that 'the dynamics of research may have taken prevailing opinions a bit too far toward the 'Primacy of Institutions' view' (2008).

The other reason for not trying to further identify causes of economic growth is the present ambiguity of the notion of 'quality of institutions'. As we shall see, the understanding of growth-related institutions varies among investigations and the effect of various configurations of institutions and their functional components are not spelled out. When descriptive measures such as absence or presence of a particular type or component of an institution are not stated, the quality or performance of that institution hardly can be assessed in a reliable way (cf. Sengupta 2003). The development process needs a dynamic and complex societal response to the changing options and threats of a country. We shall explore this aspect in section 3.2.

3 The Quality of Institutions and the Role of the State

The establishment of secure and stable property rights has been a key element in the rise of the West and the onset of modern economic growth, as argued by North and Thomas (1973), North and Weingast (1989), and North (1990). This has led to a number of investigations which explore the impact of property rights and other institutions on economic growth, including Knack and Keefer (1995, 1997), Classens and Laeven (2003), and the International Monetary Fund (IMF 2003). The review below provides an initial survey over available indices with bearing on land tenure, especially measures of the perceived strength of property rights.

3.1 Available Indices of the Protection of Property Rights

Knack and Keefer have constructed an index using data from the International Country Risk Guide (ICRG). This index measures a country's property rights in a broad sense and includes five measures: quality of the bureaucracy, corruption in government, rule of law, expropriation risk, and repudiation of contracts by the government (on a scale from 0 to 10). The index equals the average rating between 1982 and 1995 (1995, as reported by Classens and Laeven 2003). The ICRG is updated by The PRS Group, Inc., East Syracuse, NY, and available on a commercial basis.

Classens and Laeven empirically explore the role of property rights in influencing the allocation of investable resources. 'At the firm level, our idea of property rights is the degree of protection of the return on assets against powerful competitors. This notion of property rights is different from what is common in the literature where it is typically regarded as the protection of assets against actions by government. By focusing on the asset side of a firm's balance sheet, we instead use the term property rights as referring to the protection of entrepreneurial and other investment in firm assets against actions of other firms. We argue that a firm operating in a market with weaker property rights may be led to invest more in fixed assets relative to intangible assets because it finds it relatively more difficult to secure returns from intangible assets than from fixed assets.' (Classens and Laeven 2003, p 2403)

Moreover, the authors argue that property rights matter more for securing returns from intangible assets than from tangible assets and then relate strength of property protection to the firm's investment allocation and hence to firm growth. They use two property indexes in addition to the one by Knack and Keefer, namely the rating of protection of property rights from the Index of Economic Freedom constructed by the Heritage Foundation. Property rights are rated in each country on a scale from 1 to 5. The

score is based, broadly, on the degree of legal protection of private property, the probability that the government will expropriate private property, and the country's legal protection of private property. The index equals the median rating for the period 1995 to 1999. The third index of property rights is taken from the Global Competitiveness Report of the World Economic Forum. It provides an index of property rights on a scale from 1 to 7 in 2001. These indices are rendered in an Appendix for each of about 35 countries. The conclusion includes the finding that the effect of insecure property rights on the asset mix of firms is economically as important as the lack of access to external financing, because it impedes the growth of firms to the same quantitative magnitude.

Also the International Monetary Fund, IMF, offers an empirical analysis of the relation between institutional quality and development. The analysis uses three measures of institutions. These indicate, first, the quality of governance, including the degree of corruption, political rights, public sector efficiency, and regulatory burdens; second, the extent of legal protection of private property and how well such laws are enforced; and third, the level of institutional and other limits placed on political leaders (2003). The source of quality of governance is (Kaufmann et al. 1999, cf. 2002). As for the protection of property, the source is the Heritage Foundation mentioned above.

As mentioned at the outset, the above listing of available data may provide for an initial ranking of countries according to protection of property rights, but due to the present incomplete understanding of institutions, such ranking is deferred. However, the conception of protection of property rights by Classens and Laeven triggers a note. They state that 'property rights is the degree of protection of the return on assets against powerful competitors' (2003, p 2403). This conception calls to mind an intervention by Danish governmental officials during the 1780s, in order to enforce the economic rights of tenants relative to their landlords (Stubkjær 2008, p 247). Most of the investigations mentioned above seem to adopt the point of view of foreign direct investors and the likely agent of abuse of power is the target country government, rather than peers. Broadening the scope of agents of abuse, a measure of the quality of economic institutions could be the availability of an impartial and skilful third party, who would monitor asset transactions in a cost-effective way and thereby reduce abuse of power. Moreover, a measure of the quality of political institutions could be the concern for including economic actors at the edge of the present market (e.g. the Danish tenants) into the market.

3.2 The Management of Institutional Change for Growth

To stay in the market, an economic agent needs to react skilfully to options and threats, and the same apply when the economic agent is a country. The present inequality in economic development is partly the result of conscious action, as evidenced by economic history. For example, the English Tudor plan succeeded during 1485–1603 in building a wool manufacturing sector through enlightened policy, as tariff protection was combined with the intention and the factual establishment of new and competitive industries (Jomo and Reinert 2005, p 16). England gradually established itself as the workshop of the world. By importing raw materials and exporting manufactured goods, it experienced the most dramatic increase in wealth the world had yet seen. When the colonies wanted to establish their own manufactures, most were prevented from implementing this, despite the acclaimed benefits of the free market (Jomo and Reinert 2005, pp 11, 102, with reference also to Chang 2002). Only the United States were strong enough to set up, when they got independence from Britain, a protection scheme of tariffs and bans on imports to allow ‘infant industries’ to become internationally competitive. In the 19th century American tariffs were among the highest in the world. It is remarkable that it was only after World War II that the US significantly liberalized its trade and took up the cause of free trade. (Jomo and Reinert 2005, p 102)

The Continental European countries, who could not benefit from gold mines like Spain, initiated during 18th century surveys of their assets in the context of bringing about reforms to promote the common weal of the country: development of agriculture and manufactures, increase of population and development of their skills in terms of professions and inventions, regulation of foreign trade to obtain a balance of payment, and in addition provided institutional inventions like fire insurance and later a mortgage scheme. Such ‘Cameralist’ measures were widely discussed and lectured at several, mostly German universities (Jomo and Reinert 2005, pp 53–59). The national ambition of ‘keeping up with the neighbours’ provides motivation, both for the early general education, reported on in section 2, but also for the establishment of cadastres as a base for quasi-rational development, besides the taxing purpose.

Above, the well educated populace in former British colonies in Africa was related to the fact that restrictions in the British trade with her colonies were lifted in 1846, providing the colonies freedom in trade. However, freedom in trade may well co-exist with the restrictions in production mentioned by Reinert. Overall, this brief and perhaps incomplete exercise in economic history may give an impression of the complex societal interactions needed to climb the ladder of economic growth. The configuration of political institutions needed to mould a country’s economic institutions to

achieve growth seems only partly identified and the configuration is likely to vary across cultures.

4 Creating Capital with Covered Bonds

In recent years, investors in emerging markets have looked for alternatives to government bonds, creating an interest in covered bonds, or financial products comparable to the US mortgage-backed securities (Chiquier 2004). ‘Covered bonds’ is the term mostly used in Anglo-American language on European varieties of mortgage bonds, especially the German Pfandbriefe and the Danish Realkreditobligationer. To illustrate the amount of capital created, in 2003 the total volume of outstanding mortgage loans amounted to 101% of GDP in Denmark and 81% in the USA (Frankel 2004, p 97). German, Danish, and US mortgage systems were compared in a study, issued by the UN Economic Commission for Europe. Among others, differences in gross borrowing costs were estimated and the European versions found notable cheaper (UNECE 2005, p 46).

The Danish mortgage bonds and the related practise of securing the bonds through a mortgage deed, recorded in the land registry, developed from the Cameralist tradition of 18th century and were implemented through an act on mortgage credit associations from 1850. ‘The Danish market has a long and very well-documented history and no bankruptcy or liquidation [of mortgage credit institutions]... has ever taken place.’ (Moody 2002, p 7). Thus, this lending practice had and still has a number of benefits (cf. Nyboe Andersen 2003):

- It introduces the mortgage associations, now mortgage-credit institutes, as a third party between the borrower and the investor, thereby eliminating undue influence. Moreover,
- the mortgage-credit product gives even small borrowers the opportunity to finance homes and businesses on financial-market terms.
- The so-called ‘balance principle’, instituted by the Mortgage Credit Act, requires a high degree of accordance between the terms of lending and the issued bonds and thereby practically eliminates bankruptcy. The balance principle is fundamentally a very simple way of managing risk.

Likely as a consequence, Danish mortgage practises have attracted international interest and presently a Danish mortgage system is being introduced in Mexico (VP Securities 2005; Lozano 2008).

The above mentioned reports on European and US practices reveal a remarkable difference: The US / Washington Consensus point of view is the investor and the context is the market. The point of view in Danish tra-

dition is the tenant, the owner, or the borrower, and the context is the state. This is not to say that markets in Denmark are not operating or are ignored, rather: they are regulated. The lucky invention was the introduction of the not-for-profit mortgage credit institutions as a skilful, impartial mediator between borrower and investor (Stubkjær 2008, p 256). However, before mortgage institutions can be transplanted to other societies, research is needed to identify quality measures for the organisations which 1) identify units of real estate, 2) record rights in land, and 3) perform the foreclosure process in case of default, because the mortgage system depend on these.

5 Conclusion

The essay set out to investigate the causes of economic growth, considering the wide variation across countries. The use and development of technology to increase productivity is generally considered the basic cause of growth, together with the extent of a formal schooling, which includes technology and reason. Moreover, two mechanisms emerged which contribute to explaining the spatial variation: The channels for dissemination of news and the quality of institutions, respectively. The channels materialize in terms of international trade routes and the volume of news media. The news spread more freely through the English speaking channels of commerce. The amount of news media and the level of education indicate the capacity for absorbing news and to rephrase the news to fit the local context.

The quality of institutions seems best described in the terms of institutional economics, but the relevant institutional variables have not yet been empirically sorted out. Moreover, economic growth presupposes a recurrent moulding of institutions in order to catch new options and respond to varying risks. The scattered evidence provided spans the dominating position of neo-liberalism during the last decades and an alternative, development economics position, which allows for a more active and appreciated role of the state. This may revive interest in the 18th century ‘Cameralist’ institutions, including cadastre and mortgaging schemes.

As an alternative to the Washington Consensus focus on individual ownership and the market as instruments of growth, the presented evidence points to the beneficial effect of regulating the market through skilful, impartial mediators. This evidence is not only of Danish origin, as the contribution by Claessens and Laeven (2003) remind us of ‘powerful competitors’.

Acknowledgements

Dr. Johannes Luef, president and CEO of VP Securities A/S, Copenhagen, kindly supported a proposal to investigate relations between cadastre, mortgage, and poverty alleviation.

References

- Berger J (2007) Warum sind einige Länder so viel reicher als andere? Zur institutionellen Erklärung von Entwicklungs-unterschieden. *Zeitschrift für Soziologie* 36(1): 5–24
- Bruce JW (1998) Review of tenure terminology, Tenure Brief, Nr. 1, July 1998. Land Tenure Center, University of Wisconsin, Madison. <http://digital.library.wisc.edu/1793/22013>
- Chang H-J (2002) *Kicking Away the Ladder: Development Strategy in Historical Perspective*, Anthem Press, London
- Chiquier L, Hassler O, Lea M (2004) Mortgage Securities in Emerging Markets. World Bank Policy Research, Working Paper No. 3370
- Claessens S, Laeven L (2003) Financial Development, Property Rights, and Growth. *The Journal of Finance* LVIII(6): 2401–2436
- Dale PF (1976) *Cadastral Surreys within the Commonwealth*, H.M.S.O. London
- Deininger K (2003) *Land Policies for Growth and Poverty Reduction*, World Bank Policy Research Report, Oxford University Press, New York cf. <http://go.worldbank.org/A9HSAOTB60>
- De Soto H (2000) *The Mystery of Capital: Why Capitalism Triumphs in the West and Fails Everywhere Else*, Basic Books
- Easterlin RA (1981) Why Isn't the Whole World Developed? *The Journal of Economic History* 41(1 (Special Issue: The Tasks of Economic History)): 1–19
- FAO (1998) *Contemporary thinking on Land Reforms*, A paper prepared by the staff of the Land Tenure Service in Rural Development Division, Rome, 1998. <http://www.landcoalition.org/docs/odfaomon1.htm>
- Frank AU (1996) An object-oriented, formal approach to the design of cadastral systems. In: Kraak MJ, Molenaar M (eds) *Proceedings of 7th International Symposium on Spatial Data Handling, SDH '96, Advances in GIS Research II*, Delft, The Netherlands (August 12–16, 1996), IGU, Vol. 1, pp 5A.19–5A.35
- Frank AU (2008) A Case for Simple Laws. In: Smith B, Mark D, Ehrlich I (eds) *The Mystery of Capital and the Construction of Social Reality*, Open Court, ISBN: 978-0812696158
- Frankel A, Gyntelberg J, Kjeldsen K, Persson M (2004) The Danish mortgage market - As housing finance evolves, are there reasons to follow the Danish model? *BIS Quarterly Review*, March 2004, pp 95–109
- Grier RM (1999) Colonial legacies and economic growth. *Public Choice* 98: 317–335

- IMF (2003) World Economic Outlook: Growth and Institutions, A Survey by the Staff of the International Monetary Fund, April 2003, <http://www.imf.org/External/Pubs/FT/weo/2003/01/>
- Jomo KS, Reinert ES (eds) (2005) *The Origins of Development Economics: How Schools of Economic Thought Have Addressed Development*, Zed Books, London New York
- Kaufmann D, Kraay A, Zoido-Lobato P (1999) Aggregating Governance Indicators, World Bank Policy Research Working Paper No. 2195 World Bank, Washington, see WPs No. 2196 and 2772, the latter with an update for 2000–2001
- Knack S, Keefer P (1995) Institutions and economic performance: cross-country tests using alternative institutional measures. *Economics and Politics* 7(3): 207–227
- Knack S, Keefer P (1997) Does social capital have an economic payoff? A cross-country investigation. *Quarterly Journal of Economics* 112(4): 1251–1288
- Lozano L (2008) Mexican structured finance rides out the global credit storm, *Euro money*, March, 2008 (Latin America)
- Maddison A (2001) *The World Economy: A Millennial Perspective*. Development Centre Studies, OECD, Paris
- Mitchell D, Clarke M, Baxter J (2008) Evaluating land administration projects in developing countries. *Land Use Policy* 25(4): 464–473
- Moody (2002) Danish mortgage bonds: Highly secure financial instruments, Moody's Investor Service, Special Comment, May 2002
- Navratil G, Frank AU (2004) Processes in a Cadastre. *International Journal on Computers, Environment and Urban Systems* 28(5): 471–486
- Navratil G, Frank AU (2007) Hierarchies in Subdivision Processes. In: Zevenbergen J, Frank AU, Stubkjær E (eds) *Real Property Transactions*, IOS Press, Amsterdam, pp 221–240
- North DC (1990) *Institutions, Institutional Change and Economic Performance*, Cambridge University Press, Cambridge, UK
- Nyboe Andersen B (2003) Speech at the Annual Meeting of the Association of Danish Mortgage Banks, 30. April 2003
- Paldam M, Gundlach E (2008) Two Views on Institutions and Development: The Grand Transition vs the Primacy of Institutions. *Kyklos* 61(1): 65–100
- Sengupta, R (2003) *The World Economic Outlook, April 2003: A Review of the IMF's Assessment on Relation between Economic Growth and Institutions* http://www.ideaswebsite.org/themes/world/may2003/print/prnt090503_World_Economic_Outlook.htm
- Simpson SR (1976) *Land Law and Registration*. Cambridge University Press, London
- Stubkjær E (2008) The Institutionalization of Real Property Rights: The Case of Denmark. In: Smith B, Mark D, Ehrlich I (eds) *The Mystery of Capital and the Construction of Social Reality*, Chicago, USA, Open Court Publishing Company, pp 229–259
- Temple J, Johnson PA (1998) Social Capability and Economic Growth. *Quarterly Journal of Economics* 113(3): 965–990

- UNECE (2005) Housing Finance Systems for Countries in Transition - Principles and Examples, United Nations, Economic Commission for Europe, Geneva.
http://www.unece.org/hlm/prgm/hmm/hsg_finance/housingfinance.pdf
- VP Securities (2005) VP Securities Services to set up mortgage credit system in Mexico
<http://www.uk.vp.dk/C1256CF300390A8E/0/75FAEF723F8917C6C1257075002A6A03>
- Zevenbergen J, Frank AU, Stubkjær E (eds) (2007) Real Property Transactions - Procedures, Transaction Costs and Models, IOS Press, Amsterdam, Netherlands, ISBN: 978-1-58603-581-5

Index

A

Accuracy, 102
Attribute accuracy, 102, 104

B

Balance principle, 276
Basic set of movements, 159
Batch process, 85

C

Cadastral, 265–277
Cardinal directions, 172, 210
Cartography, 222
Cellular automata, 256
Change, 207
Chorem discovery, 225
Chorem layout, 225
Chorems, 224
ChorML, 229
Classification, 10, 48
Classification of paths, 155
Code of ethics, 28
Cognition, 209
Cognitive spatial representations,
239
Communication, 107, 240, 245
Completeness, 102
Composition, 171
Computational geometry, 90
Computing with words, 72, 97
Concept, 65
Conceptualization, 6
Confidence, 176
Consistency, 124

Corpus of geographic
knowledge, 20
Crisp object, 140
Crust, 86

D

Data completeness, 110
Data integration, 123
Data quality, 40
Decision-making, 40, 104, 115
Degree distribution, 191
Delaunay triangulation, 80, 190
Differential equations, 258
Dual graph, 190
Duality, 82
Dynamic algorithm, 89
Dynamic attributes, 173
Dynamics, 151

E

Economic performance, 268
Emergency plan, 260
Epistemology, 3
Error, 101
Error propagation, 104, 108
Euclidean space, 84
Eudaimonia, 31
Event, 208
Extension principle, 98

F

Field, 82
Fitness for use, 109, 137
Flooding application, 142

- Folksonomy, 71
- Formal ontologies, 123
- Fuzzy object, 140
- Fuzzy representation, 107
- Fuzzy set, 148

- G**
- Generalized Constraint Language, 72, 97
- Geodetic datum, 70
- Geographic information, 3
- Geographic knowledge, 225
- Geonoemata, 17
- Geosemantics, 20
- Geosensor, 206
- Geospatial training games, 254
- Geostatistics, 115
- Granulation, 42, 97
- Graph, 180
- Grounding, 7, 69

- I**
- Image mining, 136
- Imperfection, 39
- Imprecise possibility distribution, 95
- Imprecise probability distribution, 95
- Incompleteness, 105
- Incremental algorithm, 85
- Information content, 237
- Information object, 5
- Institution, 273
- Intelligent agent, 245
- Interoperability, 123, 252

- K**
- Kinetic algorithm, 89

- L**
- Land administration, 266
- Latent Dirichlet Allocation, 123
- Learning cycle, 253
- Level of detail, 157
- Lineage, 102
- Linear referencing, 183

- Location-based services, 169
- Logical consistency, 102, 104

- M**
- Metadata, 101, 111, 124
- Metaquality, 102
- Model completeness, 110
- Monte Carlo uncertainty propagation analysis, 115
- Moving object, 154, 171
- Moving point, 171

- N**
- Natural language, 96
- Natural language processing, 23
- Network, 179, 189
- Networks of constraints, 68
- NL-Computation, 97

- O**
- Object, 42, 81, 208
- Observation, 39
- Ontological constraint networks, 64
- Ontology, 3, 20, 40, 63 top-level, 21

- P**
- Personalizing, 7
- Planar graph, 191
- Positional accuracy, 102, 104
- Positional error, 107
- Practical wisdom, 31, 35
- Precisiation, 97
- Predicates, 173
- Primacy of institutions, 272
- Probabilistic latent analysis, 123
- Profession, 33
- Propagation of variances, 108
- Property rights, 273, 274
- Purpose, 4

- Q**
- Quad-Edge, 85, 90
- Qualitative knowledge, 10
- Qualitative spatial relations, 210

Quality of institutions, 273
 Query language, 174

R

Random noise, 41, 45
 Random planar graph, 192
 Reason, 4
 Remote sensing, 135
 Representation, 6
 Responsibility, 29
 Risk management, 116
 Route information, 236

S

Sampling theorem, 53
 Scale, 52
 Scale-free graph, 190
 Semantic accuracy, 102
 Semantic conflicts, 18
 Semantic consistency, 124
 Semantic engineering, 63
 Semantic reference system, 21
 Semantics, 8, 64, 156, 209
 generalized-constraint-based, 97
 Semantic similarity, 131
 Semantic triangle, 66
 Serious games, 251
 Similarity, 124
 Skeleton, 86
 Social construction, 43
 Spatial data quality, 101, 148
 Spatial decision making, 219
 Spatial granularity, 157
 Spatial partition, 177
 Spatial patterns, 137
 Spatio-temporal
 behavior, 154
 data, 253

data types, 173
 features, 170
 predicates, 173

Stereology, 135

T

Taxonomy, 152
 Teleology, 3
 Temporal accuracy, 102
 Terrain model, 82
 TerraLib, 211
 TerraME, 211
 Tessellation, 79
 Time, 153, 171, 246, 258
 Time representation, 262
 Time scale, 260
 Topological relationship, 172
 Topology, 79, 82
 Trajectory, 174, 208
 Triangulated Irregular Network, 86
 Turing test, 240

U

Unbiasedness, 138
 Uncertainty, 72, 95, 112, 136, 176
 Uncertainty management, 115
 Undirected graph, 189

V

Vague object, 140
 Virtue ethics, 27
 Visualization, 108
 Vocabulary, 224
 Voronoi diagram, 80, 190

W

Wayfinding communication, 245