LECTURE NOTES IN GEOINFORMATION AND CARTOGRAPHY

LNG&C

A. Ruas · C. Gold (Eds.)

Headway in Spatial Data Handling

Springer

Lecture Notes in Geoinformation and Cartography

Anne Ruas · Christopher Gold (Eds.)

Headway in Spatial Data Handling

13th International Symposium on Spatial Data Handling

Springer

Editors

Dr. Anne Ruas
Inst. Géographique National-IGN
COGIT Laboratoire
2-4 avenue Pasteur
94166 Saint-Mandé Cedex
France
anne.ruas@ign.fr

Dr. Christopher Gold
University of Glamorgan
Dept. Mathematics and
Computing
Pontypridd, M. Glam.
United Kingdom CF37 1DL
cmgold@glam.ac.uk

ISBN: 978-3-540-68565-4 e-ISBN: 978-3-540-68566-1

Lecture Notes in Geoinformation and Cartography ISSN: 1863-2246

Library of Congress Control Number: 2008928568

Typesetting: Camera-ready by the editors

Cover design: deblik, Berlin

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Foreword

Geographic information is a key element for our modern society. Put simply, it is information whose spatial (and often temporal) location is fundamental to its value, and this distinguishes it from many other types of data, and analysis. For sustainable development, climate change or more simply resource sharing and economic development, this information helps to facilitate human activities and to foresee the impact of these activities in space as well as, inversely, the impact of space on our lives. The International Symposium on Spatial Data Handing (SDH) is a primary research forum where questions related to spatial and temporal modelling and analysis, data integration, visual representation or semantics are raised.

The first symposium commenced in 1984 in Zurich and has since been organised every two years under the umbrella of the International Geographical Union Commission on Geographical Information Science (http://www.igugis.org).

Over the last 28 years, the Symposium has been held in:

1st - Zürich, 1984
2nd - Seattle, 1986
3rd - Sydney, 1988
4th - Zurich, 1990
5th - Charleston, 1992
6th - Edinburgh, 1994
7th - Delft, 1996
8th - Vancouver, 1998
9th - Beijing, 2000
10th - Ottawa, 2002
11th - Leicester, 2004
12th - Vienna, 2006

This book is the proceedings of the 13th International Symposium on Spatial Data Handling. The conference was held in Montpellier, France, on June 23rd to 25th 2008, in conjunction with and prior to SAGEO, the annual French conference on Geomatics. All the papers in this book were submitted as full papers, and received blind reviews from three members of the Programme Committee. 63 papers were submitted and the 36 that are included here are of a high standard, as well as being accessible to everyone who is doing research in the science of geographical information. This year we have chosen to promote the five best reviewed papers that can be noted by a * on the table of contents. These very enthusiastic and original papers deal with ontology, classification, data matching, 3D and spatial process description.

This publication ‘Headway in Spatial Data Handling’ is the fourth in the Springer-Verlag series which members of the commission hope will continue successfully in the future.

Acknowledgements

This conference has been organised by a network from four geographical locations: Montpellier (Thérèse Libourel), Brest (Thomas Devogele), Paris (Anne Ruas) and Wales (Chris Gold).

We thank all the members of the Programme Committee for their extraordinary effort to return detailed reviews on time.

We warmly want to thank Nicolas Moyroud for helping us with the web site with a constant and positive availability, Thomas Devogele for editing and for his contact with the publisher and Therese Libourel for her energy in organising the conference in Montpellier (and for editing the Latex files).

A special thank you to Victor for enhancing the pleasure of our working together.

Anne Ruas and Chris Gold
31 March 2008

Table of Contents

Shape and Spatial Relations 1

Classification

Classification and Image Analysis

Process Modelling

Generalisation and Multiple Representation

3D and Relief

3D

Ontology

Uncertainty and Matching

Shape and Spatial Relation 2

Road and Navigation

Programme Committee

Chairs
Anne Ruas and Chris Gold

Dave Abel
Bénédicte Bucher
Eliseo Clementini
Alexis Comber
Thomas Devogele
Gregory Elmes
Peter Fisher
Andrew Frank
W. Randolph Franklin
Michael Goodchild
Hans Guesgen
Francis Harvey
Claire Jarvis
Bin Jiang
Christopher B. Jones
Brian Klinkenberg
Menno-Jan Kraak
Michael Leitner
Chao Li

William Mackaness
Paola Magillo
Hervé Martin
Martien Molennar
Sébastien Mustière
Peter van Oosterom
Henk Ottens
Donna Peuquet
Ross Purves
Sanjay Rana
Andreas Riedl
Juri Roosaare
Monika Sester
Bettina Speckmann
Monica Wachowicz
Robert Weibel
Qihao Weng
Anthony Yeh
Sisi Zlatanova

Local Organizing Committee

Chairs
Thérèse Libourel and Thomas Devogele

Céline Berger
Patrick Bisson
Jean-Paul Bord
Michel Casteigts
Stefano Cerri
Laurent Chapelon
Jean-Pierre Chery
Jean-Christophe Desconnets
Michel Deshayes
Frédéric Huynh

Pascal Kosuth
Arnaud Martin
Pierre Maurel
André Miralles
Isabelle Mougenot
Nicolas Moyroud
Michel Passouant
Michel Sala
Georges Schmitt
Jean-Philippe Tonneau

A Study on how Humans Describe Relative Positions of Image Objects

Xin Wang[1], Pascal Matsakis[1], Lana Trick[2], Blair Nonnecke[1],
Melanie Veltman[1]

[1] Department of Computing and Information Science,
University of Guelph, Guelph, Ontario Canada, N1G 2W1
e-mail: {xin, matsakis, ltrick, nonnecke, mveltman} @uoguelph.ca

[2] Department of Psychology
University of Guelph, Guelph, Ontario Canada, N1G 2W1
e-mail: matsakis@cis.uoguelph.ca

Abstract

Information describing the layout of objects in space is commonly conveyed through the use of linguistic terms denoting spatial relations that hold between the objects. Though progress has been made in the understanding and modelling of many individual relations, a better understanding of how human subjects use spatial relations together in natural language to is required. This paper outlines the design and completion of an experiment resulting in the collection of 1920 spoken descriptions from 32 human subjects; they describe the relative positions of a variety of objects within an image space. We investigate the spatial relations that the subjects express in their descriptions, and the terms through which they do so, in an effort to determine variations and commonalities. Analysis of the descriptions determines that common elements of spatial perception do indeed exist between subjects, and that the subjects are quite consistent with each other in the use of spatial relations.

Keywords: Spatial relations, natural language, spatial cognition, human information processing

1 Introduction

Spatial information is understood and conveyed through the use of spatial relations, which describe how one object in a scene or an image is located in relation to some other object. Spatial relations have been studied in a number of different disciplines, including computer science, geographic information science, cognitive science and linguistics. Within these disciplines, spatial relations are generally considered to fit into one of three categories: topological, including relations like *OVERLAP* and *SEPARATE*; directional, including relations like *ABOVE*, *BELOW*, *RIGHT* and *LEFT*; and distance, including relations like *NEAR* and *FAR*. In natural language, relations such as these are referred to using a variety of different terms and phrases. Most spatial terms, including ones we are very familiar with, like *near* or *beside,* possess semantics which are far more nuanced than might be expected at first glance.

The perception of spatial relations is determined by many factors, including the point of observation and any intrinsic axes of image objects (Herskovits 1986); and any additional objects or associated context (Regier 1992). Additionally, the dimensionality of the space and objects in question determines the relations that can be used – in a 3D space, relations like *BEHIND* and *IN FRONT OF* may be appropriate. In addition to these factors, individual differences like gender (Linn & Petersen 1985) and handedness (Halpern 1986; Mark *et al.* 1995) may affect how one forms mental models of spatial phenomena and assigns meanings to spatial concepts. Anthropologist Hall (1966) found that a subject's experience of space, and hence perception of spatial relations, are affected by culture. This was confirmed by Montello (1995), in his critical discussion of the significance of cultural differences in spatial cognition. Based on these factors, two human subjects may perceive the same concept quite differently, and hence describe it differently (Mark *et al.* 1994; Mark *et al.* 1995; Worboys 2001). As early as 60 years ago, Whorf and Sapir (1940) proposed that language influences or constrains the way in which people think; this work is known as the Sapir-Whorf hypothesis. Although it is not clear whether or how such effects apply to spatial relations, there are significant distinctions in the use of spatial terms in different languages. A number of cognitive and linguistic scientists have informally described how spatial relations are expressed in natural language. Talmy (1983) suggested that in linguistic descriptions, the spatial disposition, i.e., the site and orientation, of one object, referred to as the *figure* (or *argument*), is always characterized in terms of one or more other objects selected from the remainder of the scene, referred to as the *ground* (or *reference*). The ground objects are used as a fixed reference from which the position of the figure is described. Talmy also pointed out that any natural language has only a limited number of words available for describing the spatial relations in an infinite number of spatial layouts, and each of these words actually represents a family of layouts that all share certain abstract characteristics.

Additionally, some research into the computational modelling of spatial relations has taken human perception into account. The most common method of capturing human perception is as follows: To begin, subjects are presented with a small number (usually less than 100) of images, referred to as *configurations*, containing basic shapes. Subjects are then given a set of spatial relations (usually less than 10), and for each configuration, they are required to either answer a series of yes or no questions (i.e., whether a given relation describes the configuration (Robinson 1990)), or to rate a list of spatial relations based on how well they describe the configuration individually (Gapp 1995; Wang & Keller 1997, 1999; Zhan 2002). The weaknesses in these methods are obvious. Firstly, the data used to train the system is collected by having all subjects describe a small number of configurations using the same relations, and consequently, only a small number of relations and configurations are applicable. Secondly, as pointed out by Landau (1996), the English lexicon of spatial prepositions numbers above eighty members, not considering terms used to describe compound spatial relations, or uncommon relations. Hence, it would be implausible for a given subject to test all of these relations for each configuration, yet any practices of limiting spatial relations to a given smaller set may bias the subject and subsequently, the results. Additionally, spatial relations are not independent from each other, rather, a variety of relations occur and interact in a given natural language description. We are more interested in which relations users refer to when describing an image, and how they use them together (i.e., in the context of image retrieval).

The goal of this research is to design and carry out an experiment to better understand how human subjects naturally describe the relative positions of objects in a series of images. The primary research question is: when describing different *configurations*, what spatial relations do people refer to, and how (i.e., using what terms)? In order to design the experiment so that we capture human perception of a number of different relations in a variety of circumstances, we must collect descriptions of a large number of varied configurations, desirably more than 1000. Under this condition, obtaining enough information from one subject would prove a cumbersome task, and we investigate the use of data collected from multiple subjects. Even though cognitive studies show that individual differences in spatial perception exist among different people, we hypothesize that common elements of perception may be found among different subjects (if this were not the case, humans would not be able communicate spatial concepts to one another). Thus, two more questions arise: How significant are the variations between descriptions given by different subjects? What are the common elements of perception? To ascertain commonalities and variations in spatial perception, we are interested in measuring the consistency of use of spatial relations in descriptions given by numerous subjects; for any given configuration, our focus is on the presence of spatial relations in subjects' descriptions.

The remainder of this paper is organized as follows: Section 2 describes the experiment design. Section 3 outlines the collection of the descriptions and the extraction of relations from these descriptions. Section 4 discusses the results of analysis and Section 5 concludes the paper with discussion of limitations and future work.

2 Experiment design

Assume we want to design a computer system capable of providing linguistic descriptions of relative positions of image objects the way a given person would. A system like this would be of use in a number of practical applications. In many robot vision scenarios, the robot's understanding of its environment will include some representation of spatial information, and the robot must then communicate this to a human user. Similarly, some Content Based Image Retrieval (CBIR) systems use spatial information in indexing and natural language in searching. In both of these scenarios, training a system to provide accurate, human-like descriptions will increase the quality of the interaction. As mentioned previously, obtaining enough information for system training from one subject is not reasonable, and we investigate the concept of a prototypical perception, based on common elements found in descriptions from different subjects. If we can determine a prototypical perception, and train the system using this data, a later point in time, the system could be fine-tuned to one individual perception through training with one user.

Hence, the goal of this research is to determine any common elements of perception, and the significance of any variations between key elements of descriptions provided by different subjects. We aim to validate the existence of a prototypical perception, and gain a better understanding of what the elements of this perception are (i.e., the spatial relations and corresponding linguistic terms people commonly refer to). We also create and demonstrate an appropriate method for collecting natural language descriptions that could be used to train a system that would be able to learn and generate descriptions according to a prototypical perception.

2.1 Generating configurations

In an effort to simplify the problem and eliminate the influence of factors such as point of observation and dimension, two and only two 2D image objects are considered. The two objects are abstract shapes, with no intrinsic axes or context associated with them. An object pool containing 25 different shapes was created, and is illustrated in Figure 1. The shapes represent regular and non regular convex shapes (O1 to O10), and simple and more complex concave shapes (O11 to O25).

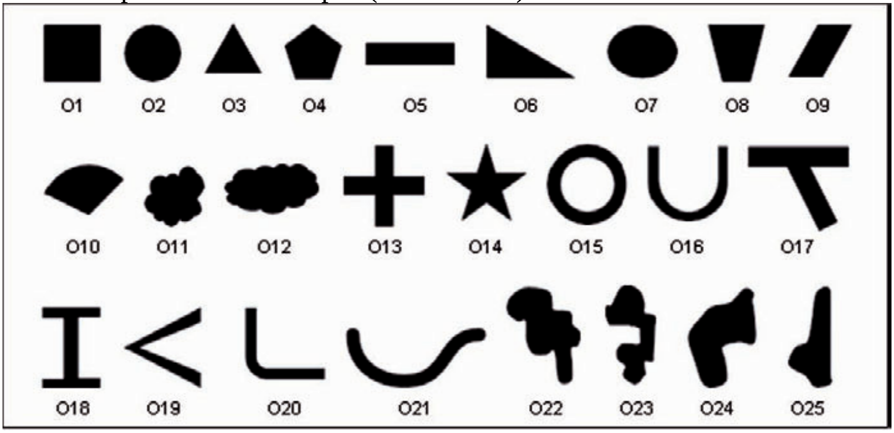


Fig. 1. The Object Pool

Using these 25 shapes, configurations were then generated using the following method: First, we randomly draw two objects from the object pool, with replacement. Secondly, each object is zoomed by a random zooming factor μ between 30% and 300%, and rotated by a random angle θ. Note that the values of μ and θ may be different for the two objects. One object, selected randomly, is then coloured grey and the other is coloured black. Finally, the transformed objects are randomly placed inside the image space, a 500 x 500 white background, 20 times to create a set of 20 configurations containing the same objects. If there is an intersection between the two objects, it is coloured a dark grey (between the grey shade and the black). The above method is repeated to create 68 such configuration sets, and thus we have in total 1360 configurations for experimentation.

2.2 Describing Task

Because we are interested in natural language descriptions (i.e., we do not want to constrain the subjects by providing a check list of terms or relations), we must properly design the experiment and communicate the describing task to the subjects in such a manner as to ensure that they focus their descriptions on relevant spatial information. To achieve this, each subject is tasked with describing configurations in the context of a game that loosely emulates image retrieval, as illustrated in Figure 2. The game is described to the subject using the following scenario: “Imagine that you are playing a game with a friend. You have a set of configurations and your friend has another set. For each configuration in your set, there are one or more similar configurations in your friend’s set. Now, imagine that you are on the phone with your friend. He/she cannot see your set, and you cannot see his/hers. Please describe the configuration shown on the computer screen so that your friend is able to find similar configuration(s) in his/her set.” Note that terms like size, shape, relative position, and absolute position are actually never pronounced, and no verbal example is given (only visual examples), thereby minimizing potential biases.

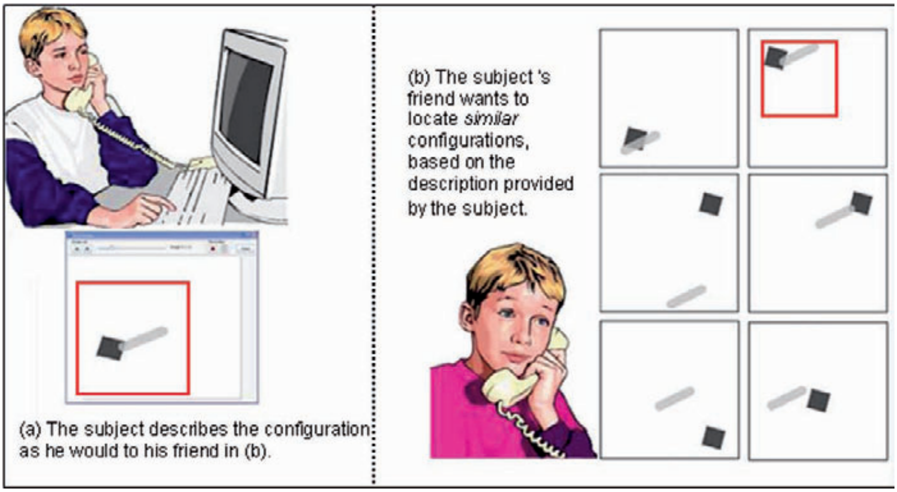

(b) The subject 's friend wants to locate *similar* configurations, based on the description provided by the subject.
(a) The subject describes the configuration as he would to his friend in (b).

Fig. 2. The Image Retrieval Game

Through learning the concept of the game, the subject is made aware of the following important characteristics of their task:

1. The objects are of the same size, shape, and colour in all configurations within a configuration set; with this in mind, the subject will presumably avoid describing the objects features, and instead will describe where the objects are.

2. For two configurations to be considered similar, the objects can be anywhere in the image, as long as their relative positions are the same, i.e., as illustrated in Figure 2, the objects may be shifted together in the space. Knowing this, the subject will presumably avoid providing information about the absolute positions of the objects, in favour of describing relative positions.
3. The relative positions of the objects in similar configurations may not be identical, just similar; with this in mind, the subject will presumably avoid excessively long or detailed descriptions.

3 Collecting descriptions and extracting spatial information

3.1 Collecting Descriptions

Approval to conduct the experiment for data collection was granted by the Research and Ethics Board at the University of Guelph in April of 2007. 32 participants, ranging in age from 18 to 45 years, took part in the experiment. To eliminate influence of language and cultural factors, it was required that subjects' first language be Canadian English.

After being introduced to the concept of the image retrieval game, each of these subjects described a total of 60 configurations from six different sets. 40 of the configurations were from 2 sets unique to each subject; these 1280 descriptions (32 participants x 40 configurations) provide sufficient information for system training, and allow us to obtain some statistics on spatial relations used and terms used to refer to them. Additionally, each subject was tasked with describing the first 5 configurations in 4 configuration sets common to all subjects. The 640 descriptions of the common configurations provide information for the study of consistencies and variations among subject's descriptions. These allow us to validate the existence of a prototypical perception, and get an idea of what elements are involved in this prototypical perception.

3.2 Extracting Target Information

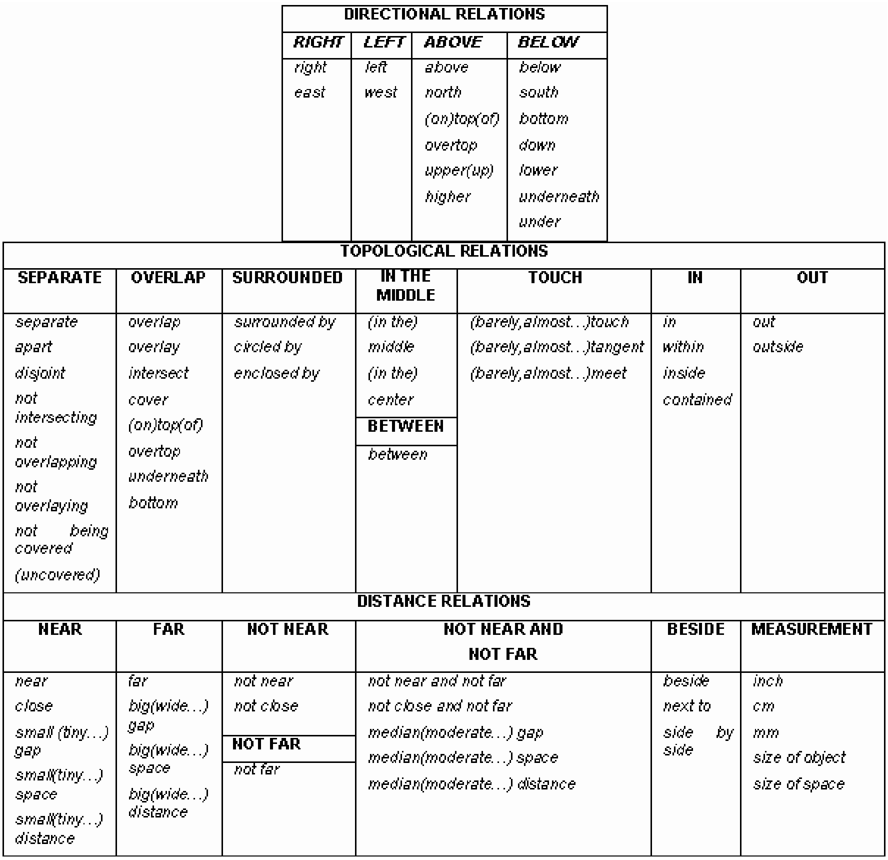
In total, 1920 spoken descriptions were collected. To capture from these spoken descriptions the information we are interested in, manual processing is required. In order to reliably extract and encode this information for analysis, constraining procedures must exist at this point. The extraction task was modelled based on the following observations, made in a preliminary review of the descriptions.

Not surprisingly, we found the descriptions to be in a variety of forms and grammar structures, and some of the information provided to be irrelevant or unusable. We specify *target information* as information about the position of one object (the argument) relative to another (the referent). The following description is an example of what is considered target information: “The grey object is to the right and below the black object.” In this description, the reference object is the black object, and the relative position is described by the terms *right* and *below*, which denote spatial relations. Non-target information includes the following:

- Information about the shapes, sizes and orientations of objects. For example: “There are two objects. They are both star-like shapes. But the grey object is smaller than the black object.” Because the goal of the description task is to describe the relative position of the objects, all of the information provided in this description is irrelevant.
- Information about the absolute positions of objects. For example: “The grey image is on the right side, towards the top of the page.” The information provided by this description, about the grey object’s absolute position, is irrelevant to the task of finding one or more *similar* configurations.
- Information related to or dependent on other configurations. For example: “The gap [between the objects] is much smaller than in the previous image.” This information is relevant, but it is not exploitable; within the context of this work, one and only one configuration is considered at a time.
- Information that is confusing or involves the use of abstract concepts. For example: “If a vertical line rejoins at the lower edge point of the dark object, it will pass through the lower center point of the light object.” The information provided by this description may be relevant, but it is not usable, because spatial relations are not explicitly involved, nor can they be reliably implied.

Many different terms were used to describe the same relation, i.e., *north* and *higher* both refer to the spatial relation *ABOVE*. We counted more than 50 distinct terms in the initial review of the descriptions, not including grammatical variations (e.g., *intersect, intersecting*), negative expressions (e.g., *not near, no overlap*), or linguistic hedges (e.g., *barely, almost*). A preliminary list of the most commonly used terms, denoting 19 different relations was generated. These *prelisted relations*, and their associated terms and categorizations, are provided in Table 1.

Table 1. The Prelisted Relations



DIRECTIONAL RELATIONS			
RIGHT	***LEFT***	***ABOVE***	***BELOW***
right *east*	*left* *west*	*above* *north* *(on)top(of)* *overtop* *upper(up)* *higher*	*below* *south* *bottom* *down* *lower* *underneath* *under*

TOPOLOGICAL RELATIONS						
SEPARATE	**OVERLAP**	**SURROUNDED**	**IN THE MIDDLE**	**TOUCH**	**IN**	**OUT**
separate *apart* *disjoint* *not intersecting* *not overlapping* *not overlaying* *not being covered* *(uncovered)*	*overlap* *overlay* *intersect* *cover* *(on)top(of)* *overtop* *underneath* *bottom*	*surrounded by* *circled by* *enclosed by*	*(in the)* *middle* *(in the)* *center* **BETWEEN** *between*	*(barely,almost...)touch* *(barely,almost...)tangent* *(barely,almost...)meet*	*in* *within* *inside* *contained*	*out* *outside*

DISTANCE RELATIONS					
NEAR	**FAR**	**NOT NEAR**	**NOT NEAR AND NOT FAR**	**BESIDE**	**MEASUREMENT**
near *close* *small (tiny...) gap* *small(tiny...) space* *small(tiny...) distance*	*far* *big(wide...) gap* *big(wide...) space* *big(wide...) distance*	*not near* *not close* **NOT FAR** *not far*	*not near and not far* *not close and not far* *median(moderate...) gap* *median(moderate...) space* *median(moderate...) distance*	*beside* *next to* *side by side*	*inch* *cm* *mm* *size of object* *size of space*

We found that in some descriptions, spatial relations are implied. Consider the following description: “The grey object is overlapping the bottom left part of the black object.” Although the subject has not said that the grey object is *BELOW* or to the *LEFT* of the black object, one might accurately deduce that this is the case, based on the description of the overlapping regions. Similarly, it was also observed that subjects at times referred to spatial relations between object parts (“The grey object lies to the left of *the upper half of* the black object”), as opposed to considering the objects in general (“The grey object lies to the left of the black object”).

Based on these observations, an interface was developed to allow a Research Assistant (RA) to extract target information by answering the following questions while listening to each description:

Q1. Does the description contain any target information?

Q2. Which object is the reference and which one is the argument?

Q3. Does the description involve any of the prelisted spatial relations? What terms are used to describe the relation?

Q4. Is the relation referred to explicitly or implicitly?

Q5. Is the relation between parts of the objects, or the entire objects?

Q6. Does the description involve relations that are not in the provided list? For each non-prelisted relation: Is the relation referred to explicitly or implicitly? Is the relation between parts of the objects, or the entire objects?

To assist the RA in his/her task, the interface provides the prelisted relations and associated terms illustrated in Table 1. However, these lists of relations and terms are by no means exhaustive, and the RA is trained and encouraged to extend them. The interface allows for the audio description to be paused and repeated, to allow the RA to correctly perform his/her task. Viewing the configuration is possible, but is strongly discouraged - the RA is instructed that this option should be accessed only when he/she feels that further clarification on a description is required. This encourages reliable extraction of the descriptions provided, and minimizes potential biases.

Although most descriptions maintain a consistent reference object, there are some in which it is not stated explicitly which object is the reference (e.g., "The black and grey objects are intersecting each other"), or the objects are used alternately as the referent (e.g., "The black object is below the grey object. The grey object is close to the black object"). For consistency, in both of these cases, the RA is instructed to select the *black* object as the referent by default. In the second case, the RA must also enter what he/she deems to be the *semantic inverse* (Freeman, 1975) of any relations in which the grey object is used as the reference object. In the example above, the RA will choose the relations *ABOVE* (the semantic inverse of *BELOW*) and *CLOSE*. Although this inversion step requires some extra effort on the part of the RA, it is required for comparison of descriptions of the same configuration provided by different subjects.

Although the information collected pertaining to Q4, Q5, and Q6 is not involved in the current analysis, we do plan to use it in future research. Also, in the framework of this work, linguistic hedges are ignored, and so is the order in which spatial relations are referred to in the description. For example, no distinction is made between "The grey object is *to the right* of the black object" and "The grey object is *mostly to the right* of the black object". Also, no distinction is made between "The grey object is *to the right and below* the black object" and "The grey object is *below and to the right* of the black object." In an effort to ensure reliability in the results, two RAs were assigned to process each of the 1920 descriptions independently, resulting in two independent data sets.

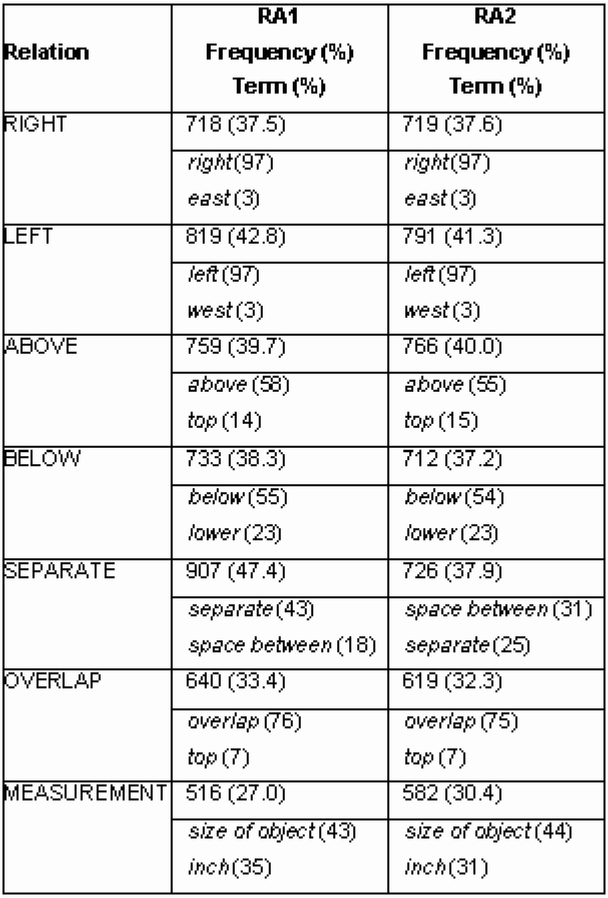
4 Data analysis

4.1 Agreement on spatial information extraction

Each RA found that 1914 of the 1920 spoken descriptions contain target information (Q1). The computed average number of spatial relations provided in each of these descriptions is 3.26 in the first RA's data set, and 3.2 in that of the second RA. The minimum number of relations provided in a description is 1 in both data sets, and the maximum number of relations is 4 in the first RA's data set, and 6 in that of the second RA. The most frequently used prelisted relations, along with the two most commonly used terms for each one, are illustrated in Table 2. The relation *RIGHT* was used in 718 descriptions (37.5% of the 1914 descriptions) according to the first RA, and the second RA extracted *RIGHT* from 719 descriptions (37.6% of the 1914). Both RAs found that of the descriptions that included reference to the relation *RIGHT*, the term *right* was used in 97%, and the term *east* was used in 3%. Clearly, from Table 2, the two RAs reach strong agreement that directional relations dominate over topological and distance relations in terms of frequency of use. Many of the discrepancies that exist in this table can be explained by the implicit attribute of some relations provided in descriptions, i.e., for a given description, one RA may have deduced some relations from the description that the other RA did not.

In 1790 (93.5%) of the 1914 descriptions containing target information, the RAs agreed on which object is the reference object(Q2), and the two objects are seldom used alternately as the referent (2.3%). Since it is difficult to judge whether the RAs agreed on spatial information extraction (Q3 to Q6) when they have failed to agree on the referent, we based further analysis about the agreement between the data sets on 1790=1914-124 descriptions. These 1790 descriptions, along with the 19 listed spatial relations, resulted in a total of 34010 answers to question Q3. The RAs gave congruent positive answers (i.e., the relation was included in the description) in 4949 cases (14.6%), and congruent negative answers (i.e., the relation was not found in the description) in 28075 cases (82.6%); thus agreement was achieved on the answer to question Q3 in 97.2% of the cases.

Table 2. The Frequency of Use of the Prelisted Relations



Relation	RA1 Frequency (%) Term (%)	RA2 Frequency (%) Term (%)
RIGHT	718 (37.5)	719 (37.6)
	right(97) *east* (3)	*right*(97) *east* (3)
LEFT	819 (42.8)	791 (41.3)
	left (97) *west* (3)	*left* (97) *west* (3)
ABOVE	759 (39.7)	766 (40.0)
	above (58) *top* (14)	*above* (55) *top* (15)
BELOW	733 (38.3)	712 (37.2)
	below (55) *lower* (23)	*below* (54) *lower* (23)
SEPARATE	907 (47.4)	726 (37.9)
	separate (43) *space between* (18)	*space between* (31) *separate* (25)
OVERLAP	640 (33.4)	619 (32.3)
	overlap (76) *top* (7)	*overlap* (75) *top* (7)
MEASUREMENT	516 (27.0)	582 (30.4)
	size of object (43) *inch*(35)	*size of object* (44) *inch*(31)

4.2 Agreement between descriptions

In measuring the agreement between the relations extracted from each description, we focus our analysis on only the prelisted relations. We also consider knowing whether or not a relation is involved in the description more important than knowing the specific attributes (the term used, whether the relation was referred to explicitly or implicitly, and whether the relation was between whole objects or object parts) of that relation. For each congruent description, we say that two RAs reach agreement on spatial relations only if the set of spatial relations extracted by the first RA, S1, and that by the second RA, S2, satisfy the following condition: S1⊆ S2 or S2⊆ S1 (S1≠∅ and S2≠∅ because we are considering congruent descriptions). In total, 1617 (90.3%) of the 1790 congruent descriptions are considered to have such agreement, and therefore are regarded to be *reliable*. For each reliable description, we can then merge the information extracted from both RAs, and let S = S1∩S2 be the merged set of spatial relations associated with the description. The actual term used for each relation is discarded, and if the relation is implicitly referred to according to S1 or S2, then it is considered implicitly referred to in S; this is because explicit is the default, and the RA must intentionally input that the relation is implicit, so we consider that it is likely not done in error. For the same reason, if the relation is between object parts according to S1 or S2, it is considered between object parts in S.

4.3 Inter-subject variations

Inter-subject variations reflect how different subjects describe configurations, and are measured based on the descriptions of the 20 configurations that are common to all subjects, illustrated in Figure 3. Although all 32 participants provided descriptions for these common configurations, some are not considered to be reliable. The results presented here are therefore based on the remaining 563 reliable descriptions. The number of reliable descriptions for each of the 20 configurations varies from 23 to 30, with the average number 28.

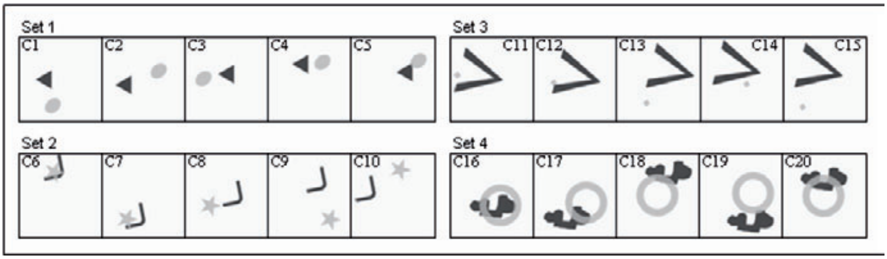


Fig 3. The Common Configurations

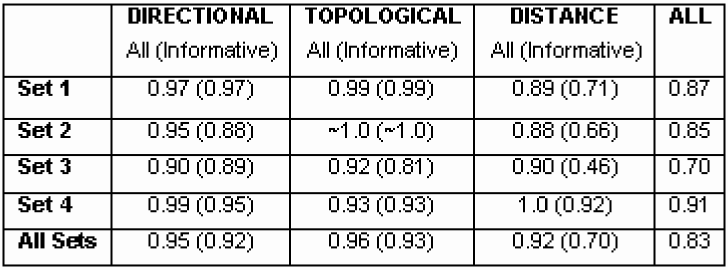
We first analyzed the data looking for tendencies in the subjects' use of different types of spatial relations, e.g., directional, topological and distance relations. The number and category of the specific relations provided in each of the 563 merged strings were counted. From the number of *N* reliable descriptions for a given configuration, we count the number of times spatial relations in a given category are used. We then normalize these values to percentages to determine the frequency of use. A high percentage (close to 100%) means that nearly every subject included this type of relation in his/her description of the configuration. A low percentage (close to 0%) denotes that very few of the subjects involved this type of relation. In both of these cases, the subjects have similar tendencies. A median percentage (close to 50%), however, indicates the subjects have very different tendencies, since about half of them involve this type of relation and about half of them do not.

We found that subjects' tendencies to involve different types of spatial relations vary from configuration to configuration and from set to set. For example, subjects' tendencies to involve directional relations, while they are similarly strong for configurations in Sets 1 and 3 (the average frequencies for directional relations are 99% and 97% respectively) are weak for C6 and C16 (only 49% and 39% of the descriptions for these configurations involved directional relations). Not surprisingly, for these two particular configurations, 93% and 100% of the descriptions involved topological relations. The subjects seldom involve distance relations when describing configurations in Set 4, and furthermore, none of the subjects mention distance relations when describing C16. It seems that subjects have similar tendencies to involve directional relations throughout describing different sets of configurations, however, they have very different tendencies to involve topological and distance relations when describing, especially, the first three sets of configurations. One possible reason is that the first three sets are simpler than the fourth, causing some subjects to only use one or two types of spatial relations when describing them.

To expand further on these observations, the consistency between subjects in their use of specific relations, and categories of relations, is computed for each of the 20 configurations. To do so, we consider two subjects, A and B, who have both provided a reliable description for a given configuration, where A's description provides *P* relations, and B's description provides *Q* relations. The consistency between the relations provided by A and B is defined as c(A,B) = $R/min(P,Q)$, where *R* is the number of relations common to both descriptions. The value of c(A,B) reaches 1 when the set of relations used by one subject entirely includes that used by the other. The consistency in the use of a given category of relations is computed in a similar fashion: $c_{\mathrm{TYPE}}(A,B) = R_{\mathrm{TYPE}} / \min(P_{\mathrm{TYPE}},Q_{\mathrm{TYPE}})$. In cases where one subject does not provide any relations of a certain type, $\min(P_{\mathrm{TYPE}},Q_{\mathrm{TYPE}}) = 0$, and $c_{\mathrm{TYPE}}(A,B)$ is set to 1, because one subject's failure to include a certain type of relation that was provided by the other does not constitute disagreement. The consistency between subjects in describing the same configuration is measured as the average of the consistencies between all possible pairs of subjects. For instance, the consistency between three subjects A, B and C in their use of topological relations is $[c_{TOP}(A;B) + c_{TOP}(A;C) + c_{TOP}(B;C)] / 3$. For *n* subjects, the consistency is the average of $n(n\text{-}1)/2$ values. According to this formula, we calculate the consistency between all of the subjects in their use of directional relations (c_{DIR}), topological relations (c_{TOP}), distance relations (c_{DIS}), and all relations together (c).

Overall, the subjects involve directional relations consistently, i.e., consistency >90%, in merged descriptions of configurations C1 through C5, C8 through C10, and C13 through 15; for 9 of these 11 configurations, consistency of directional relations is 100%. Topological relations were consistently involved only in the merged descriptions of configurations in Set 4, and distance relations were consistently involved only in descriptions of C6. In cases where only a small number of subjects provide inconsistent relations of a given type, the type consistency will remain quite high, and a measure of the consistency between only the informative descriptions (descriptions actually involving the type of relation) is more appropriate. When only the informative descriptions were considered, the average consistencies in the use of directional and topological relations did not change much, but the average consistencies in the use of distance relations decreased significantly. Table 3 illustrates the average consistencies between the subjects in their use of different types of relations in descriptions of the common configuration sets, and how these consistencies vary when only informative descriptions are considered.

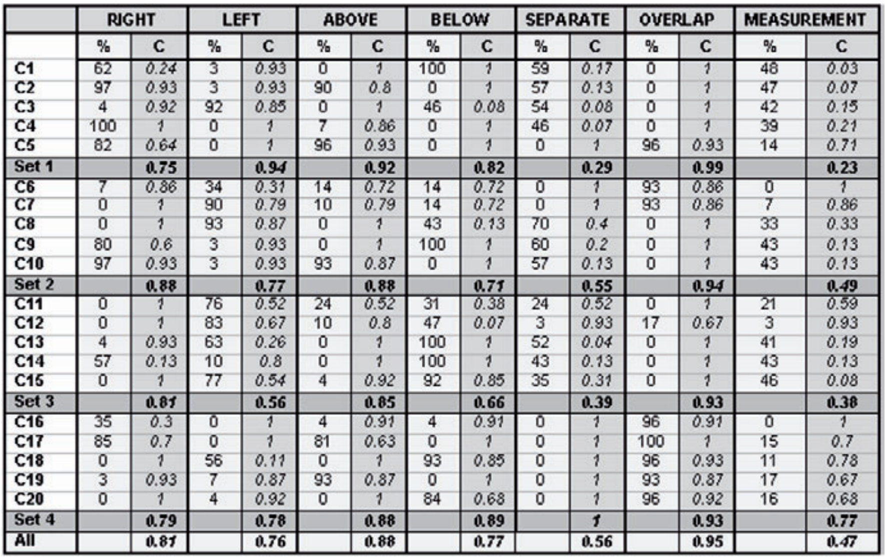
Table 3. Consistency of Use of Types of Relations



	DIRECTIONAL All (Informative)	**TOPOLOGICAL** All (Informative)	**DISTANCE** All (Informative)	**ALL**
Set 1	0.97 (0.97)	0.99 (0.99)	0.89 (0.71)	0.87
Set 2	0.95 (0.88)	~1.0 (~1.0)	0.88 (0.66)	0.85
Set 3	0.90 (0.89)	0.92 (0.81)	0.90 (0.46)	0.70
Set 4	0.99 (0.95)	0.93 (0.93)	1.0 (0.92)	0.91
All Sets	0.95 (0.92)	0.96 (0.93)	0.92 (0.70)	0.83

Next, we investigate if the subjects involve the same spatial relations when describing the configurations in the common set of 20. Note that there are 19 prelisted relations, but because none of descriptions of the common configurations make use of the relations *BESIDE* and *NOT FAR*, we only count occurrences of 17 relations. As can been seen from Table 4, which illustrates only the frequencies and consistencies of the most commonly used relations, the subjects have similar tendencies to involve most spatial relations, except for the relations *SEPARATE* and *MEASUREMENT*. For example, 59% of descriptions of C1 make use of the relation *SEPARATE*, and the consistency between subjects for this relation in describing C1 is 0.17.

Table 4. Frequency and Consistency of Relations in Descriptions of Common Configurations



	RIGHT		LEFT		ABOVE		BELOW		SEPARATE		OVERLAP		MEASUREMENT	
	%	C	%	C	%	C	%	C	%	C	%	C	%	C
C1	62	0.24	3	0.93	0	1	100	1	59	0.17	0	1	48	0.03
C2	97	0.93	3	0.93	90	0.8	0	1	57	0.13	0	1	47	0.07
C3	4	0.92	92	0.85	0	1	46	0.08	54	0.08	0	1	42	0.15
C4	100	1	0	1	7	0.86	0	1	46	0.07	0	1	39	0.21
C5	82	0.64	0	1	96	0.93	0	1	0	1	96	0.93	14	0.71
Set 1		**0.75**		**0.94**		**0.92**		**0.82**		**0.29**		**0.99**		**0.23**
C6	7	0.86	34	0.31	14	0.72	14	0.72	0	1	93	0.86	0	1
C7	0	1	90	0.79	10	0.79	14	0.72	0	1	93	0.86	7	0.86
C8	0	1	93	0.87	0	1	43	0.13	70	0.4	0	1	33	0.33
C9	80	0.6	3	0.93	0	1	100	1	60	0.2	0	1	43	0.13
C10	97	0.93	3	0.93	93	0.87	0	1	57	0.13	0	1	43	0.13
Set 2		**0.88**		**0.77**		**0.88**		**0.71**		**0.55**		**0.94**		**0.49**
C11	0	1	76	0.52	24	0.52	31	0.38	24	0.52	0	1	21	0.59
C12	0	1	83	0.67	10	0.8	47	0.07	3	0.93	17	0.67	3	0.93
C13	4	0.93	63	0.26	0	1	100	1	52	0.04	0	1	41	0.19
C14	57	0.13	10	0.8	0	1	100	1	43	0.13	0	1	43	0.13
C15	0	1	77	0.54	4	0.92	92	0.85	35	0.31	0	1	46	0.08
Set 3		**0.81**		**0.56**		**0.85**		**0.66**		**0.39**		**0.93**		**0.38**
C16	35	0.3	0	1	4	0.91	4	0.91	0	1	96	0.91	0	1
C17	85	0.7	0	1	81	0.63	0	1	0	1	100	1	15	0.7
C18	0	1	56	0.11	0	1	93	0.85	0	1	96	0.93	11	0.78
C19	3	0.93	7	0.87	93	0.87	0	1	0	1	93	0.87	17	0.67
C20	0	1	4	0.92	0	1	84	0.68	0	1	96	0.92	16	0.68
Set 4		**0.79**		**0.78**		**0.88**		**0.89**		**1**		**0.93**		**0.77**
All		**0.81**		**0.76**		**0.88**		**0.77**		**0.56**		**0.95**		**0.47**

Overall, we can conclude that the subjects are quite consistent with each other in their use of spatial relations, especially when the fact that some people may use more relations than others is taken into account.

5 Conclusions

In this work we have presented a method of capturing natural language descriptions of relative positions of image objects that reduces bias in the descriptions. The fruitful results of spatial information extraction and analysis provide a general idea of the most common terms and relations used in natural language descriptions of spatial relations between objects in images. In addition to determining these elements of a ‘prototypical’ perception, we feel that the results provide a solid foundation for further study of inter-subject variations. Although the configuration sets assigned to the subjects provided a wide variety of scenarios for the study of common elements of perception, having each subject describe the same 60 configurations (3 sets of 20) would provide further information for the study of inter-subject variations. The design of the experiment described in this paper is not flawless and some of the choices we made may be considered questionable. We present the following limitations and considerations:

The use of a single reference object for all relations provided in a description is a clear limitation. In descriptions that provide relations with mixed reference objects, where the RA was instructed to select the black object as the reference object by default, and enter the semantic inverse of relations in which the grey object was used by the subject as the reference, the relations and terms entered by the RAs may not reflect the description exactly. This instruction may have also resulted in disagreement between the two data sets as to which object was the reference; cases where the RAs do not agree on which object is the reference are disregarded in further analysis.

Because the instances of spatial relations are not uniformly distributed, the numbers of positive examples for different spatial relations are disproportionate. Additionally, because non-prelisted relations were omitted from the measurement and merging procedures, information about occurrences of rarely used relations is limited. Furthermore, we did not use all of the information collected in the experiment, and we feel the additional consideration of each new piece of information could assist in achieving a closer approximation of how humans describe images.

Acknowledgments

The authors wish to express their gratitude for support from the Natural Science and Engineering Research Council of Canada (NSERC), grant 262117.

References

Freeman, J. (1975) “The Modeling of Spatial Relations”, *Computer Graphics and Image Processing*, (4),156-171.

Gapp, K. P. (1995) “Angle, Distance, Shape, and their Relationship to Projective Relations”, *Proceedings of the17th Annual Conference of the Cognitive Science Society*, San Diego, CA, 112-117.

Hall, E. T. (1966) *The Hidden Dimension*, New York: Doubleday.

Halpern, D. F. (1986) *Sex Differences in Cognitive Abilities*, Hillsdale, NJ: Lawrence Erlbaum Associates Press.

Herskovits, A. (1986) *Language and Spatial Cognition: A Interdisciplinary Study of the Prepositions in English*. Cambridge, England: Cambridge University Press.

Landau, B. (1996) "Multiple Geometric Representations of Objects in Language and Language Learners", in P. Bloom, M. Peterson, L. Nadel, and M. Garrett (eds), *Language and Space*, Cambridge: MIT Press, 317-363.

Linn, M. C. and Petersen, A. C (1985) "Emergence and Characterization of Gender Differences in Spatial Ability: A Meta-Analysis", *Child Development*, 56(6), 1479-1498.

Mark, D. M., Comas, D., Egenhofer, M. J., Freundschuh, S. M., Gould, M. D. and Nunes, J (1995) "Evaluating and Refining Computational Models of Spatial Relations Through Cross-Linguistic Human-Subjects Testing", in Frank, A. U. and Kuhn, W. (eds), *Spatial Information Theory: A Theoretical Basis for GIS*, number 998, Springer-Verlag, Lecture Notes in Computer Sciences: Berlin, 553-568.

Mark, D. M. and Egenhofer, M. J. (1994) "Modeling Spatial Relations Between Lines and Regions: Combining Formal Mathematical Models and Human Subjects Testing", *Cartography and Geographic Information Systems*, 21(4),195-212.

Montello, D. R. (1995) "How Significant Are Cultural Differences in Spatial Cognition?", in Frank, A. U. and Kuhn, W. (eds), *Spatial Information Theory: A Theoretical Basis for GIS*, Springer-Verlag, Lecture Notes in Computer Sciences: Berlin, 485-500.

Regier, T. P. (1992) *The Acquisition of Lexical Semantics for Spatial Terms: A Connectionist Model of Perceptual Categorization"*, PhD thesis, University of California at Berkeley, USA.

Robinson, V. B. (1990) "Interactive Machine Acquisition of a Fuzzy Spatial Relation", *Computers and Geosciences*, 16(6), 857-872.

Talmy, L. (1983) "How Language Structures Space", in Pick, H. and Acredolo, L. (eds), *Spatial Orientation: Theory, Research, and Application*, New York: Plenum Press, 225-282.

Wang, X. and Keller, J. M. (1999) "Fuzzy Surroundedness", *Fuzzy Sets and Systems*, 101(1), 5-20.

Wang, X. and Keller, J. M. (1997) "Human-based Spatial Relationship Generalization through Neural Fuzzy Approaches", *Proceedings of the Sixth IEEE International Congress on Fuzzy Systems*, Barcelona, Spain, 1173-1178.

Whorf, B. L. (1940) "Science and Linguistics", *Technological Review*, 42(6), 229-231, 247-248.

Worboys, M. F. (2001) "Nearness Relations in Environmental Space", *International Journal of Geographical Information Science*, 15(7), 633-651.

Zhan, F. B (2002) "A Fuzzy Set Model of Approximate Linguistic Terms in Descriptions of Binary Topological Relations Between Simple Regions", in Matsakis, P. and Sztandera, L.M. (eds), *Applying soft computing in defining spatial relations*, Physica-Verlag, Heidelberg, Germany, 179-202.

Perceptual Sketch Interpretation

Markus Wuersch[1], Max J. Egenhofer[2]

[1] uLocate Communications Inc., 60 Canal St., Boston, MA 02114, USA
e-mail: markus@wuersch.net

[2] National Center for Geographic Information and Analysis, Department of Spatial Information Science and Engineering, Department of Computer Science, University of Maine, Orono, ME 04469-5711, USA
e-mail: max@spatial.maine.edu

Abstract

An automated extraction of regions from sketches can be of great value for multi-modal user interfaces and for interpreting spatial data. This paper develops the *Perceptual Sketch Interpretation* algorithm, which employs the theory of topological relations from spatial reasoning as well as good continuity from gestalt theory in order to model people's perception. The Perceptual Sketch Interpretation algorithm extracts regions iteratively, removing one region at each a time, thus making the remaining sketch simpler and easier to interpret. The evaluation of the algorithm shows that the use of gestalt theory empowers the algorithm to correctly identify regions and saves processing time over other approaches.

1 Introduction

Spatial data are being collected constantly and in large amounts. Interpreting these data poses a shear never-ending task of gaining information from raw data. This task, when dealing with spatial data, relies heavily on feature extraction. Once information can be quickly extracted from spatial data, spatial analysis can be performed based on the resulting information. The assistance of spatial analysis and its results greatly supports society in many challenging tasks and endeavors, such as emergency management, resource management, economic impact studies, and health risk assessment. Automated feature extraction is, therefore, of great importance when dealing with spatial data.

This paper defines a perceptual feature extraction algorithm to successfully identify regions in a sketch, a particularly challenging task for interacting with and using visual information. The goal is to obtain from a visual presentation (i.e., a sketch) exactly what people perceive in such a sketch. This process is called *perceptual sketch interpretation*. The success of feature extraction methods depends largely on the scope of the geometric objects that may be handled (Bennamoun and Mamic 2002). The scope of the PSI algorithm is limited to simple regions in a sketch. Sketches that describe a highly patterned texture (e.g., checkerboard) are outside the scope of this work, because reliably identifying regions in such cases is impossible without additional knowledge.

Automatic extraction of features from a sketch that was originally drawn on paper has to address analog-to-digital conversion in order to execute the feature extraction algorithm. Converting a paper sketch into a digital environment is possible through scanning, edge detection, and vectorization, for which plenty commercial tools are available; therefore, this work only addresses the task of extracting features from vectorized representations. It is also assumed that during the analog-to-digital conversion, the resulting vector representation of a sketch is topologically cleaned (i.e., removing overshoots, undershoots, and slivers).

There is a discrepancy between the elements contained in a sketch and the elements that people perceive. In a sketch, only lines are explicitly present, while regions are perceived by grouping together lines that form closed loops. People are typically very good at perceiving such sequences of lines as regions. What seems to be such a simple task for people, however, has proven to be complex to be formalized so that a machine could carry out that task automatically and reliably. The challenge of feature extraction lies in recovering features undamaged and free of breaks and in successfully grouping them according to the object to which they belong (Bennamoun and Mamic 2002). This paper describes a perceptually supported algorithm for extracting regions from a sketch without prior knowledge about drawing sequences and without interactive human-computer interaction. The sketch in Figure 1 will be used throughout this paper as a running example to illustrate the steps of the region extraction algorithm.

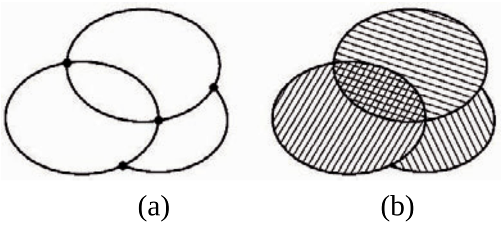


Fig. 1. A sample sketch: (a) the original sketch-nodes highlight the intersections of drawn lines and (b) the identified regions.

Solving feature extraction from sketches will be of great value for research in several domains that deal with visual data, such as computer vision and feature extraction from satellite images or aerial photographs, as well as multi-modal user interfaces, such as spatial-query-by-sketch (Egenhofer 1996), if static sketches *in lieu* of real-time sketches are used as queries.

The remainder of this paper reviews related work in Section 2. Underlying principles from gestalt theory and spatial reasoning are summarized in Section 3. Section 4 introduces the perceptual sketch interpretation algorithm, followed by the description of a prototype system (Section 5) and an evaluation with 24 sketches, and their intended meanings, collected from human subjects (Section 6). Section 7 draws conclusions and suggests future work.

2 Related work

Feature extraction involves image processing, computer vision, and machine intelligence. The most relevant work is briefly reviewed in this section.

Saund (2003) uses a maximal turning path and smooth continuation between lines to identify closed or nearly closed regions. The identified figures are either accepted or rejected based on a measure for a good gestalt. This perceptually closed path finding algorithm, however, requires prior domain knowledge for a successful interpretation.

PerSketch is a perceptually supported sketch editor (Saund and Moran 1995), which offers users suggestions when editing an object in a sketch. In doing so, PerSketch tries to read the users' mind. The algorithm picks objects based on geometric properties (e.g., closure, parallelism, corners, and T-junctions). Research on building these rules can be found in the computer-vision literature (Mohan and Nevatia 1989, Sarkar and Boyer 1993).

CANC2 (Mohan and Nevatia 1992) is a computer vision system that identifies object edges from a vectorized image. The set of vectorized edges of an image is reduced to object edges by applying gestalt laws (e.g., proximity, continuity, symmetry, closure, and familiarity), thus eliminating noise. Identified edges are grouped into non-overlapping object surfaces.

The use of sketching as human-computer interaction mode is explored in *Sketching Spatial Queries* (Blaser 2000), which aims at building a spatial query from a sketch input. The query processor computes similarity values to any other sketch and returns sketches that are similar to the input sketch. Similarity between sketches is computed based on completeness, geometry, topology, metric, and directions of objects and topological relations. Sketching is also used in multi-modal interfaces. Oviatt et al. (1997) use pen input to convey location information. Likewise, Quickset (Cohen 1997) links spoken and pen input.

Based on Wuersch's (2003) use of gestalt principles to extract regions from sketches, Waranusast (2007) forms regions from sketches drawn on PDAs. Such sketches also provide the temporal information about the drawing, offering further heuristics to extract regions successfully.

3 Underlying principles

The region-extraction algorithm developed in this paper is based on the theory of *topological relations* and on *gestalt theory*. Both are briefly explained and complemented with refinements and formal definitions.

3.1 Topological Relations in Sketches

One of the main objectives that spatial reasoning can serve is to change a spatial representation into a different format (translation and interpretation) (Vieu 1997). In the case of interpreting a sketch, this means to identify spatial information of a sketch's elements and to use this information to form new objects. Topology is a most critical part in identifying significant spatial information (Egenhofer and Mark 1995, Kuipers 1979). The 9-intersection (Egenhofer and Herring 1991) provides a framework for identifying formally binary topological relations. It distinguishes eight topological relations between two 2-discs and 33 topological relations between two lines in R^2. This paper considers topological relations between two regions, between two patches (regions in a partitioned space), and between two simple lines.

Regions are homeomorphic to 2-discs and, therefore, two regions in a sketch can have any of the eight possible topological relations. When identifying regions, however, it is impossible to distinguish two regions with the topological relation *equal*.

Partitions are defined as subdivisions of space that consist of cells in the most general case, where any two distinct cells do not have a common interior (Egenhofer and Herring 1991). Patches are not true partitions, because each hole is treated as a separate patch, inside or contained by another patch. Topological relations between two patches (Figure 2) are limited to disjoint, meet, covers, coveredBy, inside, and contains, as well as the dimensional refinements of meet, covers, and coveredBy, referring to the dimension of the shared boundaries of two patches, that is, 0-meet, 1-meet, 0-covers, and 0-coveredBy (Egenhofer 1993). Any two regions with the relation 1-covers or 1-coveredBy form two patches with the relation meet. Two overlapping regions form three or more patches with the topological relation meet between each pair of adjacent patches. Alike equal relations between two regions, this relation is not detectable between two patches.

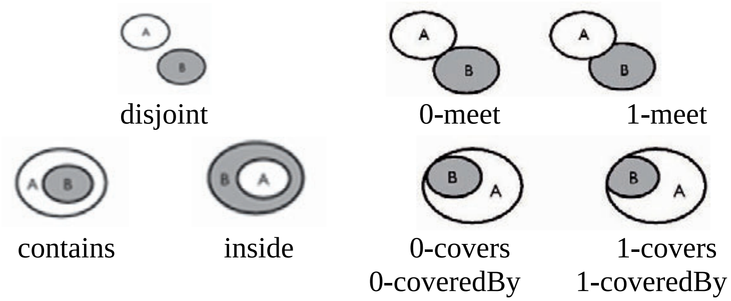


Fig. 2. Possible topological relations between two patches in a sketch.

The extraction of regions from sketches requires a topologically clean sketch (i.e., no crossing lines, overshoots, undershoots, or slivers), reducing the set of possible binary topological relations between two lines from thirty-three (Egenhofer 1993) to three (Figure 3). The distinction of meet-once (Figure 3b) and meet-twice (Figure 3c) arises from the number of non-empty intersections between the boundaries of the lines.

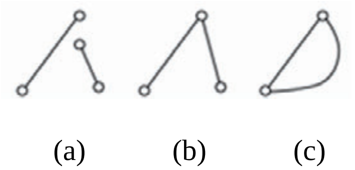


Fig. 3. Possible topological relations between two lines in a sketch: (a) disjoint, (b) meet-once, and (c) meet-twice.

The node degree—that is the number of incoming and outgoing lines at each end of a line —yields a further refinement of the meet relation, which is expressed as m1, m2, m3, m4, etc. (Figure 4).

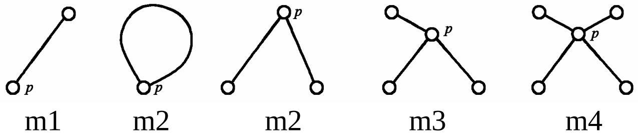
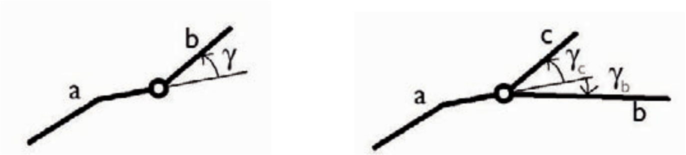


Fig. 4. Intersection types of lines with metric information about the number of incoming and outgoing lines at point *p*.

3.2 Continuity and Good Gestalt

The law of good continuity and the notion of good gestalt from gestalt theory (Koffka 1935, Wertheimer 1923) are of great importance for grouping lines in a sketch, as these gestalt properties often describe people’s perception. Gestalt theory, however, only provides a descriptive theory, but not specific computational processes (Zhu 1999). These theories are briefly explained and complemented with a formal definition.

The law of *good continuity* states that two lines are more likely to be grouped together if one line is perceived as the continuation of the other. In this paper, continuity is expressed by the angle γ formed by two lines, *a* and *b*, that meet (Figure 5a). This angle is then compared to a threshold resulting in either continuity or discontinuity. In cases where more than two lines meet, we rely on the symmetric property of continuity to find the best continuity. In doing so, the continuity angle γ is examined from both directions, that is, from *a* with *b* as the continuing line and from *b,* with *a* as the continuing line (Figure 5b).



(a) (b)

Fig. 5. Continuity for (a) two lines and (b) three lines.

In gestalt theory, the law of *pragnanz* defines that if a perceptual field is disorganized when an organism first experiences it, the organism imposes order on the field in a predictable way. This predictable way is in the direction of a good gestalt, which refers to the simplest, most stable figure possible (Zabrodsky and Algom 1994, Zhu 1999). When describing a good gestalt people use such properties as continuity, regularity, and symmetry. Here we use the continuity property to evaluate a good gestalt of a region. For a qualitative gestalt value, each absolute continuity angle is compared with a continuity threshold. If the angle is lower than the threshold, it contributes to the overall gestalt value with a plus, otherwise with a minus. The sum of all pluses and minuses describes the gestalt value of a region.

4 The perceptual sketch interpretation algorithm

The Perceptual Sketch Interpretation (PSI) algorithm cycles through the following three steps: (1) identifying patches, (2) identifying regions, and (3) extracting and removing the region with the best gestalt. These steps are repeated until all the regions are identified, that is, when no patches are left in the sketch. By iteratively removing any identified region, the remaining sketch becomes less and less complex to interpret.

The PSI algorithm makes three assumptions derived from gestalt theory, which are vital for the result returned by the algorithm:

- *Assumption 1:* Good continuity is a major factor in people's perception to organize visual input into meaningful objects.
- *Assumption 2:* By using the notion of good continuity to identify regions in a sketch, the set of identified regions contains at least one region that corresponds to people's mental model of the same sketch.
- *Assumption 3:* From the set of identified regions, the region with the best gestalt corresponds to a region of people's mental model of the same sketch.

4.1 Identifying Patches

Geographic information must be embedded in a reference system for time, space, and attribute (Chrisman 2001). Feature extraction from sketches, however, can only make use of information about space. Based on information about space, a tracking algorithm that traces along lines and continues consistently in the same direction when reaching an intersection point (e.g., always turn left, or always turn right) identifies boundaries of patches. Patches in a sketch can be used as building blocks for any region. Such a region is built as the union of two or more patches (e.g., two overlapping regions are interpreted as three patches) or a patch is itself a region.

4.2 Identifying Regions

To identify regions in a sketch means to group and union the patches into the region they form. A first step in identifying regions is to extract regions that are formed by only one patch. Such cases correspond to scenarios with a topological relation other than 1-meet (i.e., disjoint patches and patches that do not share any boundary segment with any other patch).

The law of good continuity allows the algorithm to identify regions in a sketch based on Assumption 1: two patches *A* and *B* are likely to form a new region if a segment of patch *B*’s boundary appears as the continuation of a segment of patch *A*’s boundary. Regions are identified by finding two lines that form a good continuation, starting at any segment of any patch’s boundary, here called the *starting line*. The two patches containing the two lines that form a good continuation are combined to build a new region. This process is repeated until a closed boundary is found or no further continuous boundary lines can be found. At this point, the patches used so far are combined to form a region. By repeating this task for any line in the sketch, a set of regions is created that are candidates to be extracted. This set is generally much smaller than the set of all possible regions in a sketch and does not necessarily contain all the regions to be extracted.

An identified region has to satisfy a set of conditions to be a valid region. First, the interior of the region has to be connected, which leads to the conclusion that only patches that share a boundary segment (i.e., the patches have a 1-meet topological relation) can be identified to form new regions (Figure 6a and 6b). Second, the region that is formed as the union of two patches has to contain the two lines that formed the good continuity. In order to satisfy this constraint, both lines cannot be the shared boundary between the two patches, as the shared boundary is not contained in the union of the two patches (Figure 6c and 6d). Third, the resulting new region must be simple, that is, it has no holes, separations, or spikes.

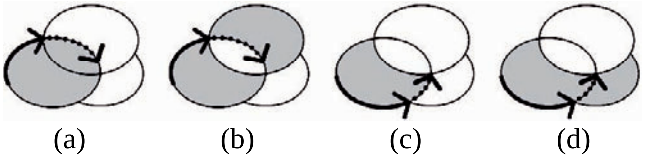


Fig. 6. An example sketch with a starting line (solid arrow) and a continuing line (dotted arrow); (a, c) conditions are satisfied, (b, d) not satisfied.

At the intersection of three lines (i.e., an m3-intersection), there are two possibilities on how the PSI algorithm should proceed if no continuing lines are found. One could argue that the next line in the current patch should serve as the continuing line, because such intersections occur likely where two patches meet (Figure 7). Alternatively, the PSI algorithm simply stops and proceeds to the next patch. The latter approach follows the idea of finding a continuous boundary and, therefore, this approach is chosen for the evaluation of the PSI algorithm.

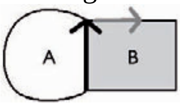


Fig. 7. Continuing at m3-intersection when no continuous boundary is found as an alternative to stopping.

4.3 Extracting Regions with Best Gestalt

Removing at each iteration of the algorithm only the region with the best gestalt value can lead to a more accurate identification of any region left in the sketch. Because the patches are newly built after each time a region is removed from the sketch, the number of patches left in the sketch decreases. Iteratively removing an identified region from the sketch is crucial for a successful interpretation of all the regions in a sketch (Figure 8).

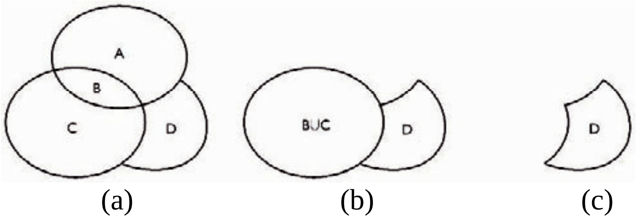


Fig. 8. A sample sketch: (a) the original sketch, (b) after one iteration, and (c) after two iterations.

A region is removed from a sketch by first removing its boundary. In some cases, however, only some parts of the regions boundary can be removed, because its remaining parts are still used for building other patches. In order to outline a rationale on deciding what boundary segments can be safely removed, the segments are classified into line types. Each line type is described by the number of patches that the segment is part of and by the intersection type of each end of the segment (Table 1).

Table 1. Classification of line types

Classification	Intersection Type		Patches
	End 1	End 2	
A	2	2	1
B	3	3	1
C	3	3	2
D	3	4+	1
E	3	4+	2
F	4+	4+	1
G	4+	4+	2

For any line type, except for types A and C, two representations can be found, one where the specific line can be removed and another one where the line cannot be removed, making it necessary to define a rationale whether or not to remove the line. Further analysis shows that lines of type A should not be present in a sketch at this point of the PSI algorithm. Since these lines form a closed loop that was identified as a region, the lines were already removed from a sketch. Lines of type C are the common boundary segments of two patches. Removing such a line when removing the boundary of one of the patches always results in a semi-open set in the sketch and, therefore, lines of type C are kept in the sketch in any case.

For the remaining line types B, D, E, F, and G it is uncertain whether or not to remove the specific line. The difference between cases where the line can be removed and cases where the line cannot be removed is in the number of regions that the line is part of. When a line cannot be removed, it is because that line is part of one or more regions in the sketch, independent of how many patches the line is part of. This information is not available before the extraction algorithm has completed and, therefore, a different approach is chosen. First, all segments of a region's boundary are removed from the sketch, except for the segments of type C. Second, with the remaining lines in the sketch, new patches are built and checked if there are any semi-open sets (Figure 9). In that case, one or more lines that close the semi-open set have to be brought back into the sketch.

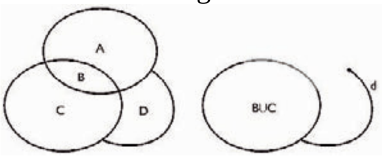


Fig. 9. After removing the region A∪B an open line d is left in the sketch.

During this process, more than one line, called the *closing line,* can be brought back into a sketch. Any such line must be part of the removed region's boundary, it must meet one or more open lines at their open end: it must not be an open line itself, it must be contained in at least one patch, and bringing back the closing line should not introduce any complex object to the sketch. If there is more than one line that meets these constraints, the line that reintroduces the least number of elements of the removed region is chosen. For example, if several lines fulfill these conditions, any line that connects more than one open line is preferred over lines that only connect to one line. In another example, a closing line introduces back into the sketch a part of the removed region's boundary, whereas another closing line introduces back a part of the removed region's boundary and a part of its interior. In this case, the first closing line is chosen, because it does not bring back any part of the region's interior. Finding the closing lines of a sketch after removing a region is an iterative process until no open lines are left in the sketch.

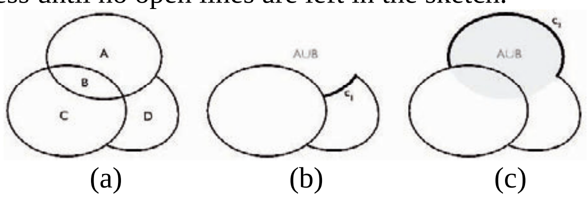


Fig. 10. The closing line: (a) a sketch with patches A-D; (b) and (c) the same sketch with the region A∪B removed. In case of (b) the closing line c_1 only brings back a part of the boundary of (A∪B) whereas in (c) the closing line c_2 also brings back a part of the interior of A∪B into the sketch.

The PSI algorithm uses a minimum and a maximum continuity threshold for identifying continuous lines. At first, the minimum threshold is set low (e.g., 10 degrees) in an attempt to identify regions with a high gestalt value. Only if no regions are identified the continuity threshold is increased until the maximum threshold is reached or until there are no patches left in the sketch.

In cases where the PSI algorithm finishes with patches left in the sketch, these remaining patches are added to the set of extracted regions in order to complete the spatial scene.

It is possible that a patch is lost after a region is removed from the sketch. Whereas the extracted regions will not cover the same space as the original sketch, the extracted regions might match with people's mental model. In this case, an option is given whether or not to fill gaps at the end of the region extraction process. In order to illustrate such a case, two partitions have been added to the sample sketch (Figure 11).

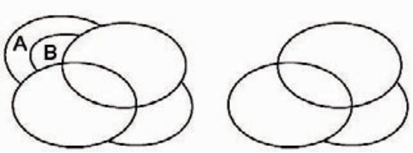


Fig. 11. Patch B is lost after removing region A.

5 Prototype

The PSI algorithm (Figure 12) was implemented in a prototype application that serves as a test bed for the model evaluation. It extracts features from a digital sketch. Any pre-processing of such a line drawing (i.e., scanning, raster-to-vector conversion, and cleaning topology) was completed with commercial hardware and software.

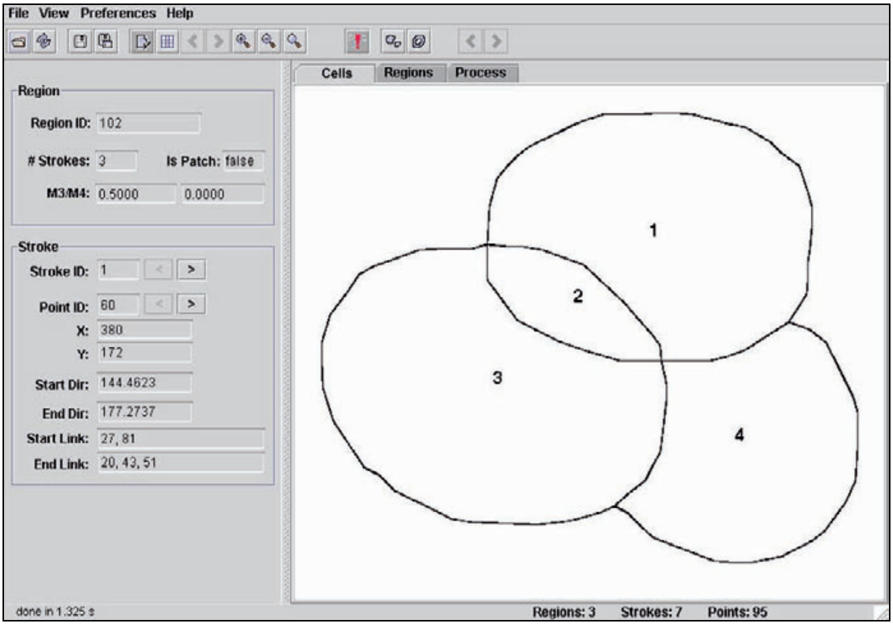
The prototype uses a map metaphor for displaying the sketch and allows a user to interact through a WIMP (windows, icons, menus, pointers) interface. A preference pane lets users adjust any setting used in the feature extraction process, such as continuity thresholds. Supported file formats are a text file containing a list of points grouped by line numbers and ESRI's interchange format (e00). Upon opening a sketch, three different views of the sketch are displayed: the original sketch with patches, the processed sketch containing regions (Figure 13), and a process view displaying feature extraction process at different stages. These visualizations help with analyzing possible errors in case the algorithm commits any misinterpretations. The interpreted sketch can be saved in the Spatial-Query-by-Sketch format (Blaser 2000), enabling a subsequent spatial query that can be executed on a set of other sketches.

6. Evaluation

The PSI algorithm was evaluated for correctness and compared to an alternative approach where, instead of using continuity, every possible region in a sketch is analyzed to identify regions. The result of the evaluation shows a significant advantage in efficiency using continuity to identify regions over that alternative approach. In addition, the PSI algorithm correctly interpreted 75% of the analyzed sketches. The comparison to this approach shows the advantage in processing load when the notion of good continuation is used.

```
  function PSI (sketch): sketch
  newSketch := empty sketch;
  newRegions := empty list of regions;
  continuityThreshold := minThreshold;
  loop
    remove from sketch patches that are not 1-meet to any other patch and
    add them to newSketch;
    find set of all possible regions in sketch;
    or find set of possible regions using continuity:
        for each line in sketch
            for each patch containing line
                find region using continuity at start of line
                and add it to newRegions;
                find region using continuity at end of line
                and add it to newRegions;
            end for;
        end for;
    end or;
    if newRegions is empty: increment continuityThreshold;
    else
        remove region from newRegions with best gestalt
            and add removed region to newSketch;
        build patches with remaining lines in sketch;
    end if;
  loop until patches is empty
        or continuityThreshold > maxThreshold;
  add unused patches to newSketch;
return newSketch;
```

Fig. 12. Pseudo code of the PSI algorithm.



Region ID: 102
Strokes: 3
Is Patch: false
M3/M4: 0.5000 0.0000
Stroke ID: 1
Point ID: 60
X: 380
Y: 172
Start Dir: 144.4623
End Dir: 177.2737
Start Link: 27, 81
End Link: 20, 43, 51
done in 1.325 s
Regions: 3 Strokes: 7 Points: 95

Fig. 13. The application window showing the processed sketch in the region tab. A region is highlighted and the corresponding attributes of that region are displayed in the sketch properties panel on the left.

6.1 Evaluation Design

In order to objectively evaluate the PSI algorithm, a set of sketches were obtained from people who were not involved in the design of the PSI algorithm. For this purpose, a web-based survey was conducted, giving the participating subjects the opportunity to draw and submit their sketch through a Web browser interface. Participants were also asked to submit their interpretation of the sketch. The collected information, therefore, contained the topological information, labels of each patch, and a description of the composition of each region (e.g., which patches are contained by a region). In total, 36 sketches were collected of which 24 did not contain any complex regions or tessellations and were selected for the evaluation.

6.2 Correctness

For each collected sketch, regions were extracted manually according to the description obtained from the online survey. The resulting spatial scene was termed *ground truth* and used to evaluate the correctness of the PSI algorithm's results. This ground truth was compared to a spatial scene identified by the PSI algorithm. The prototype for Spatial-Query-by-Sketch (Blaser 2000) was used to determine similarity values between the two sketches. The ground truth was used as a query input, operating on the interpreted sketch created by the PSI prototype. Spatial-Query-by-Sketch returns similarity values (0% to 100%), which quantify the accuracy of the region extraction process: a similarity value of 100% shows correct interpretation of the sketch, less than 100% indicates a deviation from a correct interpretation (100% is a theoretical value and because of rounding errors the actual received similarity value for identical sketches were 99.9%). From the 24 analyzed sketches, 18 (i.e., 75%) were interpreted correctly (Table 2).

6.3 Advantage of the Continuity Approach

The PSI algorithm uses the notion of good continuation to identify what patches should be combined to form regions with the best possible continuous boundary. An alternative to this approach would analyze all possible regions in a sketch.

Test sketches were processed again using all possible regions in a sketch. First, the results of this process are compared to the ground truth and the results from the extraction process of the continuity-based approach. This approach using all regions produced 15 correct interpretations (i.e., 62.5%), a less accurate result than the continuity-based approach—the PSI algorithm performed on these samples 12.5% better in absolute numbers, and 20% better with respect to the success rate of all-regions approach (Table 2).

Table 2. Correctness of the PSI algorithm

	Using continuity	Using all regions
Correctly interpreted sketches	75%	62.5%
Average similarity	93.9%	92.4%
Smallest similarity	37.6%	46.4%
Largest similarity	99.9%	99.9%

The analysis of the processing times of both approaches clearly shows the advantage of the continuity-based approach as it executed on average 119 times faster than when analyzing all possible regions in a sketch (Table 3). This difference in processing time is due to the often very large numbers of possible regions that can be extracted from a sketch.

Table 3. The processing load of the PSI prototype (processing was done on a computer running Windows XP, 2.99 GHz processor, and with 1 GB of RAM)

	Using continuity	Using all regions
Average processing time [sec]	0.2	26.2
Average number of regions analyzed	11	2,056

6.4 Shortcomings

The six sketches that were incorrectly interpreted using the continuity approach were analyzed in more detail and two reasons for an incorrect interpretation were found: either an incorrect region was identified as having the best gestalt, or a region was removed incorrectly. In two cases, the rationale for removing an identified region's boundary from the sketch returned incorrect results, while in five cases continuity was not the major factor used by people's perception to order visual input so that regions were identified that should not have been extracted. Regions with a regular shape (e.g., squares and rectangles) were not identified or were not classified as having a good gestalt. The set of identified regions, however, contained at least one region to be extracted, thus Assumption 2 (Section 4) holds.

7 Conclusions and future work

We developed an algorithm to extract features from sketches. This algorithm makes use of the law of good continuity and the notion of a good gestalt to identify regions and to rank these regions by their gestalt value. A prototype implementation was used to evaluate the PSI model from which we can draw conclusions and suggest future work.

7.1 Results

The results of the PSI algorithm's evaluation lead to three major conclusions:

- By using continuity to identify a set of regions from patches, the resulting set contains at least one region that corresponds to a region in people's mental model. This conclusion confirms Assumption 2 (Section 4), supported by the results of the model assessment. Because this assumption clearly holds for the approach of analyzing all possible regions and because the results of the assessment have shown evidence that using continuity produces equal or better results, it can be concluded that the assumption also holds for the continuity approach. This conclusion is also supported by the analysis of the incorrect interpreted sketches, where other reasons were identified as the cause of incorrect results.

- Good continuity is, amongst other gestalt laws, one of the major factors used by people’s perception to order visual input into meaningful objects. This conclusion is drawn because the PSI algorithm has correctly interpreted 60% to 75% of the test sketches. Sketches that were misinterpreted, however, ask for additional reasoning other than the notion of good continuity. This conclusion refers to Assumption 1, but also applies to Assumption 3 (Section 4), which states that the region with the best gestalt has a corresponding region in people’s mental model. Clearly, this assumption depends on the definition of a good gestalt, which in turn depends on the gestalt laws used.
- The continuity approach to identify regions is preferred over analyzing all possible regions. The analysis of the number of correctly interpreted sketches and the analysis of the processing times strongly support this conclusion.

7.2 Future Work

The model assessment indicates possible future research topics as well as refinements and extension to the current model. Further analysis using the PSI prototype could be performed using a variety of algorithm preferences. Such analysis could show correlations between scene characteristics and distinct preferences of the algorithm. Where such correlations exist, the PSI algorithm could be tailored towards different types of sketches. Additional analysis of the algorithm’s settings could also reveal which settings are the most relevant and would allow us to aim future work at the most important parts of the algorithm.

The PSI algorithm was tested on a set of sample sketches that were hand drawn. Further valuable results could be gained by evaluating the PSI algorithm on data other than sketches. Such data could be vectorized aerial photographs, satellite imagery, or any other data type that can be transformed into a vector representation.

The current data model requires a completely clean topology of the sketched lines in order to apply qualitative reasoning as it is described in this work. If a scene is to be analyzed on a more detailed level, however, metric aspects as well as direction information become relevant. Incorporating such refinements for the topological relations would possibly result in more accurate sketch interpretations. It would also allow for a purely automated process, as it would rely less on generating a clean topology of the scanned sketch.

Drawing errors, such as overshoots, undershoots and slivers, are corrected for by a cleaning function before the actual region extraction process commences. In doing so, some information that could reveal more details about the possible regions in a sketch might be compromised. For example, slivers indicate that a line was drawn twice. In cases where drawing errors occur, they could give better insights on regions in a sketch thus improving the result of the region extraction. In the example of a line drawn twice, the PSI algorithm could use this information to make sure that the line is used in two different regions.

The analysis of the shortcomings of the PSI algorithm has shown that the rationale of removing an identified region's boundary, outlined in section 0, could not be relied on at all times. The result of the PSI algorithm could be improved by refining this rationale.

The PSI algorithm uses the notion of good continuity to identify a set of possible regions from a sketch and to describe a region's gestalt. The laws of organization also define other principles (e.g., regularity, symmetry, proximity, co-linearity, co- circularity, parallelism, closure, similarity, and simplicity) that can possibly be used instead or in addition to continuity. Research on using laws of organization in computer vision can be found in Lowe (1990), Mohan and Nevatia (1992), Park and Gero (1999), Saund (2003), Saund and Moran (1995), Zabrodsky and Algom (1994), and Zhu (1999). The analysis of the shortcomings of the PSI algorithm showed that such an extension of the algorithm could lead to a better performance of the PSI algorithm. For example, because regular shapes were not identified or were not assigned a good gestalt, regularity could be of great value for this algorithm.

While these recommendations for future work show room for improvement of the PSI algorithm, the algorithm showed convincing results supporting the perceptual approach of interpreting sketches.

Acknowledgments

This work was partially supported by the National Geospatial-Intelligence Agency under grant numbers NMA201-01-1-2003.

References

Bennamoun M, Mamic G (2002). Object recognition. Springer-Verlag, London

Blake A, Isard M (1998). Active contours. Springer-Verlag, London

Blaser A (2000). Sketching spatial queries. Ph.D. Thesis, University of Maine, Orono, ME

Blaser A, Egenhofer M (2000). A visual tool for querying geographic databases. In Di Gesù V, Levialdi S, Tarantini L (eds), AVI 2000—Advanced visual databases, Salerno, Italy, pp 211-216

Chrisman N (2001) Exploring geographic information systems. John Wiley, New York

Cohen P, Johnston M, McGee D, Oviatt S, Pittman J, Smith I, Chen L, Clow J (1997). Quickset: multimodal interaction for distributed applications. Proceedings of the fifth ACM international multimedia conference, pp 31-40

Egenhofer M (1993). A model for detailed binary topological relationships, *Geomatica*, 47(3&4), 261-273

Egenhofer M (1993). Definitions of line-line relations for geographic databases. *IEEE data engineering bulletin* 16(3), 40-45

Egenhofer M (1996). Spatial-Query-by-Sketch. In Burnett M and Citrin W (eds) VL '96: IEEE symposium on visual languages, Boulder, CO, 60-67

Egenhofer M (1997). Query processing in Spatial-Query-by-Sketch. Journal of visual languages and computing 8(4): 403-424

Egenhofer M, Herring J (1991). Categorizing binary topological relationships between regions, lines, and points in geographic databases. Technical Report, Department of Surveying Engineering, University of Maine, Orono, ME, (http://www.spatial.maine.edu/~max/9intreport.pdf)

Egenhofer M, Mark D (1995). Naive geography, In: Frank A, Kuhn W (eds), COSIT '95, Spatial information theory. Lecture Notes in Computer Science vol 988, pp 1-16

Egenhofer M, Shariff AR (1998). Metric details for natural-language spatial relations, ACM transactions on information systems 16(4): 295-321

Koffka, K (1935). Principles of gestalt psychology, Harcourt, Brace and Company, New York

Kuipers B (1979) Modeling spatial knowledge. Cognitive science 2(2): 129-153

Lowe D (1990) Visual recognition as probabilistic inference from spatial relations, In Blake A, Troscianko T (eds), AI and eye, John Wiley, New York

Mohan R, Nevatia R (1989). Using perceptual organization to extract 3-d structures, IEEE transactions on pattern analysis and machine intelligence 11(11): 1121-1139

Mohan R, Nevatia R (1992) Perceptual organization for scene segmentation and description, IEEE transactions on pattern analysis and machine intelligence 14(6): 616-635

Oviatt S, DeAngeli A, Kuhn K (1997) Integration and synchronization of input modes during multimodal human-computer interaction. Proceedings of the conference on human factors in computing systems (CHI '97), pp 415-422

Park S-H, Gero J (1999) Qualitative representation and reasoning about shapes, In Gero J, Tversky B (eds.), Visual and spatial reasoning in design, Key Centre of Design Computing and Cognition, University of Sydney, Sydney, Australia, pp 55-68

Sarkar S, Boyer K (1993) Integration, inference, and management of spatial information using Bayesian networks: perceptual organization. IEEE transactions on pattern analysis and machine intelligence 15(3): 256-274

Saund E (2003) Finding perceptually closed paths in sketches and drawings. IEEE transactions on pattern analysis and machine intelligence 25(4): 475-491

Saund E, Moran T (1995) Perceptual organization in an interactive sketch editing application. International conference on computer vision (ICCV '95*)*, IEEE Computer Society Press, pp 597-604

Shariff AR, Egenhofer M, Mark D (1998) Natural-language spatial relations between linear and areal objects: the topology and metric of English-language terms, International journal of geographical information science 12(3): 215-246

Vieu L (1997) Spatial representation and reasoning in artificial intelligence, in Stock, O. (ed.) Spatial and Temporal Reasoning, Kluwer, Dordrecht, pp 5-41

Waranusast R (2007) Perceptual-based region extraction from hand drawn sketches. Proceedings of the third IASTED international conference, advances in computer science and technology, Phuket, Thailand, pp 222-227

Wertheimer M (1923) Laws of organization in perceptual forms, In Ellis W (ed.), A source book of gestalt psychology, Routledge & Kegan Paul, London, pp 71-88

Wuersch M (2003) Perceptual sketch interpretation. M.S. thesis, University of Maine

Zabrodsky H, Algom D (1994) Continuous symmetry: a model for human figural perception. Spatial vision, 8(4): 455-467

Zhu S-C (1999) Embedding gestalt laws in Markov random fields—a theory for shape modeling and perceptual organization. IEEE transactions on pattern analysis and machine intelligence 21(11): 1170-1187

The Shape Cognition and Query Supported by Fourier Transform

Tinghua Ai, Yun Shuai and Jingzhong Li

School of Resource and Environment Sciences, Wuhan University,
129 LuoYu Road., Wuhan, P. R. China, 430072
email: tinghuaai@gmail.com, shuaiyunsh@126.com, lilidegm@gmail.com

Abstract

As an important function of GIS, spatial query covers not only the extraction of geometric, topologic or semantic information but also the retrieval of spatial cognition related information. This study presents a shape based spatial query way that is formally described as: *Select {O_i} From DataBase Where O_i.shape LIKE Template At_Degree* <C_i>. To extend the new operation *LIKE* in formal SQL language, the shape representation and measurement have to be built. This study aims at polygon object offering a Fourier transform based method to compute the degree of shape similarity. Through cognition experiment builds the membership function of fuzzy term *LIKE*. The query experiments show the shape based query retrieves the building result consistent with human cognition.

Keywords: Spatial query, Spatial cognition, Shape measurement, Fourier transform

1 Introduction

The traditional SQL language integrates spatial concepts leading to the spatial SQL query, in which the data types have been extended from simple common ones such as integer, float, date, char string etc. to complex spatial data types such as point, line, polygon, poly-line, poly-polygon *etc*.

Operations are also extended to process spatial information including topological, Euclidean, directional and metric data. In spatial query standards, for example OGIS or SQL3, the spatial operations usually fall into three categories (Shekhar and Chawla 2002): (1)the geometric operations which are spatial reference determination, envelope computation, boundaryextraction, simple judgment and others; (2)the topological operations which are the boolean judgment of relationships of equal, disjoint, intersect, touch, cross, within and contains; and (3) the spatial analysis which includes buffer, union computation, intersection extraction, convex-hull generation and others.

However, the spatial query involves other operations not covered in the classification above. Spatial query answers not only the geometric, topologic or semantic information but also some spatial cognition related information. We usually need to find an object to match the mental symbol in our memory by human reaction. The cognitive structures and process are part of the mind, which emerges from a brain and nervous system inside of a body that exists in a social and physical world (Freksa 1991). The spatial shape and spatial pattern recognition and query is such a question that the retrieval depends on the human cognition reaction rather than the properties of existed entity. For example, we want to extract some buildings which are “T” shaped or “U” shaped from a spatial database. Shape allows to predict more facts about an object than other features, e.g. color. Thus, recognizing shape is crucial for object recognition and also plays an important role in spatial query. From the perspective of spatial cognition requirements, the SQL query need to continuously extend operations so that it can compare and extract cognition results.

The shape based spatial query usually behaves as the identification that one object is like another one or two objects are similar in shape structure. The retrieved result is usually uncertain depending on human’s emotion, background knowledge and perception abilities. The shape template and the degree to which extent two objects are similar vary with different persons. Thus spatial shape can only qualitatively compare with fuzzy properties. In this process, the key question is to mathematically describe the shapes and to derive a similarity measurement to compare the shapes under the idea of fuzzy properties.

In multimedia and image processing domain, the shape representation and measure is active generating a lot of methods and algorithms (Hu 1962; Latecki and Lakämper 2002; Bengtsson and Eklundh 1991; Jones and Ware 1998). These methods aim at region, boundary and structure respectively. For the description of global shape properties, there are geometric parameters including size, perimeter, convex perimeter, elongation, roughness, and compactness, etc. For polygon object, the turn function or bend angle function based on contour points can be applied to measure region shape with the invariance to translation, rotation and scale (Latecki and Lakämper 2002). In these fields, the moment based algorithm is an efficient method to represent shape aiming at area boundary, which includes invariant moments, higher order moments and generalized complex moments (Kim Y-S. and Kim W-Y.1997).

In image database field, the shape measure is conducted on the basis of pixel or raster grid through the integration of set of pixels to get a complete shape concept. But in GIS database which mainly stores vector data, the shape measure directly faces independent geometric entities, such as line, polygon, arc, point cluster etc. So the shape measure method in GIS domain is different from that of image database. For pixel data, it is easy to be transformed to the frequency domain representation which usually acts as the basis of shape measure. In image database, the object boundary is usually represented as the chain code whose change along the boundary tracking or central angle can be converted to frequency domain.

Contrast to the image pixel data, this study aims at the GIS vector data investigating the shape representation and studying the application of shape based query for building feature database. The applied algorithm is Fourier transform which has been widely used in image data analysis. But in this study Fourier transform is based on the continuous function of vector data rather than the discrete pixel chain. The query way behaves as the template match which is controlled by Fourier descriptor. The rest of paper is organized as follows. Section 2 examines the characteristics of region shape taking the building feature as example. Section 3 presents the analysis of Fourier transform on vector polygon data and the formula of shape measure. The template based spatial shape query is offered in section 4 with experiment discussion. Section 5 summarizes the characteristics of this method and concludes with the future research.

2 Shape Representation

Shape is probably the most important property that is perceived about objects. It allows to predict more facts about an object than other features, e.g. color (Palmer 1999). In GIS database, shape contains some characteristics of geographic phenomena which can be mined to discover the hidden geographic principles. The famous example in geo-science is that Alfred Wegener built the theory of continental drift which was first driven by continental shape analysis. Wegener in 1912 noticed that the shapes of continents on either side of the Atlantic Ocean seem to fit together, for example, Africa and South America. In human-geography, the shape of ancient building reflects the construction culture characteristics in the corresponding era.

Aiming at different contents, shape representation can be divided into three classes, namely the boundary, the region and the structure objective respectively. The boundary shape of an object describes the complexity of curve pattern, the extension trend at one dimension. Its comparison can be measured by smooth, fitness and other computations. The region shape regards the object as a point set, a connected component and at two dimensions represents the pattern of object distribution and extension. The region shape can be described with vague terms like *elongated*, *round*, *compactness* and so on. The structure shape regards the whole object as a component of different parts and studies the organization pattern. A given object can be mapped into a graph in which nodes correspond to divided pieces and arcs encode spatial relationships by skeleton conversion at reduced dimensions. This study will investigate the region shape in GIS data query.

Compared with other domains such as image processing, computer vision and computer graphics, GIS deals with spatial objects with vector representation at larger scale. The shape representation in GIS has the following properties:

- **Abstraction.** According to Gestalt cognition principles, we look at an object first obtaining the whole complete sense to recognize the region shape, and then looking into the down details to make up the shape. It means the shape representation should firstly be abstracted through the simplification to get a generalized concept. In image processing, we use noise reduction methods to remove minor or deflection details. In GIS filed, the vector data can apply map generalization technology to get the abstracted structure (Brassel and Weibel 1988). Compared with the objective of map representation at smaller scale, the generalization aiming at shape extraction requires to be conducted at a large degree (Ai 2000; Li et al 2004). There is a special map cartogram, also called value-by-area map (Dent 1975) which depicts the attribute of geographic object as the object's area. The abstraction makes the region size completely different from the original area but the shape similar to the original. Another map called schematic map is a linear abstraction of functional networks for subway, railway, shipping lane representation through line simplification to represent the network structure and topological relationships (Avelar and Muller 2000). The main shape of network edge gets maintenance as similar as possible in the simplification process.
- **Symbolization** For region shape description, we usually establish the association between the recognized object and the template pattern in our

memory, which is familiar to us according to our knowledge and experience. This template is used as a symbol to represent the region shape (Rainsford 2002). The shape template could be the text letter the Chinese text, the familiar animal, the goods in daily life and so on. We usually say the territory of China is like a great cock and Italy a boot as shown in figure 1. The symbolization makes the shape representation easy to understand and communicate.

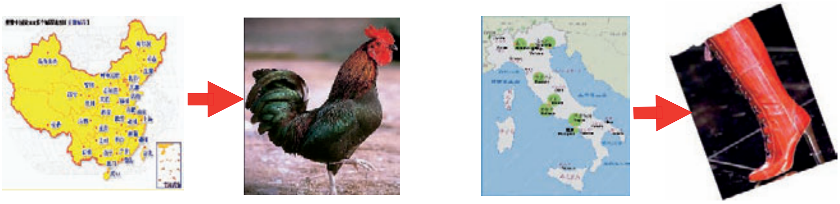


Fig. 1. Examples of symbolization of country's territory in shape representation

- **Indetermination.** The shape representation relies on the cognition of humans with different background knowledge, intending interests and emotions. So for the same region object, different persons may recognize getting different shapes. The indetermination of region shape recognition makes the shape query uncertain in template selection and the decision degree to which extent two objects are similar. The same building may be identified as L, U or V shape for different persons. The shape based spatial query should belong to fuzzy query class. For the measurement of shape similarity, the qualitative method divides the representation into such as *similar very much, moderately similar, commonly similar, slightly similar, etc.* For shape comparison, we can apply fuzzy mathematics to define the membership function for fuzzy term "similar".

Considering the above properties, shape is a very difficult concept to capture mathematically and with consistency. How to use a model to represent the shape sense in our mental world and by a single number to compute the shape measure is not easy. Desired property of a similarity function *C* should be a metric which has the characteristics: self-identity, positivity, symmetry and triangle inequality (Basri et al. 1998) and also the similarity function should be continuous and invariant to geometric transformation. A lot of algorithms have been developed to measure shape similarity aiming at different situations and requirements. Among them the moment based and Fourier transform based methods play an important role under the requirements of shape similarity above. They combine information across an entire object rather than providing information just at a single boundary point, they capture some of the global properties missing from many pure contour-based representations: overall orientation, elongation, etc.

Next sections will discuss the region shape measure by Fourier transform considering the vector data structure rather than chain coding.

3 Fourier transform and shape measure

Fourier transform is a well known data analysis of frequency domain broadly applied in image processing, such as shape representation. Many Fourier transform methods have been reported in the literature, these include using Fourier descriptor for shape analysis and, character recognition (Persoon and Fu 1977), shape classification (Kauppinen et al. 1995) and shape retrieval (Lu and Sajjanhar 1999). In these methods, different shape signatures have been exploited to obtain Fourier descriptor. The basic idea of shape representation by Fourier transform is to describe the shape in terms of its spatial frequency content. The process is firstly representing the boundary of the shape as a periodic function which is expanded in a Fourier descriptor series, and then obtaining a set of coefficients that capture the shape information. A Fourier descriptor begins with tracking the region boundary which can be represented as two data structures, namely the chain code in grid representation and the consecutive points in vector representation. Both ways meet the important condition in shape representation, namely to maintain invariance during the region translation, rotation and scaling. We will apply the vector data structure based method since GIS mainly deals with vector data.

3.1 Fourier Descriptor on vector polygon

Usually we convert a 2D areas or boundaries to 1D function by shape signature to represent shape. Generally there are four shape signatures, namely central distance, complex coordinates (position function), curvature and cumulative angular function (Persoon and Fu 1977). Here we consider position based shape signature by Fourier descriptor method.

The boundary is represented as a set of connection points. Each point *k* is regarded as a complex number pair by treating the x-axis as the real axis and the y-axis as the imaginary axis. Thus a given point which is the function of arc length *s* can be represented as

$$U(s) = x(s) + iy(s) \tag{1}$$

where *s* is the distance of arc path between the given point and the reference original point, for example b_0 as shown in figure 2 and *i* is $\sqrt{-1}$. The representation treats the plane of the region as an Argand diagram reducing a 2D problem to a 1D problem.

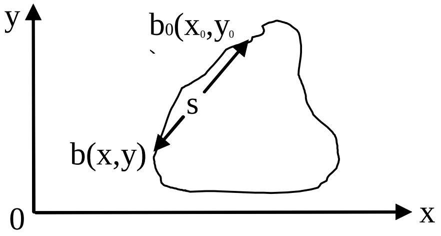


Fig. 2. The boundary representation in complex plane

Let region perimeter be *Z,* then $U(s)$ is a periodic function with

$$U(s+Z)=U(s), 0 \leq s < Z \tag{2}$$

Let $t=\frac{2\pi s}{Z}$, then the equation is modifies as

$$U(t)=x(t)+iy(t), 0 \leq t < 2\pi \tag{3}$$

$U(t)$ is a periodic function with period 2π and its Fourier expansion given by

$$U(t)=\sum_{n=-\infty}^{+\infty} p_n e^{-int}\ ,\ 0 \leq t < 2\pi \tag{4}$$

Where the coefficients of Fourier transform are

$$p_n=\frac{1}{2\pi}\int_0^{2\pi} U(t)\ e^{-int}dt\ ,\ n=0,\pm 1,\pm 2,\ldots\ldots\ldots \tag{5}$$

Suppose the region boundary have *M* points , then the boundary can be regarded as a consecutive accumulation of *M-1* segments. Let s_k be the length of arc from point *k* to reference point around the boundary.

$$s_k=\left\{\begin{array}{l} 0\ ,\ k=0\ ; \\ \sum_{\lambda=0}^{k-1}\sqrt{(x_{\lambda+1}-x_\lambda)^2+(y_{\lambda+1}-y_\lambda)^2}\ ,\ k=1,2,3\ldots M-1 \end{array}\right\}$$

Assume the change point (*x(s),y(s)*) locate between point *k* and point (*k+1*). Let γ be the distance between point($x(s), y(s)$) and point *k*, $\gamma=\sqrt{(x_{(s)}-x_k)^2+(y_{(s)}-y_k)^2}$, *k =1, 2, 3 …., M-1,*then $s=s_k+\gamma$. Let $t=\frac{2\pi s}{Z}$, the segment length between point *k* and point *k+1* be

$l_k = \sqrt{(x_{k+1} - x_k)^2 + (y_{k+1} - y_k)^2}$ and $D_k = (x_{k+1} - x_k) \big/ l_k$,and construct a lineal function across point *k* and point *k+1* and apply the integral with segment length ranging from 0 to l_k. ,then the coefficients P_n are represented as

$$P_n = \frac{1}{Z}\int_0^Z U(s)\, e^{-i\frac{2\pi ns}{Z}}\, ds = \frac{1}{Z}\sum_{k=0}^{M-1}\int_0^{l_k} U(s_k + \gamma)\, e^{-i\frac{2\pi n(s_k+\gamma)}{Z}}\, d\gamma$$

$= \frac{1}{Z}\sum_{k=0}^{M-1}(a_k + i\, b_k)$, where a_k, b_k are expressed respectively as

$$a_k = \int_0^{l_k}[(x_k + \gamma D_k)\cos(-\frac{2\pi n(s_k+\gamma)}{Z}) - (y_k + \gamma D_k)\sin(-\frac{2\pi n(s_k+\gamma)}{Z})]\, d\gamma$$

$$b_k = \int_0^{l_k}[(x_k + \gamma D_k)\sin(-\frac{2\pi n(s_k+\gamma)}{Z}) + (y_k + \gamma D_k)\cos(-\frac{2\pi n(s_k+\gamma)}{Z})]\, d\gamma \qquad (6)$$

Further expand the above formula finally getting the expression of coefficients as follows:

$p_n = A_n + i\, B_n$,whree A_n, B_n are expressed respectively as

$$\left\{\begin{array}{l}
\left.\begin{array}{l}
A_n = \frac{1}{Z}\sum_{k=0}^{M-1}\frac{l_k}{2}(x_{k+1} + x_k) \\
B_n = \frac{1}{Z}\sum_{k=0}^{M-1}\frac{l_k}{2}(y_{k+1} + y_k)
\end{array}\right\} n = 0\,; \\
\left.\begin{array}{l}
A_n = \frac{1}{2\pi n}\sum_{k=0}^{M-1}\left\{\begin{array}{l} x_k\sin\alpha - x_{k+1}\sin\beta + y_k\cos\alpha - y_{k+1}\cos\beta \\ + \frac{Z}{2\pi n\, l_k}[(y_{k+1} - y_k)\times(\sin\alpha - \sin\beta) - (x_{k+1} - x_k)\times(\cos\alpha - \cos\beta)] \end{array}\right\} \\
B_n = -\frac{1}{2\pi n}\sum_{k=0}^{M-1}\left\{\begin{array}{l} x_k\cos\alpha - x_{k+1}\cos\beta - y_k\sin\alpha + y_{k+1}\sin\beta \\ + \frac{Z}{2\pi n\, l_k}[(y_{k+1} - y_k)\times(\cos\alpha - \cos\beta) + (x_{k+1} - x_k)\times(\sin\alpha - \sin\beta)] \end{array}\right\} \\
\text{where: } \alpha = -\frac{2\pi n\, s_k}{Z}, \beta = -\frac{2\pi n\, s_{k+1}}{Z}
\end{array}\right\} n \neq 0
\end{array}\right. \qquad (7)$$

In Fourier descriptor above, we use the accumulation of *M-1* boundary segments to expand the coefficient formula and for each segment apply the integral of a continuous lineal function. Traditionally the boundary is divided into chain codes or partitioned as series of segments with a determinate step, and then expand the formula by discrete integral method. Obviously the previous method has higher accuracy due to the continuous integral.

The magnitude of coefficients P_n has rotation and translation invariance. Define a new coefficient *d(n)* to normalize the coefficients P_n , which is achieved by dividing the magnitude values of all the other descriptors by the magnitude value of the second descriptor.

$$d(n) = \frac{|p_n|}{|p_1|}, n = 1, 2, \ldots., M - 1 \tag{8}$$

Then the descriptors have also the invariance to scale. The series of coefficient *d(n)* is usually called shape parameters since they capture the main shape information.

3.2 Shape measure

According to the characteristics of Fourier transform, the Fourier descriptor approximates the original region at different accuracy and the coefficients capture the shape information with invariance to translation, rotation and scale. The coefficient P_n at different order *n* represents different frequency contents of region. The frequency domain at lower order *n* corresponds to more significant shape component. Since the shape is an abstracted representation, we can use a subset of low order coefficients to represent the overall features of the shape. The very high frequency information describes the small details of the shape. It is not so helpful in shape discrimination, and therefore they can be ignored.

Based on the coefficients *d(n),* the feature vector to index the shape is then

$$f = [d(1), d(2), d(3),\ldots d(N\text{-}1)].$$

Now for two model shapes indexed by Fourier descriptor feature f_i and f_j respectively, since both features are normalized as to translation, rotation and scale, the Euclidean distance between the two feature vectors can be used as the similarity measurement

$$dis = \sqrt{\sum_{k=1}^{N} |d_i(k) - d_j(k)|^2} \ , \tag{9}$$

Where *N* is the truncated number of harmonics needed to index the shape.

Given a region, the truncated number *N* is determined by the degree that Fourier descriptor approximates the original region. Assume the original region is *Co*, the approximated region by Fourier descriptor is *Ca* and the intersection of *Co* and *Ca* is $Co\cap a$. We define a parameter approximation degree which is computed as

$$A\_degree=area(Co\cap a)/area(Co) \qquad (10)$$

The parameter *A_degree* ranges from 0 to 1. The closer the value is to 1, the higher accuracy the approximation is. Given a determinate value for *A_degree*, e.g. 0.85, different shape needs different order expansions of Fourier descriptor to access the accuracy. According to the characteristics of coefficients *Pn* at different order (frequency domain), the approximation needs fewer orders to access a given accuracy, if the region shape intends to meet the following conditions:

- The region is close to a circle (with high compactness degree);
- The region shape is close to convex (with high convex degree);
- The boundary is smooth with few angular arches on boundary.

Figure 3 illustrates that the Fourier descriptor of different shape types has the above intension by experiments. Figure 3A vs. figure 3A' reflects the difference between shapes with complex and smooth boundary. Figure 3B vs figure 3B' reflects the difference between concave shapes and that near to convex one. Figure 3C vs. figure 3C' reflects the difference between the shape near to circle and far away from circle. The graph under each pair describes the change of approximation accuracy.

For two given regions, the shape measure should pre-determine the measure accuracy according to the above approximation degree to control the expansion order. Let n_i be stop order when the expansion just satisfies the approximation accuracy. Select the larger one from two computed expansion orders $\{n_1, n_2\}$ as the truncated number of harmonics needed to index the shape in the computation of shape similar distance. It means the shape measure needs both two original region mapped to Fourier descriptor representations at enough accuracy. To compute the shape similar distance, two vectors $f_i=[d_i(1), d_i\ (2), d_i\ (3),\ldots\ d_i\ (n1-1)]$, $f_j\ =[d_j(1), d_j\ (2), d_j\ (3),\ldots\ d_j\ (n2-1)]$ have to be matched with the same dimension. So select the large stop order as the truncated number.

4 Shape based spatial query

In this section we discuss the applications of Fourier transform based shape query to retrieve objects from spatial database. We take the building feature as the experiment data.

The shape based spatial query usually behaves as the judgment that one object is like another one or two objects are similar in shape structure.

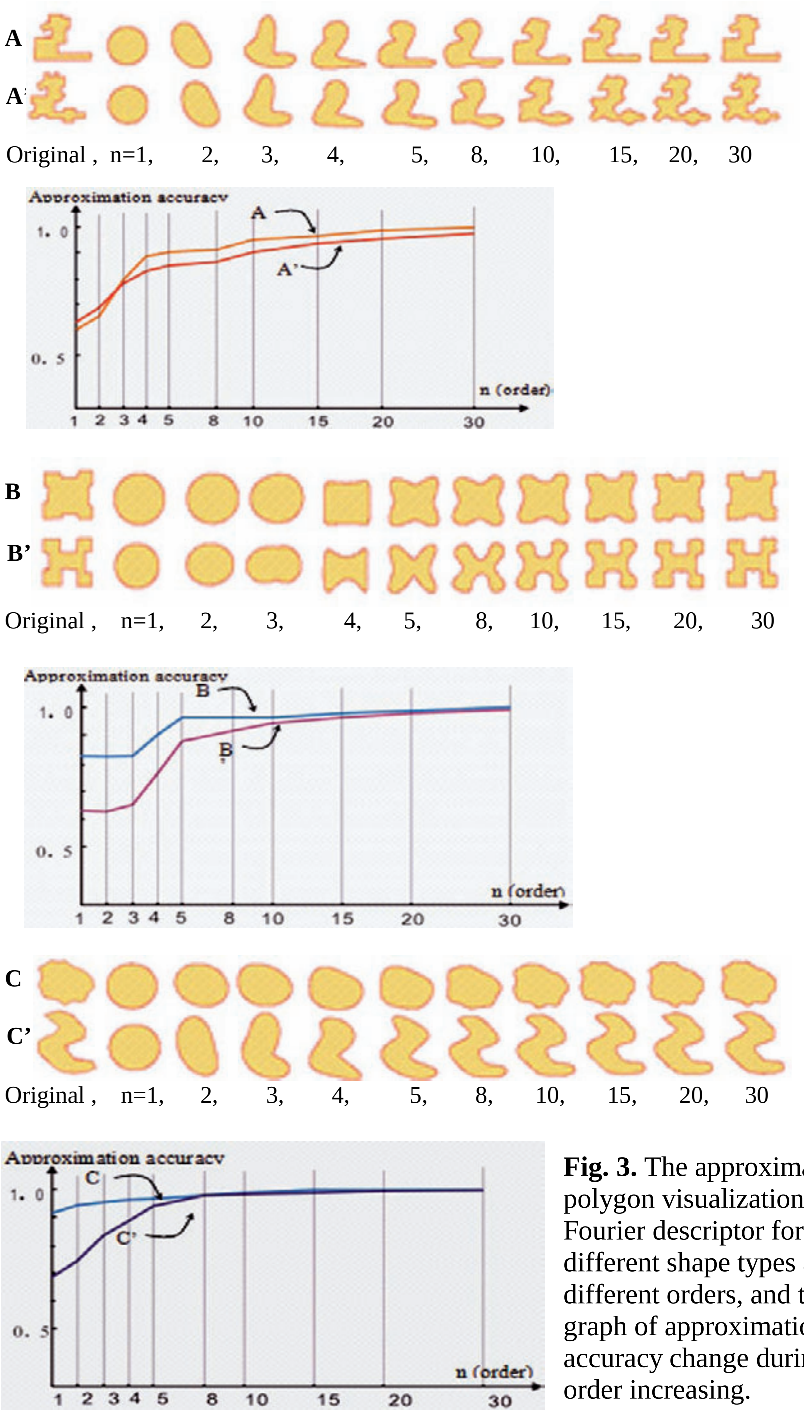


Fig. 3. The approximated polygon visualization by Fourier descriptor for different shape types at different orders, and the graph of approximation accuracy change during the order increasing.

The query can be formed as the judgment by a template shape. The SQL statement can be given as

Select {O_i} From DataBase Where O_i.shape LIKE <Template>

The operation *LIKE* is computed by Fourier transform method. The <*Template*> building in spatial database acts as some typical simple buildings which look like the shape of some letter, some special geometric construction and so on. In this study we take special letters as building template, such as *"L" "T" "U"* shape.

As mentioned in section 2, the judgment of shape similarity is indeterminate with different degrees. More precisely, the above SQL statement should be enhanced as

Select {O_i} From DataBase Where O_i.shape LIKE <Template> At_degree <c_i>

The term *At_degree* <c_i> describes the shape similarity to different extent. This query is similar to the idea of fuzzy mathematics. We use a membership function to represent the degree of "*LIKE*", which is defined on the basis of shape similar distance.

The shape similar distance between given template building and other objects to be queried is difficult to be normalized. The distance value and range depend on the shape complexity and template form. For different template building, the shape similar distance ranges quite differently. So we can not define a uniform membership function for all template buildings. Instead one template building applies one membership function.

To let the shape based query consistent with human's cognition, the quantitative representation of shape similarity distance has to be transferred to a qualitative representation which correctly reflects the spatial cognition in our mental world. For the degree to which two shapes are like, we distinguish four classes, namely *similar very much, moderately similar, commonly similar, slightly similar* denoted as c_1, c_2, c_3, c_4 respectively. The membership degree of shape similarity can be such as 0.8, 0.6, 0.4, 0.2 respectively corresponding to 4 fuzzy terms above.

By cognition experiment determines the relationship between c_i and the shape similar distance. First construct a series of building sample with reference to the template building. Based on the order of shape similar distance from small to large, let experiment participants judge the similarity degree and interrupt the linear order into 4 stages. To make building shape easily comparable, we let the shape is changed progressively facing different objective based on the idea of conceptual neighborhood. The sample data is organized in a radial structure as shown in figure 4. Summarize the judgments from different participants to determine the value of c_i. For the limitation of paper page, the detailed cognition experiment is omitted. Here we just list the result of cognition experiment for an example of "T" shaped template building. Figure 5 is the building samples with corresponding shape similar distance to the referenced template. According to the result of cognition experiment, the qualitative representation of shape similarity can be defined as the following function.

$$\text{Similarity} = \begin{cases} C_1 \text{ (very much)} & 0 \le \text{dis} < 0.4 \\ C_2 \text{ (moderately)} & 0.4 \le \text{dis} < 0.7 \\ C_3 \text{ (commonly)} & 0.7 \le \text{dis} < 1.0 \\ C_4 \text{ (slightly)} & 1.0 \le \text{dis} < 1.2 \end{cases}$$

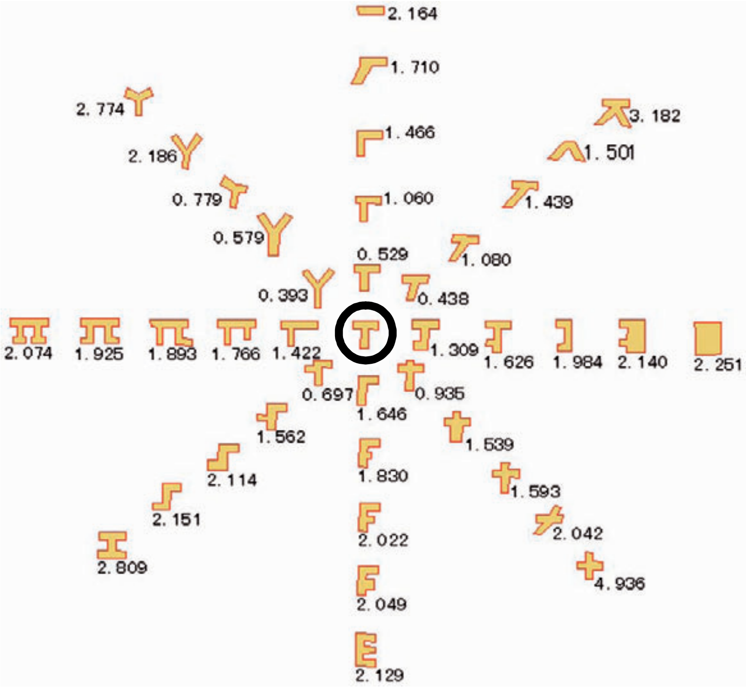
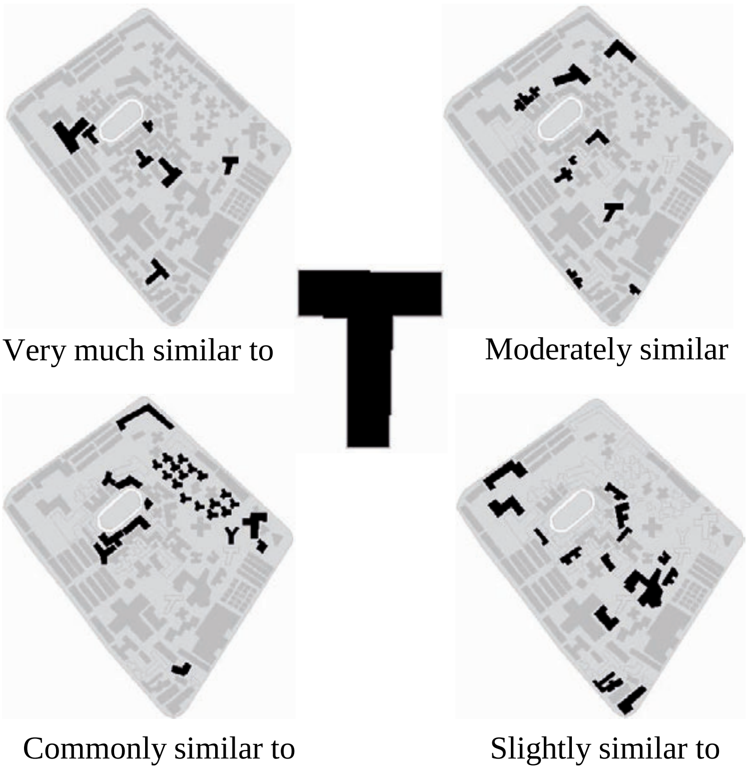


Fig. 4. The sample buildings and the shape similarity distance referenced to the central template

Based on the function of qualitative shape similarity representation and the template building, the SQL query can be actually performed by the arithmetic comparison on shape similarity distance to obtain different similar objects at degree varying from 0.0 to 1.0.

Figure 5 shows the result of shape based query from a set of building features by a “T” shaped template. Four groups of shaded polygon correspond to the buildings very much, moderately, commonly and slightly similar to the template respectively.



Very much similar to

Moderately similar

Commonly similar to

Slightly similar to

Fig. 5. The shape based query extracts buildings respectively *very much, moderately, commonly* and *slightly* similar to the template.

5 Conclusion

As a significant feature in spatial cognition, shape plays an important role in spatial data handling. Shape in some degree is the result of the evolution of geographic entity and the interaction with phenomena context in history. For example, the river shape and drainage pattern has something to do with hydrological and geological conditions in natural environment. By shape identification and analysis, we can discover some special spatial characteristics and pattern principles hidden behind the object.

Shape based query belongs to the data handling at high level with properties of abstraction, indetermination and symbolization requiring the knowledge of spatial cognition to support. We can anticipate to neither design an uniform query operation nor get a determine result by shape query, because it is too difficult in shape description, measurement and mathematically modeling.

For shape based query, the SQL statement can be formed as

Select {Oi} From DataBase Where O_i.shape LIKE <Template> At_degree <c_i>.

This study presents a method of shape query applying Fourier transform to measure shape. The region, e.g. building feature, is first approximated by Fourier descriptor series and then by a distance comparison of coefficient vector to find the similarity degree between candidate buildings and given template. As the shape similarity distance is not normalized to all templates, we develop the similarity membership function for each template and through cognition experiments to establish the classes of shape similarity as *very much*, *moderately, commonly* and *slightly* similar. Through experiments on building feature, the shape based query can extract different group of buildings consistent with human cognition. Due to the shortcomings of Fourier transform in shape representation, the above "*LIKE*" operation is sensitive to shape structure not adaptable to all shapes. For example, if the polygon is too concave, such as star shape or the boundary is too complex with too many angular, the boundary of approximated polygon may intersect. For shape signature, there are different methods in Fourier transform including central distance, complex coordinates (position function), curvature and cumulative angular function, adaptable to different situations respectively. In this study we just consider the position based method and in the future other methods need to test and make comparisons.

Acknowledgements

This research is supported by the program for new century excellent talents in university, China under the grant number NCET-05-0623.

References

Ai T, Guo R, Liu Y (2000) A Binary Tree Representation of Bend Hierarchical Structure Based on Gestalt Principles. Forer P., Yeh A.G.O., He J. (eds) Proceedings of the 9th Int. Sym. on Spatial Data Handling, Beijing, 2a30-2a43

Arkin EM, Chew LP et al (1991) An Efficient Computable Metric for Comparing Polygon Shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 13 (3), pp 209-216

Avelar S, Muller M (2000) Generating Topologically Correct Schematic Maps. Forer P., Yeh A.G.O., He J. (eds) Proceedings of the 9th International Symposium on Spatial Data Handling, Beijing, 4a28-4a35

Basri R, Costa L, Geiger D, Jacobs D (1998) Determining the similarity of deformable shapes. Vision Research, Vol. 38, pp 2365-2385

Bengtsson A, Eklundh J O (1991) Shape Representation by Multiscale Contour Approximation. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 13 (1), pp 85-93

Brassell KE , Weibel R (1988) A Review and Conceptual Framework of Automated Map Generalization. International Journal of Geographical Information Systems, 2(3):229-244

Cho-Huak T, Roland TC (1988) On image analysis by the methods of moments. IEEE Trans. On Pattern Analysis and Machine Intelligence, 10(4):496-513

Dent BD (1975) Communication Aspects of Value-By-Area Cartograms, The American Cartographer, 2(2): 154-168

Freksa C (1991) Qualitative spatial reasoning. Mark, D and Frank, A. U.(Eds.), Cognitive and Linguistic Aspects of Geographic Space. Dordrecht: Kluwer, pp 361-372

Hu Ming-Kuei (1962) Visual Pattern Recognition by Moment Invariants. IRE Transaction on Information Theory, vol. 8, pp 179-187

Jones C and Ware M (1998) Matching and Aligning Features in Overlayed Coverages. Proceedings of the 6th ACM international symposium on Advances in geographic information systems Washington, D.C., United States, pp 28-33

Kauppinen H, Seppanen T, Pietikainen M (1995) An Experimental Comparison of Autoregressive and Fourier-Based Descriptors in 2D Shape Classification. IEEE Trans. PAMI-17(2):201-207

Kim Y-S, Kim W-Y (1997) Content-Based Trademark Retrieval System by using Visually Salient Feature. In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, pp 307-312

Latecki LJ, Lakämper R (2002) Application Of Planar Shape Comparison To Object Retrieval In Image Databases. Pattern Recognition, 35(1), pp15-29

Li Z, Yan H, Ai T (2004) Automated Building Generalization based on Urban Morphology and Gestalt Theory. International Journal of Geographic Information Sciences, 18(5): 513-534

Lu G, Sajjanhar A (1999) Region-based shape representation and similarity measure suitable for content-base image retrieval. Multimedia Systems, 7:165-174

Palmer S.E (1999) Vision science: photons to phenomenology, MIT Press, Cambridge, MA

Persoon E, Fu KS (1977) Shape Discrimination Using Fourier Descriptors. IEEE Trans. On Systems, Man and Cybernetics, Vol.SMC-7(3):170-179

Rainsford D. and Mackaness W (2002) Template Matching in Support of Generalization of Rural Building, Richardson D and van Oosterom (Eds.)Advances in Sapatial Data Handling, Springer Verlag, pp137-151

Shekhar S, Chawla S (2002) Spatial Databases: A Tour, Prentice-Hall

Classification of Landslide Susceptibility in the Development of Early Warning Systems

Dominik Gallus[1], Andreas Abecker[1], Daniela Richter[2]

[1] Research Center for Information Technologies (FZI), Haid-und-Neu-Str. 10-14, 76131 Karlsruhe, Germany,
email: {gallus|abecker}@fzi.de

[2] Institute for Photogrammetry and Remote Sensing (IPF), University of Karlsruhe (TH), Englerstr. 7, 76128 Karlsruhe, Germany,
email: Daniela.Richter@ipf.uni-karlsruhe.de

Abstract

Statistical classification techniques complemented by the use of GIS have been shown to yield good results at the task of an assessment of landslide hazard/ susceptibility. In this work, several classification methods previously applied to this task are compared with respect to their performance on data sampled from distinct alpine areas in Vorarlberg, Austria. It is shown that among different types of techniques, kernel methods, including the Support Vector Machine and the Gaussian Process model, outperform techniques traditionally employed for the task. As further result, hazard maps for the study areas are generated, which can be used as input for suitable early warning systems focussing on landslide hazard.

Key words: landslide, classification, early warning system, GIS

1. Introduction

In populated mountainous regions, natural disasters resulting from large mass movements, including different types of slope movements and avalanches, are responsible for a large share of human and material damage. According to statistics from the International Disaster Database (EM-DAT), a number of more than 10000 people in total is reported to have been affected by disastrous mass movements in different alpine regions of Austria in the last 50 years, placing this type of natural hazard among the potentially most catastrophic in this part of Europe, next to large-scale flood events (EM-DAT[1] 2007).

Following heavy rainfall periods in winter/ spring of 1999 and summer of 2000, more than 250 avalanches, landslides and other slope movements have been registered in the administrative region Vorarlberg in west Austria, resulting in estimated damage in the order of 180 million €[2].

Statistical classification methods from the fields of machine learning and pattern recognition have been shown to yield good results at the task of an assessment of landslide hazard/ susceptibility. In combination with GIS, hazard maps resulting from fitting classifiers on existing data can be produced with relative ease, reducing necessary time and effort, with the advantage of introducing a principled and objective methodology.

This paper is structured as follows: In chapter 2, a short overview on basic concepts of statistical classification is provided. In chapter 3, a short review of related work is given. In chapter 4, the study areas and input data are described. In chapter 5, we present the methods used (logistic regression, both with and without backward parameter elimination, Gaussian Process models and the SVM). In chapter 6, classification results are shown and discussed. In chapter 7, we conclude with a summary, complemented by ideas for further research.

[1] http://www.em-dat.net

[2] http://www.vlr.gv.at/vorarlberg/finanzen_abgaben/finanzen/landesbudget/weitereinformationen/
rechnungsabschluesse/rechnungsabschluss2005.htm

2. Classification

2.1 The classification task

Classification can be defined as the procedure of building a mapping from a set of objects X into a set of classes C based on certain characteristics of the objects, often referred to as features or attributes.

More formally, when given a parametric or non-parametric model M (the classifier) and a set of objects x_i with known class labels t_i, $T_1 := \{(x_i, t_i) \mid i = 1...N\}$[3] (the training set), the task of classification consists of building a mapping $h : X \to C$ based on the values of the objects' features $\vec{\theta}_i$, which, when given an object $x_l \in X$, will produce the correct class label $t_l \in C$.

Since in general, $T_1 \subset X$ [4], the mapping h is an approximation of a hypothetical true mapping, subject to statistical limitations. As a consequence, the performance of the classifier will vary, depending on the size of the training set N, the (dis-)similarity of the elements in T_1 and in a (test) set of unseen objects T_2[5] and the number and choice of the objects' features.

2.2 Classifier performance

A common measure for the performance of the classifier is the misclassification error, defined as the proportion of wrongly classified objects. The misclassification error can be measured on the training set (the training error) or the test set. Because T_1 is limited, the training error tends to be biased towards sets of objects following a distributional structure similar to the one implied by T_1, and is also referred to as the apparent misclassification error. Also, since in general, $T_1 \neq T_2$, the training error will be too optimistic with respect to a hypothetical true misclassification error, resulting from training on an infinite training set.

[3] $x_i \in X, t_i \in C$

[4] $T_1 \neq X$

[5] In general, $T_1 \cap T_2 = \varnothing$

One technique commonly used to produce a less biased measure of classification performance is (k-fold) cross-validation. The procedure consists of dividing the training set T_1 into a set of k disjoint sets, each holding $|T_1|/k$ objects. After partitioning, the classifier is trained on (k-1) training sets, and tested on the remaining k-th set. This step is repeated k times, each time involving using a different permutation of (k-1) training sets, and a different test set. The effective performance of the classifier can be taken as the average of the k test errors.

Another widely used method for evaluating the classification performance in the case of binary classification[6] independent of application domain consists of plotting of the ROC (receiver operating characteristics) curve. The ROC curve is a plot of the proportion of objects of the class c_1 correctly classified as c_1 (*tpr*, the "true positives") vs. the proportion of objects of the class c_2 wrongly classified as c_1 (*fpr*, "false positives") for each value of *tpr*. It describes the ability of the classifier to correctly recognize objects belonging to c_1 (*sensitivity*) in relation to its (in-)ability to discriminate these objects from objects belonging to c_2 (1-*specificity*).

Another widely used measure contained in the ROC curve is the AUROC (area under ROC). This threshold-independent measure takes on values between 0 (no discrimination) and 1 (perfect discrimination). Like the misclassification error, the AUROC can be computed on the training set, or the test set.

3 Related work

A review of literature on statistical classification of landslide hazard/ susceptibility reveals that the multivariate logistic regression, a variant of the generalized linear model, is the most frequently chosen method (Atkinson, 1998), (Gorsevski, 2000b), (Ohlmacher, 2003). It is typically employed in an automatic stepwise model selection procedure, usually starting with a model including all available parameters, followed by their successive elimination according to a measure of parameter (in-) significance, or a measure of goodness of fit of the model penalized by model complexity. The procedure is aimed at reducing the parameter space in order to avoid overfitting – fitting too many model parameters to the training data, for which some of the features might be uninformative with respect to the class label, irrelevant for the task of discrimination, or simply noise. At

[6] $C:=\{c_1,c_2\}$

the same time, it can be used as an operational method to select “significant”, “essential” or “interesting” features.

In more recent research, several different methods from pattern recognition and machine learning have been proposed, including linear discriminant analysis (Gorsevski, 2000a), (Santacana, 2003), neural networks (Lee, 2003), (Ermini, 2005), and the Support Vector Machine (Brenning, 2005). These methods have been shown to yield good results at the task, comparable to or surpassing the performance of traditionally used techniques.

With respect to the task of classification, spatial data may give rise to certain problems. In the context of classification of spatial data, it has been shown that the property of spatial autocorrelation of grid points (pixels) may lead to invalid significance statements (Ohlmacher, 2003). In the case of the logistic regression, this has resulted in the development variants capable of modelling spatial autocorrelations (Gotway, 1997). However, it is not clear if the utility of these models – in terms of classification performance – outweighs the disadvantage of added complexity. This topic is further discussed in a comparative study (Brenning, 2005), hinting at the latter.

Another problem inherent in spatial data arises due to the property of a locality of feature occurrence. Depending on the choice of the training and test set, this can make classification of new data difficult, and estimated error rates on the training set may be hardly transferable out of its spatial (and also temporal) scope, due to the occurrence of different dispositive (geology) or triggering (precipitation, earthquake) factors. This is a major reason for the objective difficulty of the task of a prediction of landslide occurrence.

4 Data

4.1 Study area

Three study areas in Vorarlberg, Austria have been chosen on the basis of various studies and field mappings carried out in context of the project “Georisikokarte” conducted by the Department of Applied Geology (AGK), University of Karlsruhe, in cooperation with the Federal Government of Vorarlberg (Ruff, 2005). The three study areas are: Hochtannberg (HTB), a region of 114 km² size situated at the eastern border of Vorarlberg, Walgau (WAL) with an area of 105 km², and Walsertal (WST), 147 km² in size (Fig. 1).

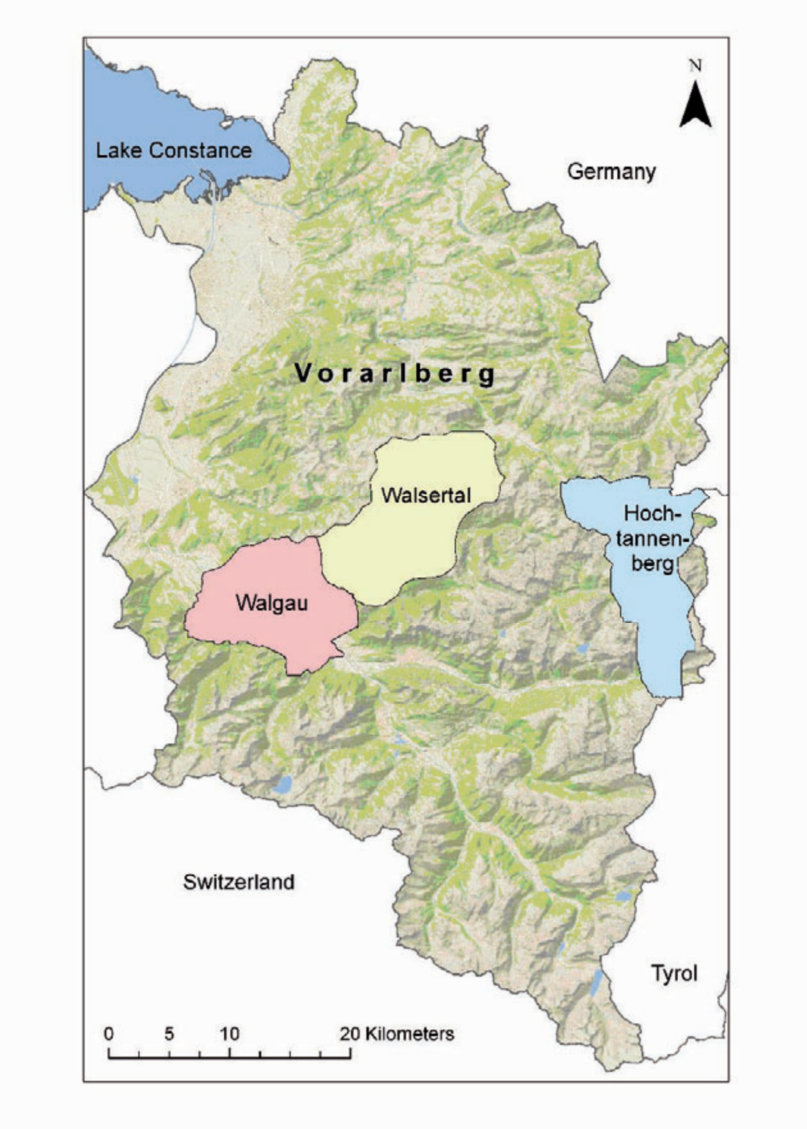


Fig. 1: Study areas in Vorarlberg/ Austria

4.2 Data & Pre-processing

Digital data used for classification was produced within ArcGIS, ArcInfo 9.2. A digital elevation model (DEM), topographic and geologic data and orthophotographs covering the whole area of Vorarlberg were provided by the Land Surveying Office Feldkirch (Austria). The DEM with a spatial resolution of 5 m was used to calculate various morphometric features, including slope, curvature, slope aspect and flow accumulation. The slope was computed as the rate of maximum change in z value from each cell (i.e., the first derivative of the surface). The curvature of a slope, referring to concavity/ convexity of a surface, was computed using the CURVATURE function within ArcGIS, with the primary output of the function resulting from subtracting the profile component from the planar component. For the slope aspect (slope direction), the output is defined as the direction of maximum rate of change in z value from each grid point. The flow accumulation grid was calculated by accumulating the slope for all grid points flowing into each down-slope grid point, and was included as an indicator of erosion effects.

Digital data on geology and tectonics of Vorarlberg is based on the geologic map by R. Oberhauser, Geological Federal Institution Vienna (scale 1:100.000)[7]. The vector data includes a number of classes describing the geology and line features for tectonic faults. The data was used to create a grid of 25 m grid point size. Additionally, a raster of Euclidean distances to faults was calculated, with the shortest distance to a tectonic fault assigned to every grid point within the area of investigation.

The Austrian land cover database for 1990, supplied by Umweltbundesamt GmbH (Corine Landcover nomenclature: Level 2, 25 ha minimum mapping unit, comparable scale: 1:100.000), was derived from satellite data (Landsat 5, TM) by means of computer-assisted visual photo-interpretation. For Vorarlberg, it consists of 12 classes, including built-up area (4), agricultural area (3), forests and natural area (3), wetlands (1) and water surfaces (1)[8].

In order to apply classification techniques to the study areas, training data was constructed drawing on results from the project “Georisikokarte”. In context of the project, active landslides were mapped and an inventory of landslides for the study areas was constructed. Within the study area Hochtannberg, 107 landslides (with a total area of 0,79 km^2) were mapped. Additional inventories were constructed for Walgau (262 slides with a total area of 0,62 km^2) and Walsertal (field mapping until 2004, 761 slides with a total area of 3,54 km^2).

In course of pre-processing, all data layers including feature layers and landslide inventories were converted into the ArcInfo ASCII grid format,

[7] http://www.vorarlberg.at/geokatalog_internet/index.htm

[8] http://www.umweltbundesamt.at/umwelt/raumordnung/flaechennutzung/corine/

with a size corresponding to the size of the corresponding study area, and a resolution of 25 m/ grid point.

5 Classification Methods

In this work, four different classification techniques were used: Logistic regression (standard and stepwise, i.e. without and with backward parameter elimination), Gaussian Process models and the SVM. Each classifier was trained on training sets $T_k=\{(x_i, t_i) \mid i=1\ldots n_k\}$, consisting of sets of grid points (pixels) x_i, each x_i having a m-dimensional feature vector ϕ_i taking on values sampled from a set of m thematic layers (ArcInfo grids)[9], and a known class label t_i, indicating if the grid point was part of an area where a landslide occurrence was registered (t_1:=1) or not (t_2:=0).

All calculations were performed within the statistical computing framework R[10], using the additional packages *MASS*, *nnet*, and *kernlab*.

5.1 Logistic Regression

Logistic regression has a tradition of being widely employed for the task of statistical classification of landslide hazard/ susceptibility. This is due to several desirable properties of the method, including conceptual simplicity, straightforward interpretability in terms of probability estimation, and a decent performance at simple tasks.

In addition, logistic regression readily lends itself to efficient and robust optimization schemes, making it one of the more efficient methods used.

In terms of binary classification, given a set of m-dimensional feature vectors ϕ_i of grid points x_i with corresponding class labels t_i contained in the training set, the posterior probability of x_i belonging to class c_1 is given by

[9] see chapter 2

[10] R: A language and environment for Statistical Computing, the R Development core team, R Foundation for Statistical Computing, Vienna, Austria, http://www.r-project.com

$$p(c_1 | \vec{\phi}_i) = \frac{p(\vec{\phi}_i | c_1) p(c_1)}{p(\vec{\phi}_i | c_1) p(c_1) + p(\vec{\phi}_i | c_2) p(c_2)} \tag{1}$$

$$= \frac{1}{1 + e^{-\left(\ln\left(\frac{p(\vec{\phi}_i | c_1) p(c_1)}{p(\vec{\phi}_i | c_2) p(c_2)}\right)\right)}} = \frac{1}{1 + e^{-\vec{w}^T \vec{\phi}_i}} = \sigma(\vec{w}^T \vec{\phi}_i) = y(\vec{\phi}_i)$$

with the logistic sigmoid function σ and the inner product $\mathbf{w}^T\phi_i$, defined as the logit, or log odds of x_i belonging to class c_1, and class c_2:

$$\vec{w}^T \vec{\phi}_i = \ln\left(\frac{p(\vec{\phi}_i | c_1) p(c_1)}{p(\vec{\phi}_i | c_2) p(c_2)}\right) \tag{2}$$

In order to find a maximum likelihood solution for **w**, usually an efficient iterative higher order optimization scheme operating on the (log-) likelihood function is employed. This is true for both methods used in this work, *glm* and *multinom*, using the IRLS[11] and BFGS optimization schemes, respectively.

In the case of the stepwise regression procedure, the *stepAIC* method (package *MASS*) was used, which, by default, uses the AIC (Akaike Information Criterion) as criterion for model selection, defined as

$$AIC := -\ln p(\vec{t} | \vec{w}) + 2m \tag{3}$$

the first term being the negative logarithm of the likelihood function given the weight vector **w** and a vector of class labels **t**.

5.2 Gaussian Process model

Gaussian Process models are a class of non-parametric probabilistic discriminative models more recently introduced in the fields of pattern recognition and machine learning (Williams, 1995), (Williams, 1998), (MacKay, 1998). From the point of view of classification methodology, Gaussian Process models are interesting since they can be naturally expressed in the conceptual framework of Bayesian inference whilst making use of a kernel (or covariance function), a function of the elements x_i of the training set, for representation.

In the case of classification, given a model trained on $T=\{(x_i, t_i) \mid i=1\ldots n\}$, a predictive distribution for the class label t_{n+1} given a x_{n+1} is defined by

$$p(t_{n+1} = 1 | \vec{t}) = \int p(t_{n+1} = 1 | y_{n+1}) p(y_{n+1} | \vec{t}) dy_{n+1} \tag{4}$$

with $p(t_{n+1}=1|y_{n+1})=\sigma(y_{n+1})$, the sigmoid logistic with argument y_{n+1}.

[11] Iterative reweighted least squares

Since the posterior probability p(y_{n+1}|**t**) is non-Gaussian, this integral is analytically intractable, and is evaluated in approximation. Here, the Laplace approach to obtain a Gaussian model for p(y_{n+1}|**t**) is used, resulting in expressions for the moments:

$$E(y_{n+1}|\vec{t}) = \vec{k}^T(\vec{t} - \vec{\sigma}) \tag{5}$$

$$Var(y_{n+1}|\vec{t}) = c - \vec{k}^T(W^{-1} + C)^{-1}\vec{k} \tag{6}$$

with **k** a vector of values from a positive semi-definite kernel function *k* evaluated on the set of tuples {($\mathbf{x}_i$, $\mathbf{x}_{n+1}$)|i=1…n}[12], **σ** a vector with the elements σ(y_i), c the value of *k* evaluated at ($\mathbf{x}_{n+1}$, $\mathbf{x}_{n+1}$), **W** a n-dimensional diagonal matrix with elements σ(y_i)(1-σ(y_i)), and **C** the n-dimensional covariance matrix of the Gaussian Process y(**x**), equivalent to the Gram matrix of the kernel *k*.

Using the above result, the integral can be evaluated, resorting to an approximation to the convolution of a logistic sigmoid σ with a Gaussian, yielding (Bishop, 2006):

$$p(t_{n+1} = 1|\vec{t}) \quad \cong \sigma(\kappa(Var(y_{n+1}|\vec{t}))E(y_{n+1}|\vec{t})) \tag{7}$$

with

$$\kappa(Var(y_{n+1}|\vec{t})) = (1 + \pi Var(y_{n+1}|\vec{t})/8)^{-\frac{1}{2}} \tag{8}$$

Gaussian Process-based regression is reported to yield good performance on different tasks from the prediction of chemical properties of molecules in drug design to Wi-Fi localization. In the area of geostatistics, an equivalent (low-dimensional) form known as kriging (Cressie, 1993) is used extensively for interpolation.

In R, the Gaussian Process model for regression and classification is implemented in the *gausspr* method, which is part of *kernlab* (Karatzoglou, 2004). In this work, the default Gaussian kernel *rbfdot* was used, in the functional form

$$k(\vec{x}, \vec{x}') = e^{(-\|\vec{x} - \vec{x}'\|^2/2\sigma^2)} \tag{9}$$

with the hyperparameter σ, which was estimated using the *sigest* estimation routine available in *kernlab*.

[12] For the sake of convenience, we will henceforth refer to $\mathbf{x}_i$ as the feature vector of x_i.

Support Vector Machine

The Support Vector machine (Cortes, 1995), (Vapnik, 1998) has emerged in the 1990's as a novel method for classification and regression. Like the Gaussian Process models, it belongs to the class of kernel methods, i.e., it can be expressed as a function of a symmetric positive semi-definite kernel *k*(**x**, **x'**), evaluated at the elements of the training set T.

The design of the SVM lends the model several properties making it an excellent tool for classification.

One such property is that for the SVM, determination of model parameters (training) takes form of convex optimization. As a consequence, the training procedure for SVM is guaranteed not to produce a suboptimal solution with respect to the global optimum.

An important distinction of the SVM to the simple linear model is the introduction of the concept of the margin. The SVM training procedure is specifically aimed at this quantity, defined as the perpendicular distance between the decision boundary and the closest of the data points. Maximizing the margin leads to a particular choice of the decision boundary, with a maximal separation of the classes. This property lends the model improved generalization capabilities with respect to unseen data.

Another property resulting in better generalization capabilities of the SVM is the adaption of a "soft" margin. With the soft margin property, the SVM is allowed to make errors in classification on the training set. Due to this property, the SVM is less sensitive to the presence of outliers in the training set, and hence, less prone to overfitting.

In the dual (kernel) formulation, the definition of the SVM is

$$y(\vec{x}) = \sum_{n=1}^{N} a_n t_n k(\vec{x}, \vec{x}_n) + b \tag{10}$$

with Lagrange multipliers a_n resulting from a quadratic optimization of the term

$$\tilde{L}(\vec{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m k(\vec{x}_n, \vec{x}_m) \tag{11}$$

subject to the constraints

$$0 \le a_n \le C$$

$$\sum_{n=1}^{N} a_n t_n = 0$$

The dual formulation of the SVM introduces the kernel function *k*, which can be interpreted as the inner product of the inputs in a feature space. Since it does not require an explicit definition of the mapping into feature space for each pair of arguments ($\mathbf{x}_i$, $\mathbf{x}_{i'}$), it is possible to substitute the inner product kernel, defined as

$$k(\vec{x},\vec{x}')=\vec{\phi}(x)^T\vec{\phi}(x') \tag{12}$$

with another symmetric, positive semi-definite kernel *k* to separate the data in some (possibly high or even infinite dimensional) induced space[13].

The direct solution of a quadratic programming problem is computationally demanding. For the above quadratic optimization problem, efficient procedures like the SMO procedure (Platt, 1999) have been developed.

In practice, the procedure is found to have a scaling with N between linear and quadratic.

A suitable model for the prediction of class probabilities can be obtained by fitting a logistic sigmoid to the output of a previously trained SVM. This method is based on an idea presented in (Platt, 2000).

Within R, there are several packages implementing SVM. In this work, the *ksvm* method in package *kernlab* was used, in the described variant of SVM (C-SVM) and the Gaussian kernel *rbfdot*. In analogy to the Gaussian Process case, the hyperparameter σ was estimated using *sigest*.

6 Results

6.1 Pre-processing

Four classifiers were compared on four independent data sets, drawing on methods described in chapter 5. In each case, the model was trained on a balanced set of 50% landslide and 50% non-landslide grid points, and subject to 5-fold cross-validation. In order to determine the influence of training set size on classification performance and stability, training sets of decreasing size were used, corresponding to fractions of the full data set size.

For each dataset, a set of m thematic layers was used including morphometric data (elevation, slope, curvature, and aspect), geology, vegetation, flow accumulation and Euclidean distance to faults. The layers were derived from a DEM, a geologic, and a geotechnical map using standard GIS functionality. In course of further pre-processing, real-valued features (elevation, slope, curvature, flow accumulation, distance to faults) were standardized. Non-continuous valued features (slope orientation, geology,

[13] This is often referred to as the kernel trick.

and vegetation) were transformed to a number of binary features equivalent to the number of values assumed on the dataset.

In the case of the third study area (WST), the data, consisting of more than > 11000 landslide pixels in addition to 228530 non-landslide pixels, was split into two (roughly) equal sized datasets corresponding to the north and the south part of the study area in order to obtain training sets of comparable size.

The size of each training set, the number of features and the overall size of the study area are summarized in Table 1.

Table 1: Training data

Study area	**Dataset size [pixels]**	**# of features**	**landslide area [pixels]**	**Non-landslide area [pixels]**	**landslide/non-landslide area [pixels]**
HTB	2400	59	1200	176277	0.0067
WAL	1960	61	980	166235	0.0058
WST_N	4950	60	2475	117592	0.0206
WST_S	6320	51	3160	110938	0.0276

6.2 Classification performance

A summary of the results (Table 2-5) shows that both kernel-based classifiers consistently outperform the logistic regression (both with and without parameter elimination). This is true for all combinations of study area and training set size. The SVM shows overall best results, achieving a classification performance greater than 80% in three out of four cases, performing best on the third dataset with full training set size (84.5%). Gaussian Process models turn out competitive, with a classification performance of less than 2% worse than the SVM on average.

Conversely, results of the logistic regression in both variants set the method apart from the SVM and Gaussian Process models. The difference in performance ranges from 4.5% in the best case (dataset 2) up to 12.2% on the first dataset with full training set size. Reducing the size of the training set seems to have less of an effect on the classification performance, with relatively constant results except for a notable decrease in case of dataset 2. In the case of dataset 1 (HTB), a comparison of training error vs. test error for training sets of decreasing size shows signs of overfitting on the part of logistic regression, with a divergence of error rates on training set (lower error) and test set. A corresponding trend, albeit less distinct, can be observed in the case of the stepwise variant.

In part, the inferior performance of the logistic regression might be considered a result of the relative simplicity of the model, which might be interpreted as a "soft" version of a threshold-based discriminant $\mathbf{w}^{T}\phi$ in the logit space. By definition of the logit:

$$w^{T}\vec{\phi}_{i} = \ln \frac{p(\vec{\phi}_{i} \mid C_{1})p(C_{1})}{p(\vec{\phi}_{i} \mid C_{2})p(C_{2})} = \sum_{j=1}^{M} \ln \frac{p(\phi_{ij} \mid C_{1})p(C_{1})}{p(\phi_{ij} \mid C_{2})p(C_{2})} \tag{13}$$

the logit space might appear as a natural feature space for the task of probabilistic discrimination. However, the relative simplicity of the model might give rise to certain problems.

First off, a linear separation based on logit threshold might not exist in the (one-dimensional) logit space, which might turn out sub-optimal for discrimination, e.g. due to outliers in the likelihood ratio for a feature ϕ_i.

Another source of error might be introduced by the model's assumption of independence of $p(\phi_{ij}|C_k)$ and $(\phi_{i'j'}|C_k)$, for j and j'[14]. This assumption does not hold in our case, with at least one example of (non-linear) functional dependence for the features (the morphometric parameters).

A different assumption of independence is implicit in the formulation of the likelihood function, defined as

$$p(\vec{t} \mid \vec{w}) = \prod_{i=1}^{N} y_{i}^{t_{i}} (1 - y_{i})^{1-t_{i}} \tag{14}$$

In this case, independence of the grid points is assumed for i, i'[15]. However, this assumption might not hold in the presence of spatial correlation, resulting in an approximation to the true likelihood.

As a final observation, an examination of the logistic regression model in the standard variant reveals missing regularization. In consequence, the method is more prone to overfitting, resulting in inferior generalization capabilities with respect to better regularized methods.

[14] $j, j' \in \{1...M\}, j \neq j'$

[15] $i, i' \in \{1...N\}, i \neq i'$

Table 2: Classification performance for study area 1 (HTB)

Classifier	**Test error [%]**			
Training set size [pixel]	**1920**	**960**	**480**	**240**
Glm	33.5	34.5	35.1	35.4
Glm_stepAIC	32.3	33.5	34.4	35.2
Gausspr	23.2	25.2	27.7	29.2
Ksvm	21.3	23.8	25.5	27.6

Table 3: Classification performance for study area 2 (WAL)

Classifier	**Test error [%]**			
Training set size [pixel]	**1568**	**784**	**392**	**194**
Glm	24.3	24.8	25.1	30.2
Glm_stepAIC	24.1	24.6	25.8	31.4
Gausspr	21.8	22.3	23.6	25.4
Ksvm	19.8	21.2	21.7	25.7

Table 4: Classification performance for study area 3a (WST_N)

Classifier	**Test error [%]**			
Training set size [pixel]	**3960**	**1980**	**990**	**495**
Glm	25.2	25.6	25.2	26.4
Glm_stepAIC	25.3	25.6	25.3*	27.1
Gausspr	17.5	19.7	21.2	24.7
Ksvm	15.5	19.8	21.9	24.3

Table 5: Classification performance for study area 3b (WST_S)

Classifier	**Test error [%]**			
Training set size [pixel]	**5056**	**2528**	**1264**	**632**
Glm	25.6	25.8	26.3	26.1
Glm_stepAIC	26	26	26.6	26.8
Gausspr	20.8	21.8	22.8	24.3
Ksvm	19.9	20.7	21.8	22.8

ROC curves

ROC curves were drawn for each combination of study area and classifier for the full training set size using the *ROCR* package. A set of plots for the four classifiers for the first study area (HTB) is given in Fig. 1-4. AUROC values computed for the curves are given in Table 7, reflecting the classification performance in Table 2 (first column). Visual inspection of Fig. 1-4 and a comparison of the AUROC for the four classifiers confirm previous observations, with the kernel methods emerging as preferred models in terms of classification performance on unseen test data.

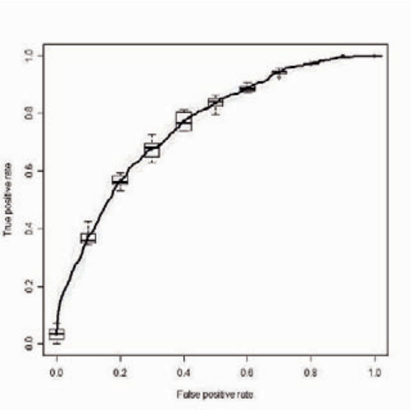


Fig. 2: HTB/ glm

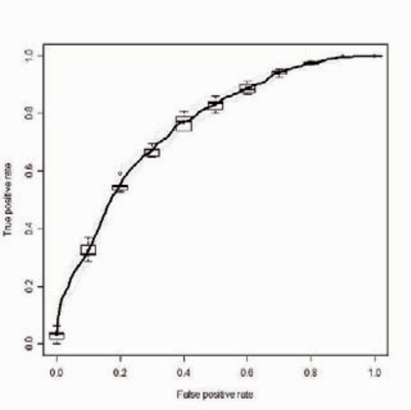


Fig. 3: HTB/ glm_stepAIC

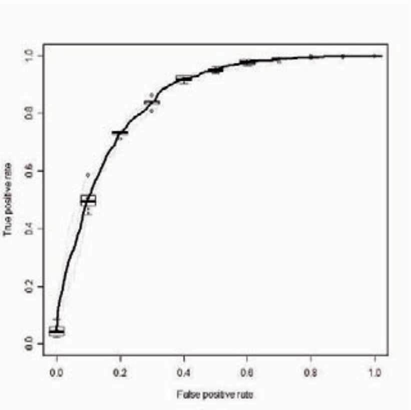


Fig. 4: HTB/ gausspr

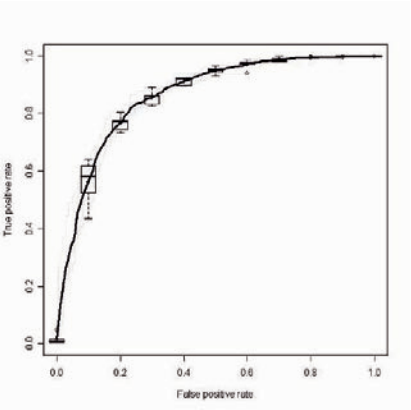


Fig. 5: HTB/ ksvm

Table 7: AUROC for different classifiers on study area 1 (HTB)/ full dataset

Classifier	**AUROC on test set [%]**					
	Run 1	**Run 2**	**Run 3**	**Run 4**	**Run 5**	**Avg.**
Glm	76.8	74.3	76.3	77.3	74.4	75.8
Glm_stepAIC	76.3	74.2	76.1	76.0	74	75.3
Gausspr	86.5	84.6	84.4	85.1	83.3	84.9
Ksvm	88.4	86.3	87.5	85.9	85.8	86.8

Hazard maps

After training, the models were used to generate hazard maps for each study area. For this task, all available training data (i.e., all 5 subsets) was used, and the trained models were applied to the whole study area, containing the total number of pixels (Table 1). Hazard maps were generated as achromatic 8 bit raster images, the value of each pixel being the output of a sigmoid (in the case of the logistic regression) or the output of the sigmoid fitted to the output of the respective model, scaled by the maximum 8 bit value. To facilitate post-processing using GIS functionalities, each map was exported in a number of raster formats, including PNG, TIFF, and ArcInfo ASCII Grid using methods contained in the *rgdal* package. Fig. 6 a/b -8 a/b show the (inverted) primary output of classification generated for the study areas HTB, WAL and WST by the best-performing SVM model, and results of ArcGis postprocessing, respectively.

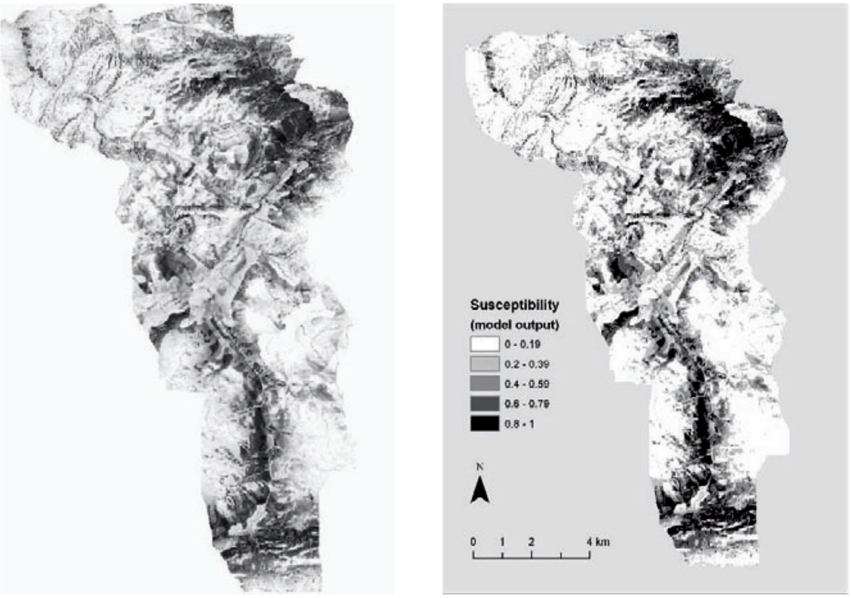


Fig. 6: a) (inverted) hazard map, R output (SVM) b) results of ArcGis postprocessing (HTB)

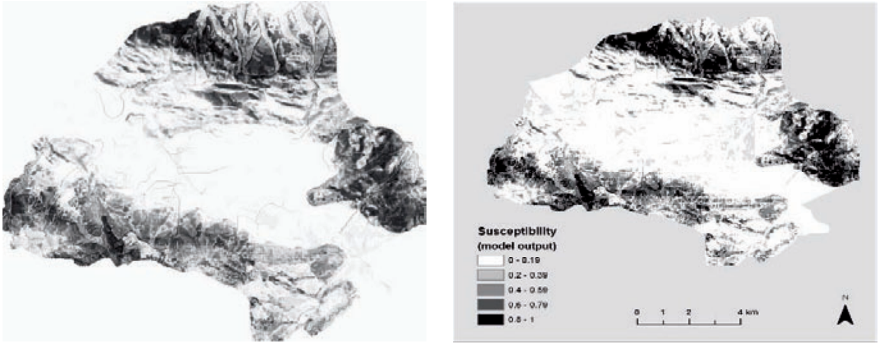


Fig 7: a) (inverted) hazard map, R output (SVM) b) results of ArcGis postprocessing (WAL)

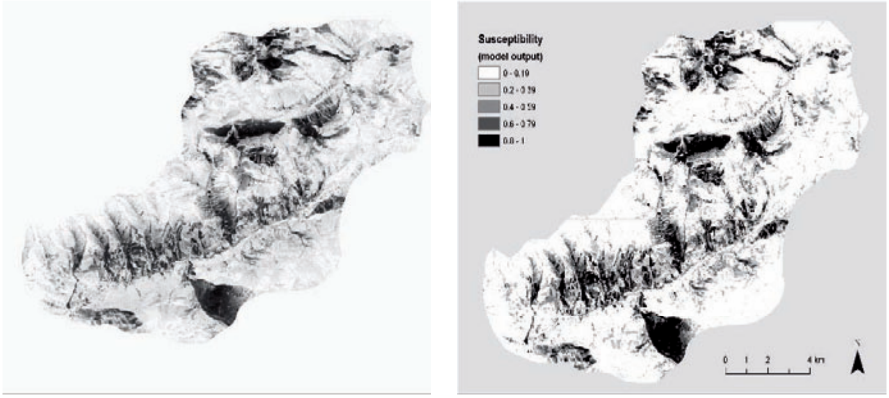


Fig. 8: a) (inverted) hazard map, R output (SVM) b) results of ArcGis postprocessing (WST)

7 Conclusions/ Outlook

In this work, several different classification techniques were applied to the task of an assessment of landslide hazard/ susceptibility. In a comparison of classification performance on unseen test data, it was shown that kernel methods, represented by the Gaussian Process model and the SVM, consistently outperform the logistic regression model, with misclassification errors for independent test datasets and different test set sizes hinting at robustness of the result.

The inferior performance of the logistic regression model can be explained by the relative simplicity of the model, which does not reflect (non-linear) dependencies inherent in the data. In addition, the lack of a built-in regularization mechanism on the part of the logistic regression explains its inferior generalization capabilities.

In contrast to the logistic regression, both kernel methods have the property of explicitly modelling a correlation between the data points by means of the kernel/ covariance function. Also, both kernel methods exhibit generalization properties superior to the logistic regression in both variants. SVM in particular are known to have improved generalization capabilities by design, resulting from the maximization of margin and the “soft” margin property. The Gaussian Process model is regularized by the Gaussian Process prior, a Gaussian with mean **0** and covariance function *k*.

Our interpretation of the results suggests several opportunities for improvement departing from the logistic regression model. Generalized linear mixed models and hierarchical models considering effects present in subsets of the data present a viable option. As an alternative, a systematic investigation of the prospective potential of kernel methods with respect to additional improvements at our task – based on a deeper understanding of the nature of spatial data – might offer opportunities for further research.

Acknowledgements

The authors thank Michael Ruff, Walter Bauer and Peter Drexel (Land Surveying Office Feldkirch) for providing the landslide inventories and thematic data for Vorarlberg.

This work was partly funded by the German Federal Ministry of Education and Research in the project “development of suitable early warning systems” (EGIFF), a part of the GEOTECHNOLOGIEN program.

References

ATKINSON, P. M., AND MASSARI, R. (1998) Generalized linear modelling of susceptibility to landslides in the Central Apennines, Italy. *Computers and Geosciences*, 24.

BISHOP, C. M. (2006) *Pattern Recognition and Machine Learning*, Springer.

BRENNING, A. (2005) Spatial prediction models for landslide hazards: review, comparison and evaluation. *Natural Hazards and Earth System Sciences*, 5.

CORTES, C., AND VAPNIK, V. (1995) Support-Vector Networks. *Machine Learning,* 20.

CRESSIE, N. A. C. (1993) *Statistics for spatial data,* New York, John Wiley & Sons.

ERMINI, L., CATANI, F., AND CASAGLI, N. (2005) Artificial Neural Networks applied to landslide susceptibility assessment. *Geomorphology,* 66.

GORSEVSKI, P. V., GESSLER, P.E., AND FOLTZ, R.B. (2000a) Spatial prediction of landslide hazard using discriminant analysis and GIS. *GIS in the Rockies 2000 Conference and Workshop.* Denver, Colorado, USA.

GORSEVSKI, P. V., GESSLER, P.E., AND FOLTZ, R.B. (2000b) Spatial prediction of landslide hazard using logistic regression and GIS. *4th International Conference on Integrating GIS and Environmental Modeling (GIS/EM4).* Banff, Alberta, Canada.

GOTWAY, C. A., AND STROUP, W. W. (1997) A generalized linear model approach to spatial data analysis and prediction. *Journal of Agricultural, Biological, and Environmental Statistics,* 2.

KARATZOGLOU, A., SMOLA J., HORNIK, K., AND ZEILEIS, A. (2004) kernlab – An S4 Package for Kernel Methods in R. *Journal of Statistical Software,* 11.

LEE, S., RYU, J.-H., MIN, K., AND WON, J.-S. (2003) Landslide susceptibility analysis using GIS and artificial neural network. *Earth Surface Processes and Landforms,* 28.

MACKAY, D. J. C. (1998) Introduction to Gaussian processes. IN BISHOP, C. M. (Ed.) *Neural Networks and Machine Learning.* Springer.

OHLMACHER, G. C., AND DAVIS, J. C. (2003) Using multiple logistic regression and GIS technology to predict landslide hazard in Northeast Kansas, USA. *Engineering Geology,* 69.

PLATT, J. C. (1999) Fast training of support vector machines using sequential minimal optimization. IN SCHOELKOPF, B., BURGES, C., AND SMOLA, A. (Ed.) *Advances in Kernel Methods - Support Vector Learning.* Cambridge, MIT Press.

PLATT, J. C. (2000) Probabilistic outputs for Support Vector Machines and comparison to regularized likelihood methods. IN SMOLA, A., BARTLETT, P., SCHOELKOPF, B., AND SCHUURMANS, D. (Ed.) *Advances in Large-Margin Classifiers.* Cambridge, Massachusetts, USA, MIT Press.

RUFF, M., KÜHN, M., AND CZURDA, K. (2005) Risikoanalyse für Massenbewegungen in den Ostalpen (Vorarlberg). IN MOSER, M. (Ed.) *15. Tagung Ingenieurgeologie.* Erlangen, Germany.

SANTACANA, N., BAEZA, B., COROMINAS, J., DE PAZ, A., AND MARTURIA, J. (2003) A GIS-based multivariate statistical analysis for shallow landslide susceptibility mapping in La Pobla de Lillet area (Eastern Pyrenees, Spain). *Natural Hazards,* 30.

VAPNIK, V. (1998) *Statistical Learning Theory,* New York, John Wiley and Sons.

WILLIAMS, C. K. I., AND BARBER, D. (1998) Bayesian Classification with Gaussian Processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 20.

WILLIAMS, C. K. I., AND RASMUSSEN, C.E. (1995) Gaussian Processes for regression. IN TOURETZKY, D. S., MOZER, M. C., AND HASSELMO, M. E. (Ed.) *Neural Information Processing Systems.* Denver, Colorado, USA, MIT Press.

Clusters in Aggregated Health Data [⋆]

Kevin Buchin, Maike Buchin, Marc van Kreveld, Maarten Löffler, Jun Luo, Rodrigo I. Silveira

Department of Information and Computing Sciences, Utrecht University, the Netherlands; email: {buchin, maike, marc, loffler, ljroger, rodrigo}@cs.uu.nl

Abstract. Spatial information plays an important role in the identification of sources of outbreaks for many different health-related conditions. In the public health domain, as in many other domains, the available data is often aggregated into geographical regions, such as zip codes or municipalities.

In this paper we study the problem of finding clusters in spatially aggregated data. Given a subdivision of the plane into regions with two values per region, a case count and a population count, we look for a cluster with maximum density. We model the problem as finding a placement of a given shape R such that the ratio of cases contained in R to people living in R is maximized. We propose two models that differ on how to determine the cases in R, together with several variants and extensions, and give algorithms that solve the problems efficiently.

Keywords: cluster, outbreak, algorithm, aggregated data.

1 Introduction

The study of geographical patterns of diseases is an important aid for the investigation of outbreaks. Analyzing the geographic nature of disease cases has been a key factor in finding the source of many outbreaks. The classical example is the outbreak of cholera in the Soho district of London in 1854 [2, 21]. The source of this outbreak, that left a death toll of more than 600 people in about 10 days, was found by John Snow, a London physician. He realized that most affected people lived around a public pump, which was later confirmed as the source. Numerous other examples have been documented since then in the literature of several fields like epidemiology, public health, preventive medicine, and medical geography.

[⋆] This research was initiated during the GADGET Workshop on Geometric Algorithms and Spatial Data Mining, funded by the Netherlands Organisation for Scientific Research (NWO) under BRICKS/FOCUS grant number 642.065.503, and has been also partially funded by NWO under the project GOGO.

Investigation of outbreaks due to both infectious and noninfectious causes (e.g., toxic exposure) can greatly benefit from the use of spatial information. Even though the role played by geography in the identification of the source depends entirely on the disease, spatial factors have a major importance for many point source outbreaks related to exposure to pollution or radiation sources (for a wide range of diseases, from respiratory illnesses to different types of cancer), as well as for airborne diseases like Legionella [6] or Q fever [8]. As these examples show, for many types of diseases, finding the source of the outbreak can be seen as a spatial data clustering problem.

In general, clustering is the process of grouping objects into meaningful subclasses (that is, clusters) so that the objects within a cluster have high similarity in comparison to one another, but are dissimilar to objects in other clusters [9, 10, 11]. Spatial clustering deals with spatial objects, and disease clustering in particular deals with (geographically referenced) cases of a disease or another health-related condition. The problem of identifying the source of an outbreak can be seen as finding a location whose neighborhood has a disease rate that is *significantly* higher than for other locations. However, the problem studied here differs from the standard spatial clustering problem in two ways.

Firstly, typical clustering algorithms consider only the absolute number of objects within clusters (that is, the density of the cluster is defined as the case count), or, in the case of density-based clustering algorithms, the number of points per unit area [7, 1]. However, in disease clustering the number of cases is not very meaningful if it does not consider the population-at-risk. For example, considering only the absolute number of cases does not take into account that the population can be clustered itself within an urban area. To adjust the case count for the population density, we define the density of the cluster as the ratio of cases to exposed people, that is, we look for clusters with high density (or *attack rate*, in epidemiology).

Secondly, in the traditional clustering problem the exact location of the cases is known. This can be the case in disease clustering, as in the two examples mentioned before. Still, in many situations the precise case location is not available. On the one hand, it is common to have a data source that does not include this information. Statistics data is very often aggregated into areas corresponding to regions like counties, zip codes, census blocks or enumeration districts, or come from sources like anonymous questionnaires where only approximate locations (like partial zip codes) are provided in the first place. On the other hand, even if the data is available, there are privacy and confidentiality considerations for not disclosing exact address information of patients [5, 4]. Although this paper focuses on aggregated data in public health, it is worth mentioning that aggregated data is also frequently used in other areas such as criminology [17], sociology [19], political science [3, 12], and geography [15].

In the public health domain, aggregated data clustering is done by statistical methods. One of the most widely used approaches for cluster detection for disease surveillance is the *spatial scan statistic* of Kulldorff and Nagarwalla [14, 13].

But they represent the aggregation regions by points; thus, the spatial scan statistic does not directly handle aggregated data. Furthermore, the candidate cluster regions (windows) are positioned only at grid points (of a predetermined grid), which simplifies the problem. Another well-known method for cluster detection is by the Geographical Analysis Machine [16]. This method assumes non-aggregated point data and only tests cluster regions based on grid points as well.

This paper begins by modeling the problem as a rectangle placement problem (a rectangle is chosen for illustration; the ideas apply equally well to a square or regular 10-gon, for example, which allows us to approximate a circular cluster region). We first present a simple model that assumes uniform distribution of both the cases and the population, and then we provide a second model with a different density measure. Some possible extensions of these models are also discussed. We then present an algorithm for the first model. It computes a location for a cluster center by considering *all* possible placements of a rectangle R over a subdivision with n regions. This is an important difference to previous approaches, which restrict the search to a finite set of points. Our algorithm is based on computing the arrangement of the combinatorially different placements of R, and on optimizing a density function within each cell of the arrangement. The total worst-case running time of the algorithm is $O(n^2)$, but we prove that under reasonable, practical assumptions on the resolution of the regions and R, the running time is only $O(n \log n)$. The algorithm is flexible enough to allow extensions to several variations. After explaining the algorithm for the first model in detail, we discuss how to adapt it for the second model, and how to incorporate variants like different shapes for the cluster region, or having two different subdivisions for the case and population data. All these variations can be easily inserted into the algorithm for the first model, although sometimes at the expense of an increase in the running time.

2 Model

In this paper we abstract the problem of finding the source of a point source outbreak as a rectangle placement problem. In the models proposed next, we are given a subdivision of the plane, consisting of a set $\mathcal{P}$ of n regions $P_1, \ldots, P_n$, and for each region P_i in the subdivision we are given two values c_i and p_i. The first value c_i represents the number of disease cases within P_i, whereas the second value p_i represents the population of P_i (for example, the number of people at risk for the disease in question).

The outbreak area is modeled by a rectangle R, and the objective is to find a placement of R such that the *density* of cases covered by the rectangle is maximized. The density in R will be defined in the following models. We

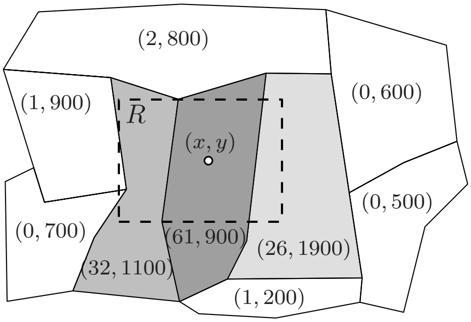


Fig. 1. Example for model I: for each region P_i, (c_i, p_i) is shown; shading visualizes density. The cases and population are assumed to be uniformly distributed inside each region. The goal is to place a rectangle R such that the density of R is maximized.

assume that we have access only to aggregated location data, meaning that the exact location of the cases is not known. The cluster rectangle R will have some fixed size and we will assume that it is axis-aligned.

We propose two basic models. The first model assumes that the distribution of the cases is *uniform* inside each region. The second model assumes a *worst-case* distribution of the cases, that is, all cases of regions intersected by R are assumed to appear inside the rectangle. Moreover, some possible variants, with different subdivisions for the case and population data, and different shapes for the outbreak area, are discussed at the end of this section.

Model I We will assume for the first model that the distributions of both the cases and the population are uniform. The density of the rectangle R is defined simply as the number of cases inside R divided by the total population in R.

More formally, let (x, y) be the position of the center of rectangle R. We will write $R(x, y)$ to refer to the rectangle R translated in such a way that its center lies at (x, y). See Figure 1 for an illustration. Finding a placement of R is equivalent to finding a value for x and y. The goal can be expressed as:

$$\max_{(x,y)\in\mathbb{R}^2} \frac{\displaystyle\sum_{i=1\ldots n} c_i \cdot f_i(x, y)}{\displaystyle\sum_{i=1\ldots n} p_i \cdot f_i(x, y)} \tag{1}$$

where $f_i(x, y) \in [0, 1]$ denotes the fraction of the area of P_i intersected by $R(x, y)$: $f_i(x, y) = Area(P_i \cap R(x, y))/Area(P_i)$.

This model seems reasonable given that we are dealing with aggregated data and the location of the cases is not known. However, there are situations in which the result obtained is not what one would expect. As an example, consider the example depicted in Figure 2. Under model I, the optimal rectangle

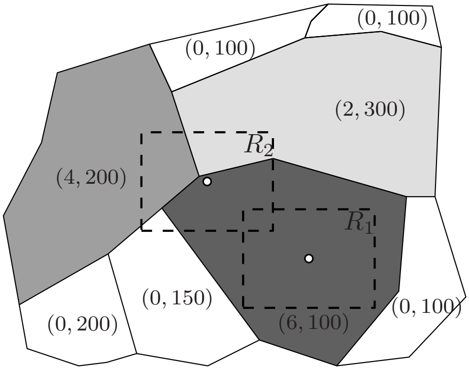


Fig. 2. The numbers in each region indicate number of cases and population. Model I will result in a rectangle like R_1, whereas model II will yield one like R_2.

lies completely inside the most dense region, like R_1. However, the situation suggests that the source of the outbreak must be close to the intersection of the three shaded regions. The next model we propose handles such situations better.

Model II As illustrated by Figure 2, the previous model does not always give the most reasonable result. In a situation like the one shown, if the real outbreak source is close to the meeting point of the three shaded regions, the assumption of uniform distribution of the cases within the regions is not valid.

To try to deal with this situation, we propose a second model where we assume that for a given location of R, all cases in each region intersected by R are concentrated inside R. This can be seen as a *worst-case* density measure for R, in accordance with the idea that most of the cases in a point source outbreak will be concentrated around the source. If the fraction of a region intersected by R is too small, we run the risk of counting more cases in a region than the number of people living in the intersection with R. To avoid this, we will take the minimum between the case count and the number of people assumed to live within that fraction of the region.

We formalize this model in the following way:

$$\max_{(x,y)\in\mathbb{R}^2} \frac{\sum\limits_{i=1\ldots n} \min\{c_i,\ p_i \cdot f_i(x,y)\}}{\sum\limits_{i=1\ldots n} p_i \cdot f_i(x,y)} \tag{2}$$

Two-subdivision variant It may be that the population information and the case information are aggregated differently. Then we would have one subdivision for the population data and another one for the number of cases. See Figure 3 for an example.

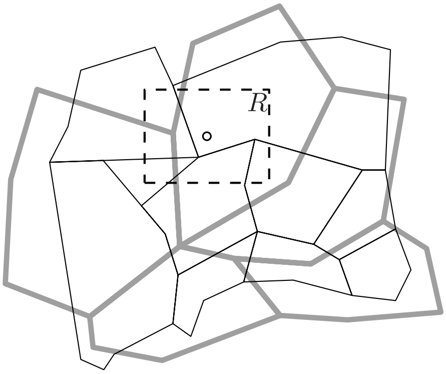


Fig. 3. A variant of the problem is where two different subdivisions are given, one for the case data and one for the population data.

We are now given a subdivision $\mathcal{P}$ of the plane comprised of n regions $P_1, \ldots, P_n$, and for each region P_i we are given a population value p_i, and in addition, we are given a second subdivision $\mathcal{C}$ comprised of m regions $C_1, \ldots, C_m$, each with a case count value c_i.

Both models I and II can be used for this variant. The equation for model I could be expressed as:

$$\max_{(x,y)\in\mathbb{R}^2} \frac{\sum\limits_{j=1\ldots m} c_j \cdot g_j(x,y)}{\sum\limits_{i=1\ldots n} p_i \cdot f_i(x,y)} \tag{3}$$

where $g_j(x,y) \in [0,1]$ denotes the fraction of the area of C_j intersected by $R(x,y)$, and $f_i(x,y) \in [0,1]$ denotes the fraction of the area of P_i intersected by $R(x,y)$. Note that in some unlikely cases, this measure can yield a value greater than 1, so it is not a real ‘density’ measure. Adaptations that address this issue are possible: the algorithm described in the next section is flexible enough to allow for a wide range of measures. In practice, however, the number of cases is expected to be much lower than the population count, and this should not be a problem.

For model II, the goal could be reformulated as:

$$\max_{(x,y)\in\mathbb{R}^2} \frac{\sum\limits_{j=1\ldots m} \min\{c_j, \sum\limits_{i=1\ldots n} p_i \cdot g_{ij}(x,y)\}}{\sum\limits_{i=1\ldots n} p_i \cdot f_i(x,y)} \tag{4}$$

where $g_{ij}(x,y) \in [0,1]$ denotes the fraction of the area of P_i intersected by $R(x,y)$ and C_j, and $f_i(x,y) \in [0,1]$ denotes the fraction of the area of P_i intersected by $R(x,y)$.

Different shapes for R In the previous model the outbreak area R was modeled with a rectangle (mainly because our algorithms are easier to describe in this case). However, depending on the characteristics of the disease under consideration, other shapes can be more appropriate. For example, for studying cases of exposure to some radiation source where the Euclidean metric applies, a disc can be better suited. For airborne diseases, when wind information is available, an ellipse with a certain rotation of the main axis can be a better choice. Both discs and ellipses (if approximated by a polygon) are variants of the outbreak area to which our algorithms can be extended, as discussed in the next section.

Resolution assumption In the analysis of the algorithm in the next section, we will not only consider the general worst-case scenario, but also a more realistic scenario. In particular, we will make a *resolution assumption*. Define the *resolution unit* r as the shortest distance between any two vertices of the region subdivision $\mathcal{P}$. Our *resolution assumption* states that there are positive constants c_1, c_2, c_3, c_4 such that (i) the distance between any vertex and any line segment not incident to that vertex is at least $c_1 r$, (ii) the length of any line segment in the subdivision is at most $c_2 r$, and (iii) the diameter of R is at least $c_3 r$ and at most $c_4 r$.

The assumption essentially states that the difference in scale between the regions, and between the rectangle R and the region subdivision are reasonable. For example, it would be very impractical to have regions that are city neighborhoods with an outbreak region of the size of the whole country. This assumption will allow to prove that in practice, the algorithms have a considerably better running time than what is provable otherwise.

Lemma 1. *The resolution assumption implies that any angle between two segments of $\mathcal{P}$ is bounded from below by a positive constant.*

Proof. Let v be a vertex of $\mathcal{P}$, and suppose there are two line segments with angle α that have v as an endpoint, and let the shorter have length l. Then the distance d between the endpoint of the shorter of the two and the longer segment will be $d = l \sin \alpha$. But we know that $l \leq c_2 r$ and $d \geq c_1 r$, which implies that $\sin \alpha \geq \frac{c_1}{c_2}$. Since c_1 and c_2 are positive constants, the lemma follows.

3 Algorithms

To solve the problems defined in the previous section, we will compute the arrangement of combinatorially different placements of the query rectangle R. The next subsection details what this arrangement looks like, and how to compute it efficiently. In the subsection that follows, we will use the arrangement to compute the optimal placement of the rectangle. Finally, we will describe how the method can be adapted to work for the other models.

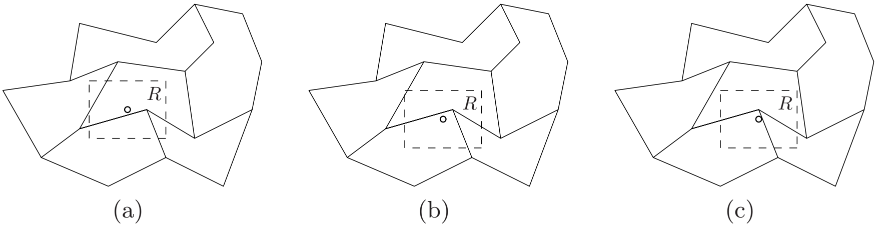


Fig. 4. Placements in (a) and (b) are are combinatorially the same, but in (c) it is combinatorially different: the set of regions intersected by R is different.

3.1 Arrangement of placements

Given a subdivision $\mathcal{P}$ and a rectangle R, we say two placements of R are *combinatorially different* if the set of pairs of edges of R and $\mathcal{P}$ that intersect are different. For example, the placements in Figures 4(a) and 4(b) are combinatorially the same, but the one in Figure 4(c) is different because the top left corner of R has moved from one cell to another. When two placements are the same, we can write the area of overlap between R and any cell $P \in \mathcal{P}$ as a closed-form function in x and y, which will allow us to optimize functions that involve this area of overlap efficiently for all combinatorially equal placements.

This combinatorial relation between placements subdivides the *placement space*—the set of possible positions for the reference point of R—into a number of regions such that inside each region, all placements are combinatorially equal. We can define and compute this arrangement for each cell $P \in \mathcal{P}$, and the total arrangement will just be the overlay of these. For a given cell P, a placement of R changes whenever a corner of R moves across an edge of P (as in Figure 4), or when a corner of P moves across an edge of R. This means the boundaries of the arrangement are exactly the line segments that arise when sliding a corner of R along an edge of P, as in Figure 5(b), or vice versa, as in Figure 5(c). The arrangement of this region is then the overlay of those two, as shown in Figure 5(d).

If we compute this arrangement for all cells of the subdivision, and compute the overlay of all of them, then this gives a new subdivision of the whole plane, such that within any cell the combinatorial structure does not change. To compute the arrangement, we collect all translated copies of the cells and of R, and note that their total number of vertices is $O(n)$ (as long as R has constant complexity). We can compute the overlay of all these polygons in $O(n \log n + k)$ time using standard methods, where k is the complexity of the final arrangement. In the worst case, this complexity can be $O(n^2)$. However, under the resolution assumption in Section 2, we can prove that the complexity is actually $O(n)$.

Lemma 2. *The complexity of the arrangement under the resolution assumption is $O(n)$.*

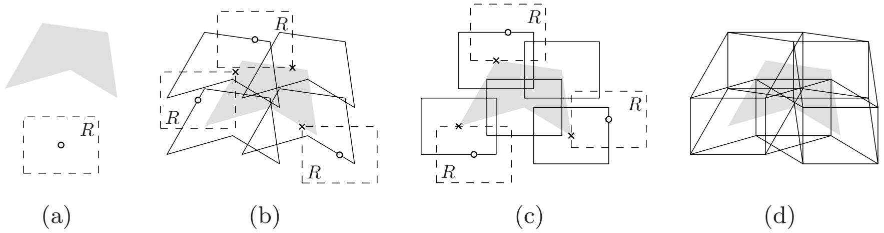


Fig. 5. (a) A cell of the subdivision (shaded), and a query rectangle R (dashed). (b) The positions of the reference point of R as a vertex of the rectangle slides along an edge of the cell. (c) The positions of the reference point of R as an edge of the rectangle slides along a vertex of the cell. (d) The total arrangement is the overlay of the previous two figures.

Proof. First, we show that R can never intersect more than a constant number of line segments of $\mathcal{P}$. Let V be the set of vertices inside R. We know that any two vertices are separated by at least $c_1 r$, and that the diameter of R is at most $c_4 r$. This means that the size of V can be at most $O((\frac{c_4}{c_1})^2)$ by a packing argument. Consider the set of line segments that intersect R, but do not have an endpoint in V. These segments must have a distance of at least $c_1 r$, and completely go through R, so there can be at most $O(\frac{c_4}{c_1})$ of them. By Lemma 1 a vertex $v \in V$ can be the endpoint of at most a constant number of line segments, so the total number of segments intersected by R is also constant.

Let p be any point in the plane. The rectangle R centered at p intersects at most a constant number s of features. This means p can be inside at most $O(s)$ different curves of the arrangement. Then we note that the regions in the arrangement corresponding to disjoint segments are pseudodiscs: the boundaries of two such regions cannot intersect more than twice. It is known that an arrangement of pseudodiscs with constant bounded depth has linear complexity [20].

3.2 Computing the optimal placement

To optimize Formula (1) over the arrangement, we first need to be able to optimize it inside a single cell. The only part of the formula that depends on x and y is the fraction $f_i(x, y)$ that describes which part of each region P in $\mathcal{P}$ is covered by R. The area of overlap can be decomposed into trapezoids, see Figure 6. The locations of the corners of these trapezoids are linear functions in x and y, so the area of each trapezoid is a quadratic function. We can then add up these functions, so that Formula (1) is in this form:

$$\max_{(x,y)\in\mathbb{R}^2} \frac{a_1x^2 + a_2xy + a_3y^2 + a_4x + a_5y + a_6}{b_1x^2 + b_2xy + b_3y^2 + b_4x + b_5y + b_6} \tag{5}$$

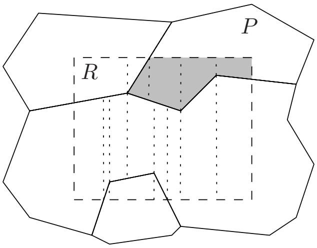


Fig. 6. The area of overlap between R and a subdivision cell P is the union of four trapezoids.

This formula can be optimized in constant time by using standard algebraic methods, or it can be numerically approximated very fast.

To find the maximum over the whole arrangement, we need to determine Formula (5) for every cell. We can of course just do this from scratch for each cell individually, but without the resolution assumption, that would require $O(n)$ time per cell, leading to a total of $O(nk)$ time (where k is the complexity of the arrangement). Instead, we will traverse the cells of the arrangement from neighbor to neighbor while maintaining some information. We maintain the numerator and the denominator of Formula (5) separately, and update them both when we move the reference point over the arrangement to a neighboring cell. Recall that the topological structure changes when a corner of R moves over an edge of P, or vice versa. Using some ideas from Reinbacher *et al.* [18], we can update the numerator and denominator in constant time when we move to a neighboring cell, basically by subtracting the contribution of quadratic functions that no longer give a trapezoid, and adding the contribution of quadratic functions that give a new trapezoid. Therefore, we spend only $O(k)$ time to determine Formula (5) for all cells, and to find the maximum. With the resolution assumption, this implies that we spend only $O(n)$ time in total.

3.3 Extensions

To solve the problem in the *second model,* our arrangements become a bit more complicated, because there is another event where the functions involved change: when the query rectangle starts containing enough area to allow all disease cases of some region to be inside it. This happens when the area of overlap between R and some region P_i becomes more than some fixed value: $Area(P_i) \cdot c_i/p_i$. This means we must add some extra curves to the placement space: the curves where the area of overlap has exactly this value. Generally this gives one closed curve (as R moves around P_i, keeping the area of overlap constant), but it could also be a collection of curves. Figure 7(a) shows how this looks in our example. In fact, the points where the pieces of this curve change coincide with the lines of the other parts of the arrangement, since this happens exactly when the combinatorial structure of the area of

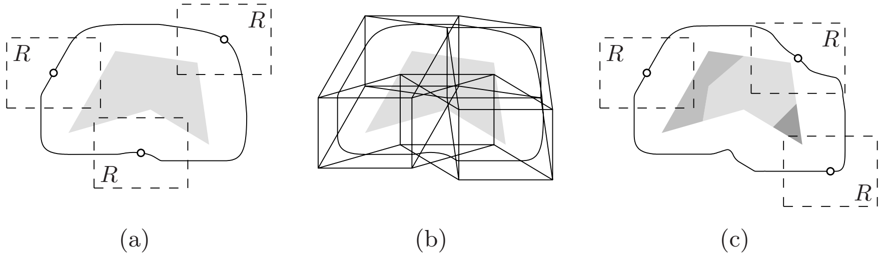


Fig. 7. (a) New curves introduced in the second model. (b) Total arrangement. (c) Non-uniform population density.

overlap changes. Within one piece, the curve is a level-set (or iso-contour) of a quadratic function, so it is a quadratic curve in the plane. Figure 7(b) shows the total arrangement. The functions we need to optimize over the new arrangement remain the same, only now we have to optimize them over cells with nonlinear boundaries. Standard numerical methods can be used to solve this problem.

When the information about the population and the disease cases are given in *separate subdivisions* $\mathcal{P}$ *and* $\mathcal{C}$ (in the first model), we can compute the overlay of the two and treat this as if it was a single subdivision. The algorithm still works without changes, but in the worst case the running time becomes as bad as $O(n^4)$. However, this will hardly occur in practice. In fact, under the resolution assumption in Section 2 we can prove that the complexity also stays linear.

Lemma 3. *In the two-subdivision variant of the problem, under the resolution assumption, the complexity of the arrangement is* $O(n)$*.*

Proof. Let l be a segment of $\mathcal{P}$. We will show that l intersects at most a constant number of segments of $\mathcal{C}$. This then implies that the overlay has linear complexity, and we can simply apply Lemma 2.

We know that the length of l is at most $c_2 r$. Sort the segments of $\mathcal{C}$ that intersect l. If two consecutive segments do not share an endpoint, then the distance between them is at least $c_1 r$. If they do share an endpoint, then this point must be at least $c_1 r$ away from l, and the angle between the segments, by Lemma 1, is at least $\arcsin \frac{c_1}{c_2}$, so the distance between them is at least some constant times r. Therefore the total number is constant.

When we have *separate subdivisions* for the population and the disease cases in the *second model,* we can still compute the overlay of the subdivisions, but now we need to be aware of the total number of disease cases in a certain *collection* of cells. Since we are assuming all cases inside a cell C are in the worst possible position, we cannot just distribute them evenly over the smaller cells in the overlaid arrangement. Instead, we compute the curves of constant overlap directly for C, while taking the finer population subdivision into account. Figure 7(c) shows an example of this situation. The curve is still piecewise quadratic, only the number of pieces now also depends on the number of smaller cells.

When our query region R is not a rectangle, but some other constant size polygon, the algorithm still works without any modifications. The number of vertices of R will appear in the running time, but not asymptotically if it remains constant.

4 Discussion

In this paper we apply computational geometry tools to solve certain disease cluster problems on aggregated data. We presented models and algorithms for finding the densest cluster in spatially aggregated data. It can be seen as an aid for finding a likely source of disease outbreaks. One model comes down to placing a rectangle such that the ratio between the cases contained within the rectangle and the population in it is maximized. The proposed algorithm solves the problem in $O(n^2)$ time, and under realistic input assumptions on the resolution of the input, in $O(n \log n)$ time. A second model uses a different assumption on the distribution of the cases within a region, and several variants (like different shapes for the cluster region or case/population data in different subdivisions) are also discussed, showing how the algorithm can be extended for those cases.

The problem addressed differs from the more traditional cluster location problems in that we do not work with the exact positions of the points but with aggregated data. As explained before, aggregated data is very often used in public health and other domains. Most of the spatial clustering algorithms do not take aggregation of the data into account. Our approach also differs from the more traditional approaches used for spatial disease clustering because we do not restrict the search of possible locations for the rectangle to a finite subset of points (like the centroids of the regions), but effectively consider all possible placements.

Several problems remain open and constitute interesting topics for further research. One of them is to incorporate more advanced density measures, to be able to use, for example, a likelihood ratio test like the one used by Kulldorff and Nagarwalla [14] for a suitable statistical model. The main difficulty lies in being able optimize such a function over a cell of the arrangement of different placements. Another interesting extension to consider is when the case data comes from different sources, for example emergency department visits and over-the-counter medication sales. Then the challenge would be not only to combine the different geographic subdivisions efficiently, but also to account for possible double-counting of the cases. Thirdly, cluster detection that includes the temporal component gives rise to new models of density where the time development of the cases in all regions may be known.

References

1. R. Agrawal, J. Gehrke, D. Gunopulus, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *Proc. ACM-SIGMOD Intl. Conf. on Mgmt. of Data*, pages 94–105, 1998.
2. H. Brody, M. R. Rip, P. Vinten-Johansen, N. Paneth, and S. Rachman. Map-making and myth-making in Broad Street: the London cholera epidemic, 1854. *The Lancet*, 356:64–68, 2000.
3. N. Cleave, P. Brown, and C. Payne. Methods for ecological inference: an evaluation. *Journal of the Royal Statistical Society, Series A*, 158:55–75, 1995.
4. L. H. Cox. Protecting confidentiality in small population health and environmental statistics. *Stat. Med.*, 15:1895–1905, 1996.
5. E. Cromley and S. McLafferty. *GIS and Public Health.* The Guilford Press, New York, 2002.
6. J. W. Den Boer, L. Verhoef, M. A. Bencini, J. P. Bruin, R. Jansen, and E. P. Yzerman. Outbreak detection and secondary prevention of legionnaires disease: A national approach. *International Journal of Hygiene and Environmental Health*, 210:1–7, 2007.
7. M. Ester, H. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. 2nd International Conference on Knowledge Discovery and Data Mining*, pages 226–231, 1996.
8. A. Gilsdorf, C. Kroh, S. Grimm, E. Jensen, C. Wagner-Wiening, and K. Alpers. Large Q fever outbreak due to sheep farming near residential areas. *Accepted for publication to Epidemiol. Infect.*, 2007.
9. J. Han and M. Kamber. *Data Mining: Concepts and Techniques.* Academic Press, San Diego, 2001.
10. J. Hartigan. *Clustering Algorithms.* John Wiley & Sons, New York, 1975.
11. A. Jain and R. Dubes. *Algorithms for Clustering Data.* Prentice Hall, Englewood Cliffs, New Jersey, 1988.
12. G. King. *A Solution to the Ecological Inference Problem.* Princeton University Press, Princeton, New Jersey, 1997.
13. M. Kulldorff. A spatial scan statistic. *Communications in Statistics: Theory and Methods*, 26,:1481–1496, 1997.
14. M. Kulldorff and N. Nagarwalla. Spatial disease clusters: detection and inference. *Stat. Med.*, 14:799–810, 1995.
15. S. Openshaw. *The Modifiable Areal Problem.* CATMOG No.38. Geo Books, Norwich, 1984.
16. S. Openshaw, M. Charlton, C. Wymer, and A. Craft. A Mark 1 Geographical Analysis Machine for the automated analysis of point data sets. *Int. J. Geographical Information Systems*, 1:335–358, 1987.
17. P. Phillips and I. Lee. Areal aggregated crime reasoning through density tracing. In *Proc. International Workshop on Spatial and Spatio-temporal Data Mining*, 2007.

18. I. Reinbacher, M. van Kreveld, and M. Benkert. Scale dependent definitions of gradient and aspect and their computation. In A. Riedl, W. Kainz, and G. A. Elmes, editors, *Proc. 12th Intern. Symp. Spatial Data Handling (SDH'06)*, pages 863–879, 2006.
19. W. Robinson. Ecological correlations and the behavior of individuals. *American Sociological Reviews*, 15:351–357, 1950.
20. M. Sharir. On k-sets in arrangements of curves and surfaces. *Discrete Comput. Geom.*, 6:593–613, 1991.
21. J. Snow. *On the Mode of Communication of Cholera*. Churchill Livingstone, London, 2nd edition, 1854.

Spatial Simulation of Agricultural Practices using a Robust Extension of Randomized Classification Tree Algorithms

J. Stéphane Bailly, Anne Biarnes, Philippe Lagacherie

UMR INRA-IRD-SupAgro LISAH, Campus de la Gaillarde, 2 place Viala,
34060 Montpellier (France)
e-mail: bailly@teledetection.fr, biarnes@supagro.inra.fr,
lagache@supagro.inra.fr

Abstract

In this paper, extensions of the classification tree algorithm and analysis for spatial data are proposed. These extensions focus on: (1) a robust manner to prune a classification tree to smooth sampling (e.g., spatial sampling effects), (2) an assessment of tree spatial prediction performances with respect to its ability to satisfactorily represent the actual spatial distribution of the variable of interest, and (3) a unified framework to aid in the interpretation of the classification tree results due to variable correlations. These methodological developments are studied on an agricultural practices classification problem at an agricultural plot scale, specifically, the weed control practices on vine plots over a 75 km² catchment in the South of France. The results show that, with these methodological developments, we obtain an explicit view of the uncertainty associated with the classification process through the simulation of the spatial distribution of agricultural practices. Such an approach may further facilitate the assessment of model sensitivities to categorical variable map uncertainties when using these maps as input data in environmental impact assessment modelling.

Keywords: CART, uncertainties, stochastic spatial simulation, robustness, predictors correlation

1. Introduction

Classification tree analysis (Breiman et al, 1984) is a very popular data mining tool that has been widely applied within the last 20 years in many different disciplines, including landscape ecology (McDonald and Urban, 2006, Fearer et al., 2007), soil science (Lagacherie and Holmes, 1997, Bui and Moran, 2001), agronomy (Tittonell et al, 2008, Gellrich et al., 2008), epidemiology (Schröder, 2006), and archaeology (Espa et al, 2007). The classification tree analysis is very popular since it accepts both categorical and continuous variables, it does not assume a model for relationship between variables (such as linear model) and it provides easily interpretable results.

The classification tree analysis, however, still has not completely solved problems that need to be addressed more thoroughly, especially when dealing with spatial data. As other non-parametric classification methods, CART is known to be sampling sensitive. This is especially true when (1) correlations between explanatory variables exist, which is often the case if these variables are derived from spatial datasets, and (2) spatial auto-correlation gives correlated individuals and redundancy in sampling. To overcome this sampling design problem, several derivative methods have been proposed (Breiman, 1996a; Breiman, 1996b; Geurts et al., 2006), and they are all based on aggregation of several classification trees with randomization. Unfortunately, if these extended methods smooth sampling effects, the advantages of CART interpretation are lost. Another well-known difficulty with CART is deciding how to prune the tree to limit overfitting. This is done in theory by evaluating each tree node from estimations of gains of purity performed from a set of so-called independent data, but, in practice, obtaining perfectly independent data is always difficult, especially with spatial data for which spatial correlation often exists. Furthermore, the evaluation of the spatial predictions provided by a classification tree can be considered as largely incomplete since it only evaluates the predictions at a local (individual) level and not with respect to its ability to predict realistic spatial structures of the variable of interest. Finally, it is noted that the use of CART for evaluating the explanatory power of candidate variables is very difficult because of the correlations that exist between these variables.

To overcome some of these problems, we propose in this paper an adaptation of the Classification Tree analysis for spatial data that have the following innovative characteristics:

- providing a robust and interpretable classification tree by selecting one tree within a collection of different trees (a forest); this selection results from a frequency analysis of the splitting nodes of the forest's trees,
- evaluating tree prediction performances with respect to its ability to satisfactorily represent the actual spatial distribution of the variable of interest,
- representing correlations of explanatory variables of different natures (categorical or continuous) in a unified framework to aid in the interpretation of the results.

This innovative classification tree analysis is illustrated and tested in a case study that aimed to model the spatial variations of weed control practices in a vine growing area located in the Peyne Valley (Hérault-South of France).

2. Methods

The methods we developed were performed in 3 steps. First, we developed a robust and randomized extension of the CART algorithm giving a robust tree. Secondly, a statistical distance that measures the performance of the tree to reproduce spatial patterns of classes was developed. Third, a synthetic correlation matrix that combines all the types of explanatory variables (i.e., predictors, the variable use to predict the class, categorical, or continuous) was proposed to aid in the interpretation of the tree.

2.1 Classification trees (CART)

The CART segmentation algorithm for classification is based on a recursive partitioning process of the multidimensional space defined by a set of k predictors, $X_{1, \ldots,}X_k$, in areas as homogeneous as possible regarding the variable, y that is being explained. y is a categorical variable having **C** modalities (1,...,c,...,C) called classes. The result is a binary hierarchical tree. The tree is characterised by several nodes $N_{j.}$ For each node N_j, the multidimensional space of the predictors is split into two subareas delineated by a value of a predictor (a threshold). The splitting rule for node N_j depends on an homogeneity measure regarding classes as, for example, the Gini index (Gini,1912), given by:

$$\text{Gini}(N_j) = 1 - \Sigma_c\, (\text{card}(y_{Nj}=c)^2/\text{card}(y_{Nj})^2). \qquad (1)$$

The tree determines a set of logical if-then conditions linking the classes to be explained to the predictors. Each terminal node of the tree, called a *leaf, contains a probability vector for each class, which sums to one. In a usual classification, the major classes are attributed to leaves (Breiman et al, 1984).*

2.2 Extension to select a robust tree within a forest

As in the bagging and random forest algorithms (Breiman, 1996a; Breiman, 1996b), we first built a "forest," i.e., a collection of **m** trees T1,…Ti,….Tm, using successive random resampling of calibration sampling sets. A calibration set is the set used to grown a tree.

The second step is a robust pruning algorithm applied over the forest of trees. Let's consider the set of p nodes forming a given tree Ti { $N_{i1,}$, $N_{i2,}$, …, $N_{ij,}$, N_{ip}} . Each node, $N_{ij,}$ is characterized by its location in the tree (its index j) and by its splitting pair (predictor, split value). At each successive location j, the principle is to select the splitting pair that exceeds a given frequency of occurrence f, calculated over the set { $N_{1j,}$, $N_{2j,}$, …, $N_{ij,}$, N_{mj} }of the nodes of the m trees having the same location. If the occurrence of several splitting pairs exceeds f, then the one having the larger f is selected.

The robust pruning algorithm is, therefore, the following :

1. Start from the **m** root nodes { $N_{11,}$, $N_{21,}$, …, $N_{i1,}$, N_{m1} }.
2. Perform a frequency analysis of the occurrence of the splitting pairs of nodes N_{i1}: if there exists a pair for at least **f*m** trees of the forest, then this node with this splitting pair is kept at location 1 (the root). If a pair does not exist, $\mathbf{N_{i1}}$ is rejected. If several pairs satisfy this criterion, the most frequent pair is kept.
3. If $\mathbf{N_{i1}}$ is kept, only the subforest having $\mathbf{N_{i1}}$ with the selected splitting pair is considered further .
4. Steps 1 and 2 are run for the following locations (2, …j,…p) until any more splitting pair is selected.

At the end of this algorithm, a single pruned tree, **Tf,** is obtained, which can be interpreted regarding splitting value and predictor pairs for each node.

Each leaf of the tree has a probability vector of dimension C that gives the occurrence of each modalities of the variable y.

Assessing the spatial pattern classes simulation resulting from robust trees

Let us assume that the spatial data that we handle and the categorical spatial field that we want to predict are in a spatial domain that can be divided into G regular cells (1,...,g,...,G) with resolution r and orientation α defined as the angle between grid axis and longitude or latitude. To assess how the obtained spatial pattern using Tf fits the observed pattern, we developed the following tools:

1. a stochastic use of the tree that, in a spatial context, both simulates a spatial distribution of classes and accounts for uncertainties in prediction,
2. a statistical distance that quantifies the dissimilarities between the observed and the simulated spatial patterns of the above defined cells, and
3. an empirical test on the spatial pattern dissimilarity differences for various robust trees (for instance, various robust trees defined for various frequency parameters **f).**

In the first step, a set of n possible spatial distributions of the variable of interest y is derived from the tree Tf. Each spatial distribution is defined by randomly allocating the individuals of each leaf of the tree to a modality c of y with respect to the probability vector of the leaf. Repeating this process n times gives n possible spatial distributions of y.

In the second step, a value for the dissimilarity between the n simulated and the observed spatial distributions of y is first computed for each cell g, and it is denoted by $\mathbf{d(y,X_{(n)})_g}$ **.**

We let $\mathbf{y = [y_1, ...,y_C]}$ be the vector that computes the percentages of observed individuals for each one of the C classes on cell g**.** In the same way, for the simulation i on cell g, we introduce the vector $\mathbf{X_i}$ **(i=1,...,n)** that computes the percentages of individuals for each one of the C classes: $\mathbf{X_i=[X_{1i}, ...,X_{ci}]}$. When concatenating the n simulations (i.e., the n vectors $\mathbf{X_i}$ **(i=1,...,n))**, we obtain the **n*C** matrix $\mathbf{X_{(n)}}$.

In classification problems, the Cohen's kappa coefficient (Cohen, 1960) is usually used as a robust distance between classes resulting from CART and observed classes in a confusion matrix. The tree validation objective is quite different here since we want to compare a distribution of simulated classifications aggregated on spatial units.

To compare a value to a distribution, it is common practice to use normalized Euclidean distances or methods that gives a score when the value falls into confidence intervals for a given probability level (Goovaerts, 2001). Due to correlation in $\mathbf{[X_{1i}, ...,X_{Ci}]}$ (the vector sums to one), we preferred to compute the dissimilarity between **y** and $\mathbf{X_{(n)}}$ for each cell using the Mahalanobis distance (Mahalanobis, 1936) given by:

$$d(y,X_{(n)})_g= [\ (y-\mu)^t\ .\ \Sigma\ .\ (y-\mu)\ ]^{0.5} \qquad (2)$$

with : $\boldsymbol{\mu} = [\boldsymbol{\mu_1, ...,\mu_C}]$, mean of $\mathbf{X_{(n)}}$ on g
and $\boldsymbol{\Sigma} =$ covariance matrix of $\mathbf{X_{(n)}}$ on g.
$\mathbf{(y-\mu)^t}$ is the transpose of $\mathbf{(y-\mu)}$

Finally, a global dissimilarity **D** over the entire domain is computed using a weighted average of the dissimilarities computed for each cell:

$$D= \ \Sigma_g\ w_g\ \ d(y,X_{(n)})_g, \qquad (3)$$

with $\mathbf{w_g}$ denoting the weight for the cell g, which is equal to the ratio between the individual counts in g and the total domain individual count.

Finally, we computed a distribution for the global dissimilarity D by repeating the process described above for the **n** spatial distributions of y simulated by the tree **Tf**. Therefore, when empirically testing the significance in the difference of spatial pattern dissimilarities for various robust trees **Tf** (e.g., obtained with different predictors) having the same parameters **r** and **α,** we can analyze the obtained global dissimilarity distributions (analysis of variance).

2.4 Visualization of correlations between various predictor types for tree interpretation

To represent correlations between predictors of different types (categorical or continuous), we developed a unified framework to aid in the interpretation of the tree results. This framework is a statistics matrix related to correlation computation for three different cases:

- the classical determination coefficient $\mathbf{R^2}$ when crossing two numerical variables $\mathbf{X_1}$ and $\mathbf{X_2}$,
- the Cramer statistic (Aaron et al., 1998), resulting from the chi-square test when crossing two qualitative variables $\mathbf{X_1}$ and $\mathbf{X_2}$ such that

$$V^2 = \chi^2/\min(p-1,\ q-1), \qquad (3)$$

with : $\boldsymbol{\chi^2}$ **:** chi-square value from contingency table
and **p, q :** number of rows, columns in contingency table

- the η^2 statistic (Haase,1983), resulting from the ANOVA sum of variances when crossing a qualitative variable and a numerical one:

$$\eta^2 = (\text{within groups variance})/\ (\text{total variance}) \qquad (4)$$

All of these statistics can be interpreted in the same way; values close to zero indicate independent variables, and values close to one indicate correlated variables.

All these methods were developed on R 2.6.0 statistical software (Ihaka and Gentleman, 1996) using the tree package (Ripley, 2007).

3. CASE STUDY

In this study, vine plots over a spatial domain corresponding to a catchment are the statistical individuals on which we want to predict a categorical variable, the weed control practice (WCP), from readily available explanatory variables or predictors.

3.1 Study site

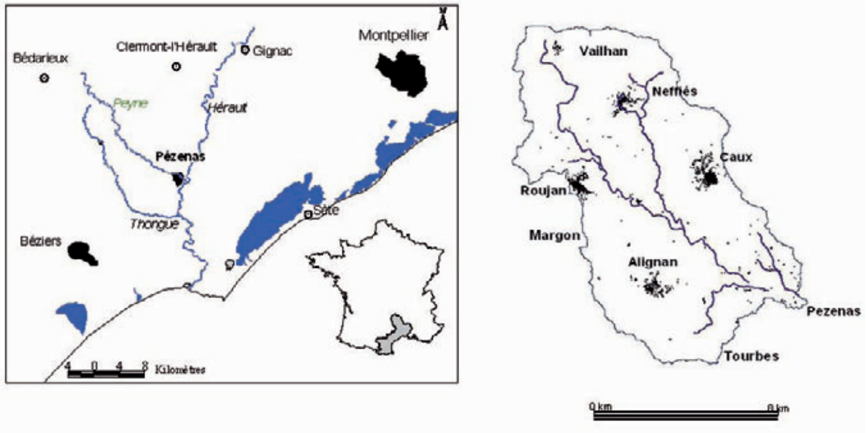


Fig. 1. Location of the study area

The studied spatial domain was the Peyne river catchment (75 km²) located in the mid Hérault valley in Languedoc-Roussillon, one of the world's largest wine-producing regions (Figure 1). This catchment suffers from serious herbicide pollution of the surface water. Studies on the pollution process show that this pollution should be in relationship with the high risk of herbicide leaching by run-off during the heavy rainfall events that are typical of the area's sub-humid Mediterranean climate (Lennartz et al., 1997; Louchard et al., 2001). They assign a crucial role to the vineyard weed control practices (WCP), which determine both the type and amount of herbicide applied and the evolution of soil surface characteristics that affect the soil’s hydraulic conductivity (Leonard and Andrieux, 1998; Hebrard et al., 2006).

The Peyne catchment incorporates all or part of the territories of eight local government areas (LGA, in France referred to as “communes”), and it is farmed by 650 winegrowers whose majority supply LGA-based cooperative wineries.

3.2 Data

A geographical database was developed that included (a) information regarding WCP and (b) a description of physical or socio-economic variables that can potentially explain the practices. The database contains a sample of 1007 geo-referenced vine plots of land, owned by 63 winegrowers and corresponding to 989 ha, i.e., about 20% of the area under vines within the Peyne valley.

Sampling scheme and data collection

The required data were gathered by surveying 63 winegrowers, selected by sampling plots along five transects perpendicular to the Peyne River. It was assumed that such a sampling adequately represented farm holdings and that the whole set of vine-plots of these farm holdings was representative of the diversity and distribution of WCP in the valley.

The survey questionnaire focused on (1) the weed control practices used in each plot and (2) the variables assumed to explain the choice of practices. In addition, the plots were precisely located on the land register map and the 1:100,000 soil map by Bonfils (1993).

Weed control practices

From the collected data, a 4-type expert-based stratification of the various WCP was performed regarding (1) their potential impact on surface runoff and (2) their intensity of herbicide use. Practice *Pa* is based on chemical weeding. In practice *Pb,* tillage alternates with chemical weed control in alleys. In practice *Pc*, the alleys are repeatedly shallow-tilled. In practice *Pd,* tillage alternates with alleys under permanent grass.

From *Pa* to *Pd* (*Pa* > *Pb* >*Pc* > *Pd*), the environmental risk is less and less strong because of a reduction in herbicide amounts (from total to partial chemical weed control) and because of the use of weed control methods that reduce more and more surface runoff.

The survey results are resumed in table 1. They show that the most common practice over the entire catchment is *Pc.* When looking on the sampled practice spatial distribution in figure 2, is clear that the WCP are not randomly distributed in space. *Pc* is the dominant practice on the left side of the Peyne River (on the right of the figure), whereas *Pb* and *Pd* are dominant on the right side of the river.

Table 1. Percentage of the different weed control strategies in the plots sample

Weed con-trol practice	Plots concerned		Land area con-		Farm holding con-	
	Number	%	Ha	%	Number	%
Pa	170	17	139	14	34	54
Pb	197	20	189	19	22	35
Pc	505	50	486	49	48	76
Pd	135	13	175	18	14	22
Total	1007	100	989	100	63	100

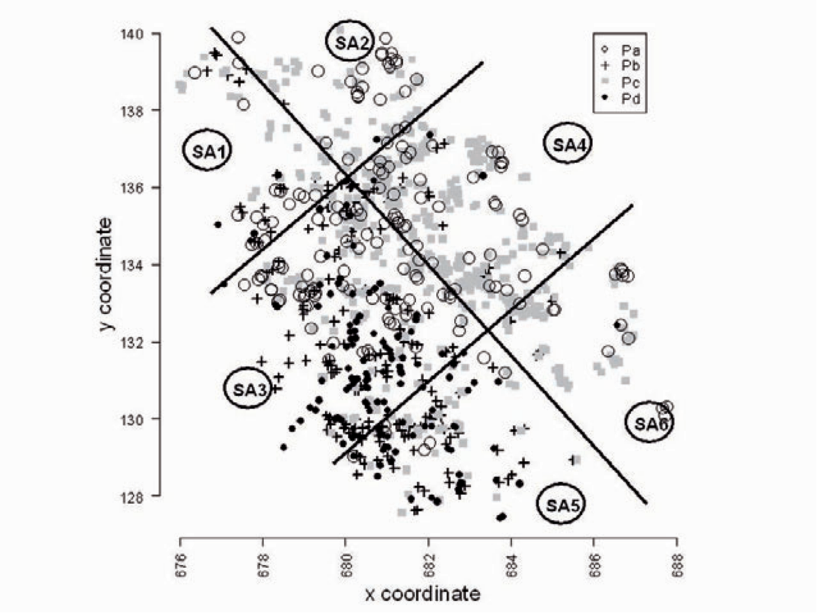


Fig. 2. Spatial distribution of the observed practices (with division of the Peyne valley into six sub-areas (SA))

Explanatory variables

To generalize the prediction of WCP from explanatory variables (predictors) throughout the Peyne catchment, we tested only variables that can be easily collected from maps, remote sensing, or geo-referenced databases.

Table 2: Explanatory variables

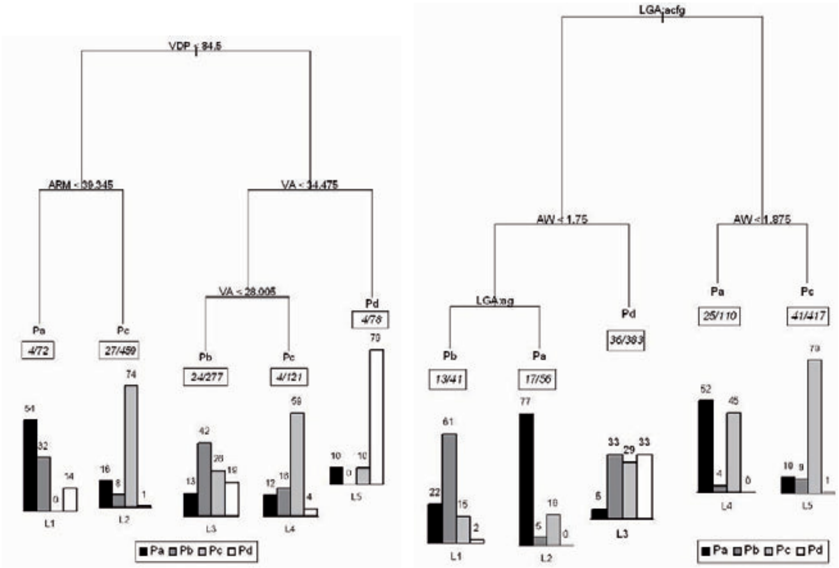
Group of variable	Variable	Modalities or range of variation
1: Characteristics of the plots	AW: alley width	1.5 to 3 m
	SOIL	1: Soil on quartzic substratum; 2: Leptic calcosol on terraced hillsides; 3: Colluvio soils on glacis; 4: Colluviosol with redoxic layer in depressions and fluviosol on alluvial flats; 5: Brunisols and fersialsols on plateau; 6: Luvic brunisols on alluvial terraces
2: Characteristics of the farm holdings	ARM : percentage of area under aromatic varieties	0 to 100 %
	VDP : percentage of wine production under Vin de Pays	0 to 100 %
	VA : vineyard area	0.3 to 62 ha
	ACT: activity	Full time, Part time, Retired
	WIN: principal winery	Cooperative winery; Private winery
3: Socio-professional environment	LGA: local government area	Alignan (a), Caux (b), Margon (c) Neffiès (d), Pezenas (e), Roujan (f), Tourbes (g), Vailhan (h)

Eight explanatory variables (predictors) were collected (table 2). These variables belong to three groups corresponding to three hypothesized levels of spatial organization of practices diversity: set 1, the physical characteristics of the plots (two variables), set 2, the structural characteristics and production priorities of the farm holdings (five variables), and set 3, the local government area (LGA) to which the plots belong. The choice of these three groups of variables was governed by the literature and the results of a previous study conducted in two of the eight LGAs of the valley (Biarnès et al. 2004).

4. Results

4.1 Robust weed control practices classification trees

In a first step, robust trees were performed on two sets of explanatory variables coming from the two main sources of available data (agricultural census or geographical databases). The tree T1 was performed on the set 2 of predictors (characteristics of the farm holdings) with a frequency parameter **f** = 95 %. T2 was performed on the sets 1 and 3 of predictors (physical characteristics and socio-professional environment) with a frequency parameter **f** = 80 %. These parameters were selected to get at least four farms in each leaf. These trees are plotted in figure 3.



Tree 1 : T1

Tree 2 : T2

Fig. 3. Presentation of the selected robust trees. Each terminal node is associated (1) to a majority practice, (2) to an inset giving the number of farm holdings and the number of plots concerned, and (3) to a lay of distribution of the four modalities of WCP (% of plots).

When comparing the trees, T1 gives the most contrasted WCP distributions (probabilities vectors) between leaves. It also presents, however, only strongly discriminant variables as shown by tree branch lengths that are related to the discriminating power of the splitting variables. Only three variables (VDP, VA, ARM) related to the economic scale and the productive choices of the farm holdings are necessary to differentiate distributions of practices. The right branch of the tree is associated with VDP oriented farm holdings (percentage of wine under VDP greater than 84 % of the total production). In these holdings, choices of practices vary according to the vine area. As a trend, winegrowers adopt practices that limit more and more polluting runoff (*Pb*, then *Pc*, then *Pd*) when increasing the vine area. In the farm holdings characterized by a weaker production of VDP (left branch of the tree), choices of weed control practices are linked to the percentage of area under aromatic varieties (ARM).

For the T2 tree, the root node is split on the values of the LGA variable, dividing the set of plots into two groups respectively located in the LGAs on the left side (right branch of the tree) and on the right side (left branch of the tree) of the Peyne River. The plots characterized by narrow alleys (AW less than 1.75 m or than 1.875 m according to the river side), are mainly associated with intensive use of herbicide (*Pa*, and even *Pb*). On the left side, the plots with wide alleys are mainly associated with practice *Pc*, whereas they are equally associated to practices *Pb, Pc*, or *Pd* on the right side of the tree.

4.2 Assessment of spatial patterns resulting from robust tree predictions

In a second step, we used the robust calibrated trees T1 and T2 to simulate spatial distributions of WCP on the set of sampled vine plots.

Figure 4 shows three examples of simulated WCP spatial distributions: a totally random spatial simulation respecting only global practices percentages in the top-left, a spatial simulation using tree T1 in the top-right, and a spatial simulation using tree T2 in the bottom-left. This figure shows that a random spatial distribution looks highly dissimilar to the observed one (Figure 2). Conversely, the stochastic use of trees gives contrasted and realistic distributions between the two sides of the Peyne River.

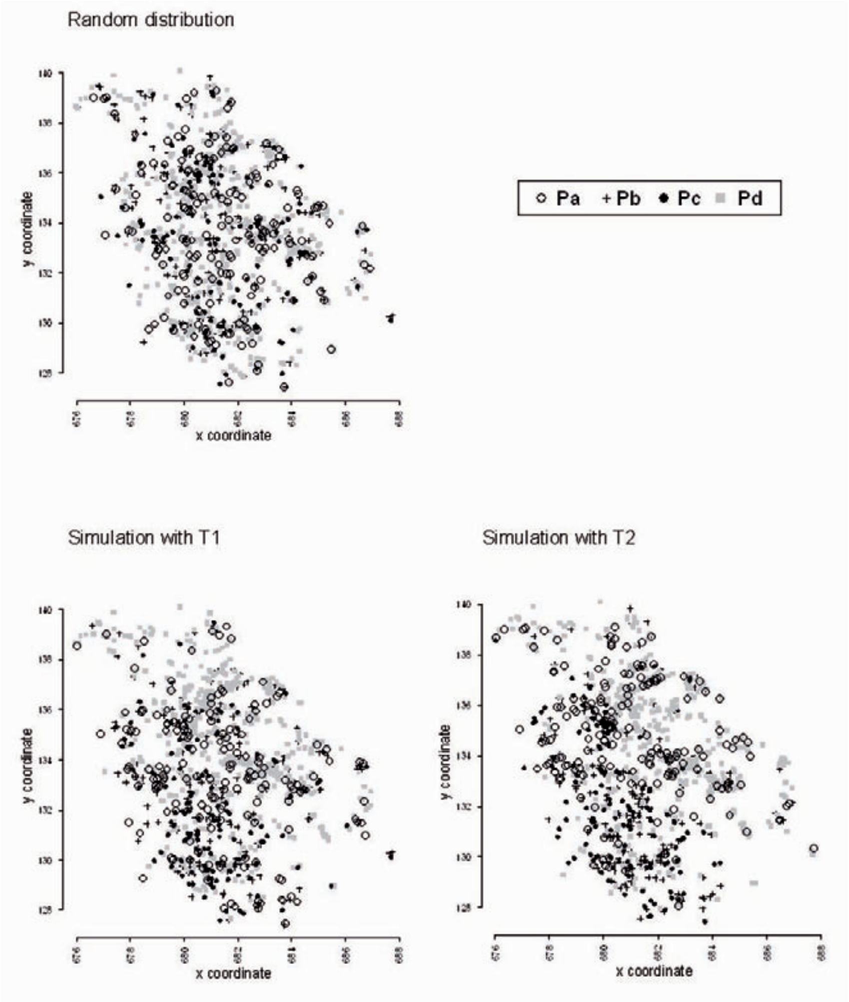


Fig. 4. Simulations of spatial distribution of practices

Global dissimilarities D between the observed WCP distributions and the n=1000 simulations were computed as explained in methods dividing the spatial domain into 6 cells (**r**=5 km, α=45°). These parameters were mainly chosen considering the two sides of the Peyne River. Computations were performed using successively trees T1, T2 and a totally random spatial simulation.

To obtain a distribution for these dissimilarities, this D computation was repeated 50 times, which yielded the statistics shown in table 3. Finally, the results show that simulations resulting from T2 are the closest to the observed distribution even though T1 was the best based on a classical statistical criteria (with purest leaves).

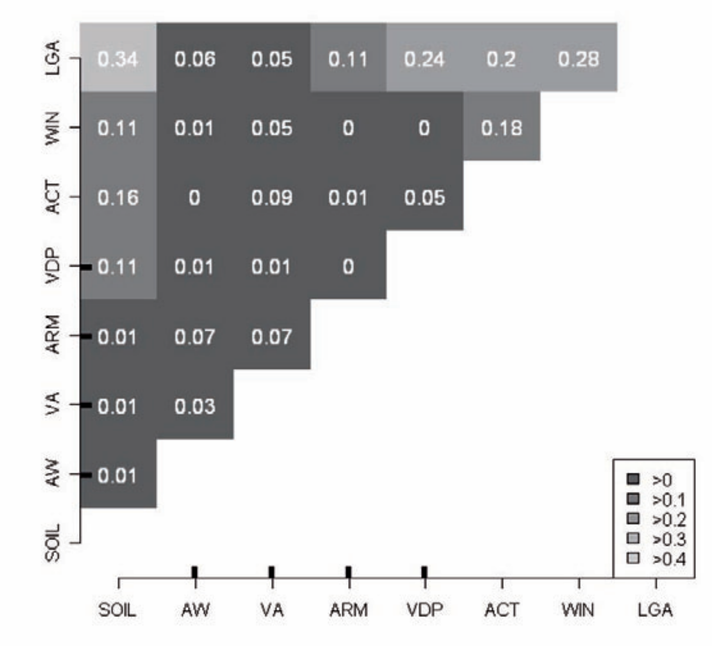
Table 3. Global dissimilarity distribution D between observed WCP spatial distribution and simulated ones

Dissimilarity statistics	Aleatory distribution	Tree 1	Tree 2
Mean	14.92	6.16	2.60
Quantile 98 %	15.27	6.23	2.64

4.3 Interpreting tree results visualizing predictors correlation

To better understand the latter result and the difference and similarities in trees T1 and T2, a square correlation matrix between all predictors was computed, as shown in table 4.

Table 4. Matrix of correlation statistics between variables. The four numerical variables have wide tickmarks on axes.



	SOIL	AW	VA	ARM	VDP	ACT	WIN	LGA
LGA	0.34	0.06	0.05	0.11	0.24	0.2	0.28	
WIN	0.11	0.01	0.05	0	0	0.18		
ACT	0.16	0	0.09	0.01	0.05			
VDP	0.11	0.01	0.01	0				
ARM	0.01	0.07	0.07					
VA	0.01	0.03						
AW	0.01							
SOIL								

This table shows that the three groups of explanatory variables that we used are not entirely independent, and it highlights the role of the LGA explanatory variable within the T2 tree. The plot variable SOIL and the holding variables VDP, WIN, and ACT are not randomly distributed among LGAs. Thus, LGA seemed to integrate various driving factors of practices, which explains its relevance in discriminating the practices:

- In the case of VDP, this may be explained by the characteristics of the wine production. This production is strongly governed by national regulations delimiting specific geographic areas for producing AOC or VDP wines (an AOC area is localized on three LGAs territories in the north-west of the Peyne Valley) and by the production orientations of the LGA-based wineries. Consequently, the LGAs of the Peyne catchment correspond to particular productive orientations of the farm holdings, which might explain the different practice distributions between LGAs.
- Correlation between SOIL and LGA comes from distinctive soil types over the Peyne Valley; LGA located on the right river side of the catchment have very clayey surface soils that almost do not exist in the LGA located on the other river side. These soils are part of the soils on the plateau described in table 2. They particularly justify the use of practices *Pb* or *Pd* due to the high risks of not having bearing capacity after a heavy rainfall event. The small area concerned by these soils does not justify by itself the extent of these practices, but studies on farm management showed an endeavour to simplify the work by limiting the range of different practices used (Aubry, 1998). A practice selected to resolve a particular problem in a particular plot may be used in other plots, which is the case with practices *Pb* and *Pd.*
- In addition, a leading role in the diffusion of practices by farm information networks has been shown by sociologists (Chiffoleau, 2005). The role of such networks and their links with LGA are being studied in two LGAs of the Peyne catchment by sociologists. Initial results showed that some of the winegrowers' information networks (proximity networks, technical advice networks) might depend on the LGA where they are living and explain the differences in practices between LGAs. In particular, the use of practices *Pb* and *Pd* by winegrowers who do not have any plots in the very clayey soil area might be linked to the LGA-based proximity network to which these winegrowers belong.

5. Conclusion

In this paper, we presented and demonstrated the interest of an adaptation of a classification tree analysis that addresses current problems that are often encountered with this method, especially when dealing with spatial data. We obtained robust trees that smooth the sampling sensitivity while keeping trees that look simple and that are easily interpretable. A specific evaluation procedure that considered the local spatial structures of the variable of interest was demonstrated to be useful when deciding between the candidate trees. Finally, the formulation of the correlations between the different variables helped to depict the complex role of a global variable (LGA) for explaining the observed spatial structure of agricultural practices.

Other problems with the application of classification trees to spatial data will need to be addressed in the future. One suggestion is to be more explicit about taking into account the autocorrelations of spatial data in tree-building algorithms as proposed by Bel et al (2005). Another suggestion is to consider the role of the uncertainty that affects both the explanatory variable and the variable to be explained, as pointed out by Lagacherie and Holmes (1997).

In this study, however, we developed a stochastic method that not only allows for simulation of the spatial distribution of the categorical variable in space but also gives a fully explicit view of the uncertainty associated with the classification process through the simulation of the spatial distribution of classes. Such an approach may further facilitate the assessment of model sensitivities to categorical variable map uncertainties when using these maps as input data in environmental impact assessment modelling.

References

Aaron, B., Kromrey, J. D., & Ferron, J. M. (1998). Equating r-based and d-based effect-size indices: Problems with a commonly recommended formula. *Annual meeting of the Florida Educational Research Association*, Orlando, FL., ERIC Document Reproduction Service No. ED433353.

Aubry C., Papy F. and Capillon A. (1998). Modelling decision-making process for annual crop management. *Agricultural Systems* 56(1): 45-65.

Bel L., Laurent J.M. , Bar-Hen A., Allard D. and Cheddadi R. (2005). A spatial extension of CART: application to classification of ecological data, *Geostatistics for environmental applications*, Springer: Heidelberg, 99-109.

Biarnès A., Rio P. and Hocheux A. (2004). Analysing the determinants of spatial distribution of weed contol practices in a Languedoc vineyard catchment. *Agronomie,* 24: 187-191.

Bonfils P. (1993). Carte pédologique de France au 1/100°000 ; feuille de Lodève, SESCPF INRA.

Breiman L., Friedman J. H., Olshen R. A. and Stone C. J. (1984). Classification and Regression Tree. London, Chapman and Hall, 358 p.

Breiman L. (1996a). Bagging predictors. *Machine learning* 26(2): 123-140.

Breiman L. (1996b). Random forest. *Machine learning* 45: 5-32.

Bui, E., and Moran, C. (2001). Disaggregation of polygons of superficial geology and soil maps using spatial modelling and legacy data. Geoderma, 103: 79-94.

Chiffoleau Y. (2005). Learning about innovation through networks: the development of environment-friendly viticulture. *Technovation* 25(10): 1193-1204.

Cohen J. (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* 20: 37–46.

Espa, G., Benedetti, R., De Meo, A., U. Ricci and Espa S. (2006).GIS based models and estimation methods for the probability of archaeological site location, *Journal of Cultural Heritage*, 7(3), July,147-155

Fearer, T.M., Prisley, S.P., Stauffer; D.F., and Keyser P.D (2007). A method for integrating the Breeding Bird Survey and Forest Inventory and analysis databases to evaluate forest bird–habitat relationships at multiple spatial scales. *Forest Ecology and Management*, 243(1), 128-143

Gellrich, M., Baur, P., Robinson, B.H. and Bebi P. (2008). Combining classification tree analyses with interviews to study why sub-alpine grasslands sometimes revert to forest: A case study from the Swiss Alps, *Agricultural Systems*, 96(1-3), 124-138

Geurts P., Ernst D. and Wehenkel L. (2006). Extremely randomized trees. *Machine learning* 63: 3 - 42.

Gini C. (1912). Variabilità e mutabilità. Memorie di metodologica statistica. Vol. 1, E. Pizetti and T. Salvemini. Rome, Libreria Eredi Virgilio Veschi, pp 211-382

Goovaerts P. (2001). Geostatistical modelling of uncertainty in soil science. *Geoderma* 103: 3-26.

Haase, R.C. (1983). Classical and Partial Eta Square in Multifactor ANOVA Designs. Educational and Psychological Measurement, 43(1), 35-39.

Hébrard O., Voltz M., Andrieux P. and Moussa R. (2006). Spatio-temporal distribution of soil surface moisture in a heterogeneously farmed Mediterranean catchment. *Journal of Hydrology* 329: 110-121.

Ihaka R. and Gentleman R. (1996). R: A Language for Data Analysis and Graphics,. *Journal of Computational and Graphical Statistics* 5(3): 299-314.

Lagacherie, P. and Holmes, S. (1997). Addressing geographical data errors in a classification tree for soil unit prediction. *Int. J. Geographical Info. Sci.* 11, pp. 183–198

Lennartz B., Louchard X., Voltz M. and Andrieux P. (1997). Diuron and simazine losses to runoff water in mediterranean vineyards. *Journal of Environmental Quality* 26(6): 1493-1502.

Leonard J. and Andrieux P. (1998). Infiltration characteristics of soils in Mediterranean vineyards in Southern France. *Catena* 32: 209-223.

Louchart X., Voltz M., Andrieux P. and Moussa R. (2001). Herbicide Transport to Surface Waters at Field and Watershed Scales in a Mediterranean Vineyard Area. *Journal of Environmental Quality* 30: 982-991.

McDonald, R.I., Urban, D.L. (2006). Spatially varying rules of landscape change: lessons from a case study, *Landscape and Urban Planning*, 74(1), 7-20.

Mahalanobis P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Science of India* 12: 49-55.

Ripley B. (2007). Pattern recognition and neural Networks. Cambrige, Cambridge University Press, 415 p.

Schröder W., (2006). GIS, geostatistics, metadata banking, and tree-based models for data analysis and mapping in environmental monitoring and epidemiology, *International Journal of Medical Microbiology*, Volume 296(1-22), 23-36.

Tittonell, P., Shepherd, K.D. , Vanlauwe, B. and Giller, K.E. (2008) Unravelling the effects of soil and crop management on maize productivity in smallholder agricultural systems of western Kenya—An application of classification and regression tree analysis, *Agriculture, Ecosystems & Environment*, 123(1-3), 137-150.

Impact of a Change of Support on the Assessment of Biodiversity with Shannon Entropy

Didier Josselin[1], Ilene Mahfoud [1], Bruno Fady[2]

[1] UMR 6012 ESPACE CNRS – Université d’Avignon
74, rue Louis Pasteur 84029 Avignon cedex 1, France
Email: didier.josselin@univ-avignon.fr, mahfoud_ilene@yahoo.fr

[2] INRA, UR629, Ecologie des Forêts Méditerranéennes (URFM)
Site Agroparc, Domaine Saint Paul, 84914 AVIGNON cedex 9, France

Abstract

The research deals with the Modifiable Areal Unit Problem (MAUP). The MAUP is a common scale effect in geostatistics relating to how a studied territory is partitioned and to the ecological fallacy problem due to spatial data aggregation. We processed a biodiversity assessment using the Shannon index on a set of remote sensing data (SPOT 5) on the Ventoux Mount (Southern France). We applied the calculation on different geographical areas, with different sizes, shapes and spatial resolutions to test the effect of support change on the biodiversity measures. We proposed a method to aggregate the data at several imbricated scales so that the loss of biodiversity due to the spatial autocorrelation can be estimated separately from the MAUP. The concept of ‘pertinent’ scale is then discussed through two biodiversity criteria, a quantitative one (the Normalized Difference Vegetation Index, which evaluates the biomass quantity) and a qualitative one (a species typology, coming from a supervised classification of remote sensing data and experts maps).

Keywords: Modifiable Unit Problem, biodiversity, pertinent scale, remote sensing data, Shannon entropy

1 Introduction

Our paper deals with the Modifiable Areal Unit Problem, also called the MAUP. This problem results from the combination of two joined processes which are well known in spatial analysis and geostatistics. For example, let us consider we are geographers wishing to measure the density of inhabitants in a set of towns. We first have to identify the limits of towns and then to count how many people live in each area. Finally, for each town that is to say for each set of grouped people, we have to process a simple ratio between this number and the town surface. At first sight, it may seem that this density assessment provides correct estimates due to the fact that the counting is related to the surface area. But in fact, it is difficult to compare small towns with large ones, because the more inhabitants there are in the area, the more accurate the density evaluation is. In other words, we can have a reasonable confidence in density measures for large towns, while the reliability of the same estimate is rather uncertain for small towns. This is a first facet of the MAUP, related to zoning or spatial partitioning dimension, also called ‘scale effect’. Moreover, if we divide a large town into different sectors, process density evaluation in the same way, and then apply the means of the densities, we surely will not obtain the same density as the one calculated for the whole town. This shows that this statistic reflects the way the territory is partitioned and the handled spatial entities influence statistical accuracy. This leads to the second aspect of the MAUP: the ecological fallacy problem. This problem is related to aggregation process that ineluctably provides a decrease in variance due to smoothing induced by grouping more and more individuals within the same sample. Both these two factors are correlated, because when we enlarge the zones, we increase the probability to include more people in towns, or more generally, more individuals in geostatistical samples. Although these effects are common to any measurement, they are also linked to the statistics we use and its robustness faced to a given sample size and the distribution shape of observed data.

In this research, we tackle the MAUP in its two main aspects: the spatial and the aggregation dimensions. We aim at evaluating the impact of the change of spatial support on biodiversity assessment processed with Shannon entropy. We propose indeed a method of making more relevant geostatistical measures, by trying to measure and adjust values to the MAUP effect. After a concise state of the art about the MAUP, we present a case study in the South of France (Ventoux Mount) and we develop a methodology to assess biodiversity using remote sensing data. The results are discussed to lead to a conclusion about the relevance of the of ‘pertinent scale’ concept.

2. The Modifiable Unit Problem

It seems that the first observation of the MAUP was stated in 1934 by Gehkle and Biehl, who noticed a relationship between correlation coefficients and spatial levels of the considered data. This was then confirmed by Yule and Kendall in 1950. These authors studied potato and wheat yield correlation for 48 English regions and demonstrated that correlations increase with scale. Robinson (1950) showed that the rank of correlations is directly linked to the size of the spatial unit. This obviously illustrated the dangers of providing individual statistics from data analyzed at an aggregated level. Numerous authors confirmed this problem (Rastetter *et al.*, 1992; Reynolds, 1998, Dusek, 2005). Openshaw made a very useful synthesis on the MAUP in 1984. More recent research defined an enlarged concept of Change Of Support Problem (COSP) (Crawford &Young, 2004) including MAUP and ecological inference problem.

Practically, most of the statistics offices of local authorities dealing with data on urban or natural areas have to handle heterogeneous entities in terms of shape and size such as counties, municipalities, environmental sites, and so on. This does not prevent them from publishing statistics even if those are marred by mistakes. As Openshaw (1984) suggested, one can deal with the MAUP in different ways. A first one is to reject any statistical information because of the MAUP. Another way is to ignore it totally. This emphasizes the difficulty to provide reliable statistics on any studied territory without any risk.

Another point of view is to take the MAUP into account and adjust for it. This obviously means to aim at reducing its impact on measurement. One way to measure the MAUP might be by comparing spatial entities with identical shapes, areas and equal number of individuals. This would imply analyzing the territorial data at rigorously identical scales. Though that case occurs very rarely due to data availability, we can notice that this approach does not avoid the MAUP in the sense that it still affects the different areas, of which we cannot assess the level accurately.

Another way of process consists in finding the best spatial partitioning or the best correspondence between the number of individuals to be grouped and the way to group them. This can be done using an arbitrary zoning process, designed by different statistical clues generally based on variance and heterogeneity. Openshaw (*op. cit.*) clearly showed that any of these statistics aiming at constituting a partition failed to solve the MAUP, and, a potentially more dangerous consequence, produced rather different thematic statistics for the same quantities of zones. This means that the criteria used to make the partition have a great effect on the results through the partition change itself. Openshaw also proposed an iterative methodology for spatial study processing an automatic zoning: ‘the idea is to use the optimal zoning approach to test hypothesis by manipulating the aggregation process’.

As far as we know, the MAUP has not been solved, neither practically, nor theoretically. It seems indeed to be a very arduous task to measure explicitly and independently the MAUP effect involving provided statistical data, while setting aside the ecological fallacy problem and the statistical robustness. It is due to the fact that observed spatial structures wear spatial autocorrelation[1](Cliff & Ord, 1973) which is itself very sensitive to the aggregation level and the considered scale (Cressie, 1993). Let us mention some research that has been developed to assess the part of aggregation effect on statistical estimates, using notably mathematical simulations (Amrhein, 1995, Reynolds, 1998). Here, we also use geocomputation for trying to eliminate, if it is possible, the MAUP from the biodiversity assessment.

The MAUP has been observed and studied on different kinds of data, such as areas, located data or remote sensing data (Marceau *et al*, 1994, Gao & Tueller, 1997), in different fields, such as landscape ecology (Jelinski & Wu, 1996) to assess landscape dynamics, biodiversity and global change.

As we shall see in the next sections, our research concerns the field of landscape ecology and required to manipulate numerous different remote sensing data, at different scales and resolutions (Baldwin *et al.*, 2004) from a multiscale point of view (Lovett *et al.*, 2006). As geographers and ecologists, we aim at understanding the spatial structure of natural vegetation. That is to say we try to catch the exact part of diversity carried by forest species and biomass spatial organization. To provide reliable measures of biodiversity, we propose a method to prevent the bias due to the MAUP and we analyze different cases in which spatial support may have some influence on biodiversity evaluation through MAUP. Size, shape and the initial resolution of the image are obviously the tested criteria. Thus, our research has two goals: a thematic one (how to efficiently assess the ‘real’ biodiversity of natural structures) and a second one (how to eliminate the

[1] In our study, *spatial autocorrelation* is considered as important when a group of contiguous pixels have the same attribute value.

MAUP from spatial measures) after having identified some of its structural factors (if any).

3. Data and biodiversity indexes applied on the Ventoux Mount, Vaucluse, Southern France

3.1. The Ventoux Mount

The Ventoux Mount is located in the South East of France, at the South-western most tip of the Alps (fig. 1). It is a Biosphere Reserve due to its location, its isolation among large plains in the Rhone valley, and because it is a remarkable habitat containing rare and endemic species: as it is at the crossroads between European and Mediterranean bioclimatic and ecological influences.

A part of our study, which is not described here, involves a spatial classification of the local species encountered on the Ventoux Mount. We designed maps using landscape surveys and numerical supervised and non-supervised classifications based on remote sensing data. This information has been used to comprehend the results of this research (Mahfoud *et al.*, 2006).

Let us list now the different elements required for our study: image characteristics and indexes used for our biodiversity estimation.

Fig 1. The Ventoux Mount with its scrub land.

3.2. Shannon entropy to assess biodiversity

Shannon's entropy, coming from information theory, requires data grouped into classes within a statistical distribution (Atlan, 2006). Shannon entropy is indeed often used in landscape ecology by biologists to describe how equitably a set of species or land uses is arranged and to measure the taxonomic richness (Farina, 2000). Shannon diversity index (H') [1] is minimal if a species is dominant, maximal when all the species are uniformly distributed. The more equitable the distribution is, the higher the index (Frontier, 1983) is. The Shannon diversity is known to be sensitive to the number of individuals and to rare events in a distribution. To evaluate how regular the distribution is, there also exists a complementary entropy index: the Shannon evenness index, which corresponds to entropy divided by the *log* of the number of classes. Our example involves the common Shannon diversity, defined by:

$$H' = -\sum_{i=1}^{m} p_i * \ln(p_i) \quad [1]$$

with p_i the frequency of the attribute i in the whole set and m the number of possible attributes.

In our study, we apply entropy to a biomass index and a classification of species processed on groups of pixels from remote sensing data.

3.3. A set of complementary images

The source image is a part of a SPOT 5 panchromatic multispectral image covering the ‘Ventoux Mount’ with a most frequent resolution of 2.5 m. The Shannon index is then calculated using a set of different images, whose goal is to evaluate the impact of the change of support on biodiversity measures (figures 2 & 3):

- images with identical extent, but taken at different initial resolution levels (image Z1 at 2.5, 5 and 10 meters of resolution)
- images of different shapes (T2 and T2M) ;
- images of different sizes (T1=Z1, T2 and T3 made by an homothetic transformation of T1) ;
- images from three different locations (Z1, Z2 and Z3) with different spatial structures and species.

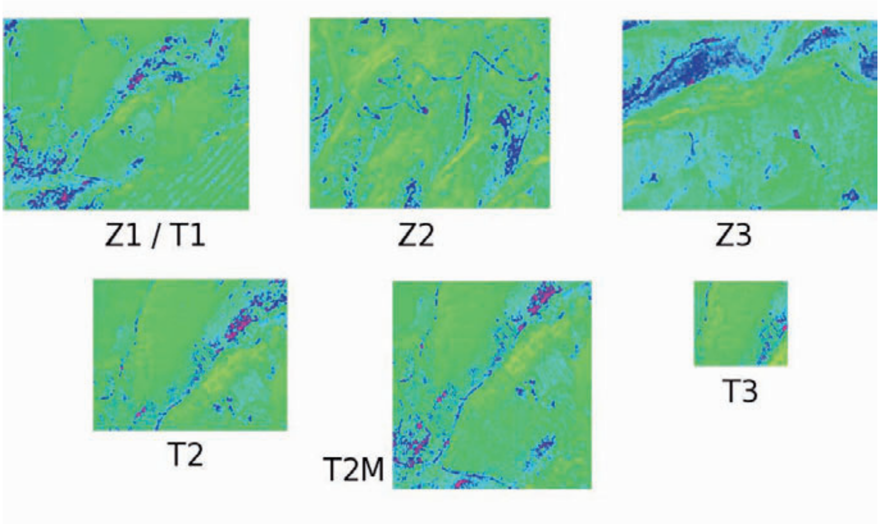


Fig. 2. The set of images for the Normalized Difference Vegetation Index (values from 0 [dark] to 255 [light]); Z1, Z2 or Z3 have the same size and shape on different locations (*location effect*); the ratio width/height is constant for T1, T2 and T3 all representing territories close to each others (*size effect*); T2M is a square in the same area and has the border same (longer) size as T2 (*shape effect*).

The software used for processing the images is the Landscape Ecology component from GRASS.

3.4. Normalized Difference Vegetation Index (NDVI) and species typology

We tested our methodology on two different attributes. One is the Normalized Difference Vegetation Index (NDVI, figure 2), which allows to make an analysis of the biomass diversity (Tucker, 1979). It is a quantitative attribute, computed as follows:

$$NDVI = (B3 - B2) / (B3 + B2) \qquad [2]$$

where B2 is the value from the red channel sensor and B3 the one from the close infrared.

The statistical distribution of pixels is built according to the number of encountered integer values. The value can go as high as 365 due to coding (there can be at most one class per integer) and this induces a possible relationship between the size of the considered area and the level of diversity: the more numerous the pixels are, the higher the probability is to have different classes of pixels and subsequent high biodiversity, due to the sensitivity of the Shannon entropy. One already can see the manifestation of the MAUP here, considering its statistical aspect.

We first processed a classification of the species on the Ventoux Mount in 10 different classes. This classification has been used for biodiversity assessment, using the same Z1, Z2, Z3, T2, T3, T2M areas and the three basic resolutions (figure 3).

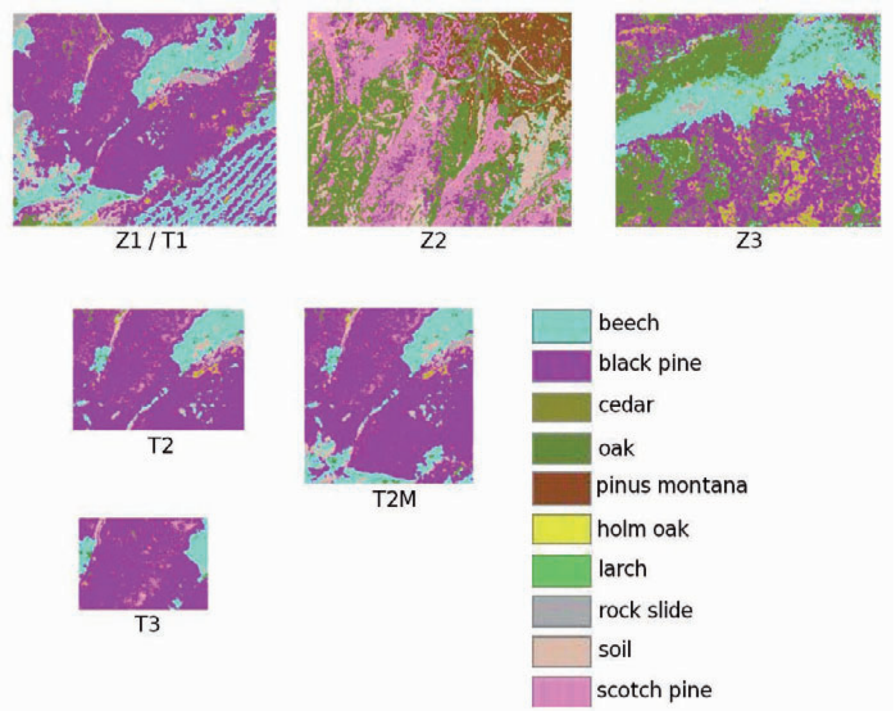


Fig 3. The set of images describing the main species and the land uses (10 classes); Z1, Z2 or Z3 have the same size and shape on different locations (*location effect*); the ratio width/height is constant for T1, T2 and T3 all representing territories close to each others (*size effect*); T2M is a square in the same location and has the same (longer) border size as T2 (*shape effect*).

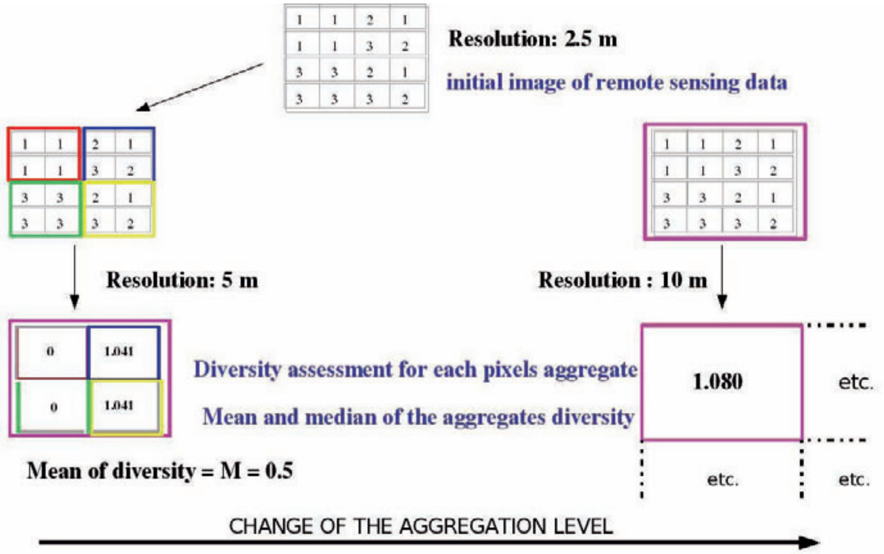
This attribute differs from the previous one in the sense that it is qualitative (nominal) so we can have a maximum of 10 classes. Besaides as we can see on figure 3, the Z1, Z2 and Z3 images have been chosen to be quite different in terms of species and levels of spatial autocorrelation. The two tested attributes are complementary and their respective diversity measures will be compared in the section showing the results.

4. A way to evaluate and to ‘prevent’ the MAUP from biodiversity assessments

4.1. Pixel aggregation methodology

For both attributes (NDVI and species classification) and for all the images, we applied a common methodology aiming at evaluating the MAUP effect and at avoiding it in the biodiversity measure. The process was based on two principles: (i) the support is changed gradually by pixel aggregation using a grid with increasing cell sizes (all cells are imbricated through scales); (ii) at each step, biodiversity is computed for each pixel aggregate (*i.e.* each cell of a given grid) and aggregate (*i.e.* cells) averaged and median diversity values are calculated for the whole image. Figure 4 depicts the process step by step.

We finally obtained a global value of diversity for different imbricated scales on the same image with its specific characteristics (size, shape, resolution, area) and a diversity criteria (NDVI or species classification). Then we are able to figure the MAUP effect.



1	1	2	1
1	1	3	2
3	3	2	1
3	3	3	2

Resolution: 2.5 m
initial image of remote sensing data
Resolution: 5 m
Resolution : 10 m
0
1.041
0
1.041
Diversity assessment for each pixels aggregate
Mean and median of the aggregates diversity
1.080
etc.
Mean of diversity = M = 0.5
etc.
etc.
CHANGE OF THE AGGREGATION LEVEL

Fig. 4. Process of pixel aggregation; example with an image of 2.5 m of initial resolution: we overlay a set of grids whose cells (*i.e.* aggregates) become larger through scales; at each aggregation level (*i.e.* ‘scale’), we process the diversity calculation for each cell (using the pixels values) and the mean (or median) of the aggregates diversities of the whole image.

4.2. Using resampling to avoid the MAUP in statistical estimates

Our methodology aims at extracting the ‘real’ diversity vale of a spatial sructure and it is based on the assumption that an diversity estimate results from three dimensions of spatial autocorrelation:

i. ‘real’ diversity using NDVI or species spatial distribution;
ii. diversity due to aggregation level (scale, ecological fallacy problem);
iii. diversity linked to the area considered for calculation (shape, size, etc.).

Our method to separate the diversity due to aggregation level from that concerning the observed attribute consisted in resampling observed pixels randomly in the same territory (same number of pixels and size of image). We then applied the same aggregation method described in 3.1 on this new image. This makes it possible to extract, by a simple difference of Shannon entropy at each scale, the loss of ‘real’ diversity carried by each territory (figure 5) and from the one due to spatial autocorrelation. Indeed, the only biodiversity estimated in such a random image is the one due to MAUP and spatial support, although the observed data image includes diversity due both to support and observation. By measuring their difference, we express the loss of diversity resulting from spatial autocorrelation. A word of caution however: the measure does not identify the intrinsic diversity of a territory according to a given attribute; in fact, it computes the loss (%) of diversity due to the spatial structure. On the opposite, the higher this clue is, the lower (bio)diversity is.

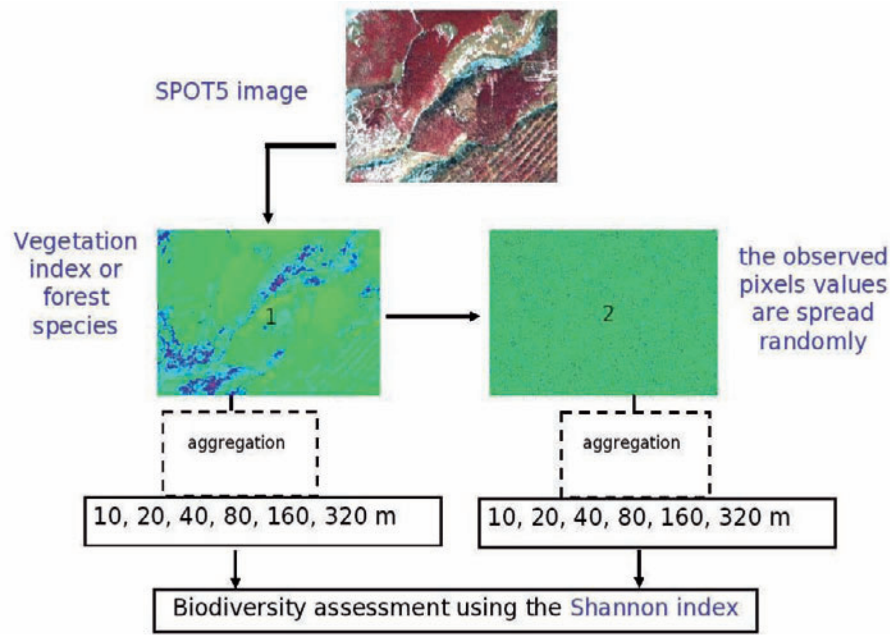

SPOT5 image
Vegetation index or forest species
1
2
the observed pixels values are spread randomly
aggregation
aggregation
10, 20, 40, 80, 160, 320 m
10, 20, 40, 80, 160, 320 m
Biodiversity assessment using the Shannon index

Fig. 5. Pixel random spreading process

Here we present results using a single random image, but they are quite similar whatever the random image is. If biodiversity estimates can be a little more accurate by processing many samples and calculating their average, it is however sufficient to use a single random image to identify the MAUP effect and to assess the real biodiversity through scales. The difference between these two kinds of diversity losses (observed *versus* random diversity) is firstly transformed into a relative measure by dividing it by the random diversity and then plotted onto a graph. This graphic representation can be used in different ways: to evaluate the part of diversity due to MAUP according to a given scale, the increase or decrease in the MAUP impact while scale is growing, or the identification of a ‘pertinent scale’ (figure 6). The relative loss of ‘real biodiversity’ is provided as a percentage of the whole diversity including MAUP.

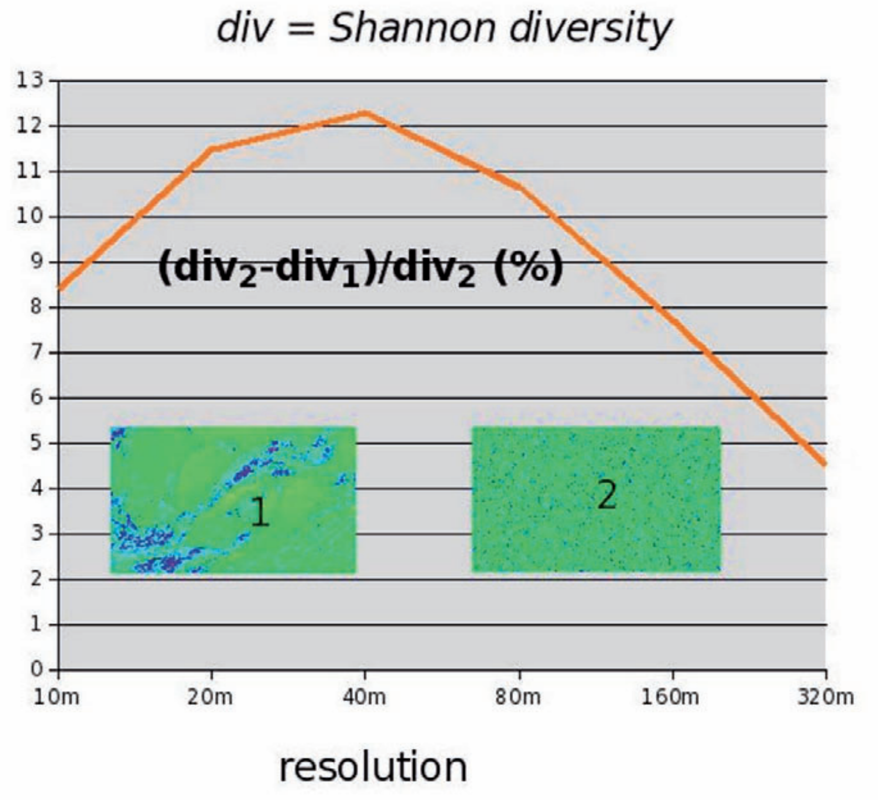


Fig. 6. Relative Shannon biodiversity loss difference (%) through the scales

5. Results

5.1. Global impact of the MAUP on biodiversity assessment

First of all let us notice that for any image (including random ones) and any tested attribute, the average values of aggregate biodiversity increased dramatically. This is the MAUP 'signature'. Some previous research about the MAUP showed that it is very frequent using different biodiversity indexes and most times much more important than one could have expected. It is an euphemism to say that spatial structure is thus not the only factor of biodiversity. In some cases, the level of aggregation explains itself up to 90% of biodiversity estimate! The MAUP effect must lead biologists to use Shannon entropy and other indexes with a lot of care. Another induced effect is visible on the distribution of pixel values: whatever the distribution shape, all data move to higher values when scale increases. That explains color change on figure 7.

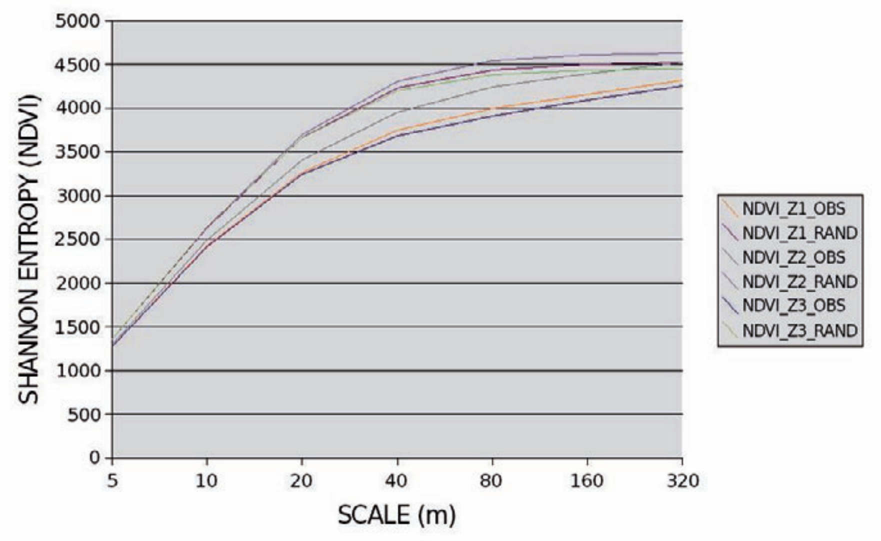


Fig.7. The Shannon entropy (X 1000) value ineluctably increases with scale growth *(OBS=observed data; RAND=randomized data)*

5.2. Do image size and shapes have an influence on the measurement of biodiversity?

All the following images show the relative loss of biodiversity as it is explained in section 3.3., figure 6.

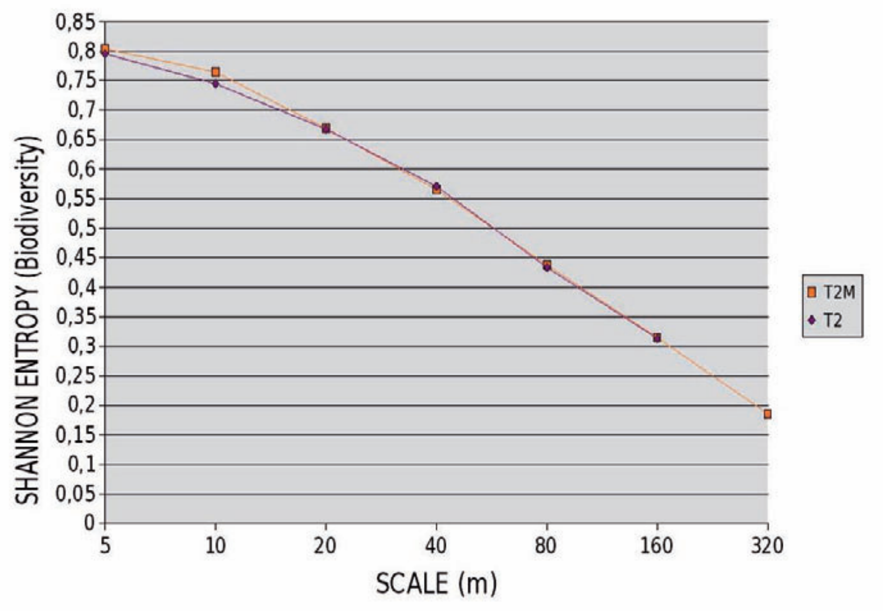


Fig. 8. Relative loss of diversity (%) through scales (m) for biodiversity images with different shapes (T2 and T2M)

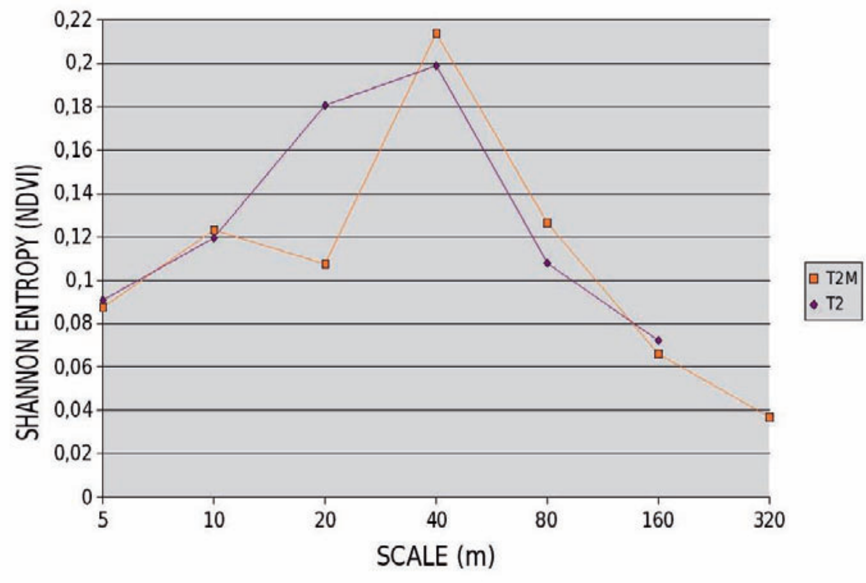


Fig. 9. Relative loss of diversity (%) through scales (m) for NDVI images having different shapes (T2 and T2M)

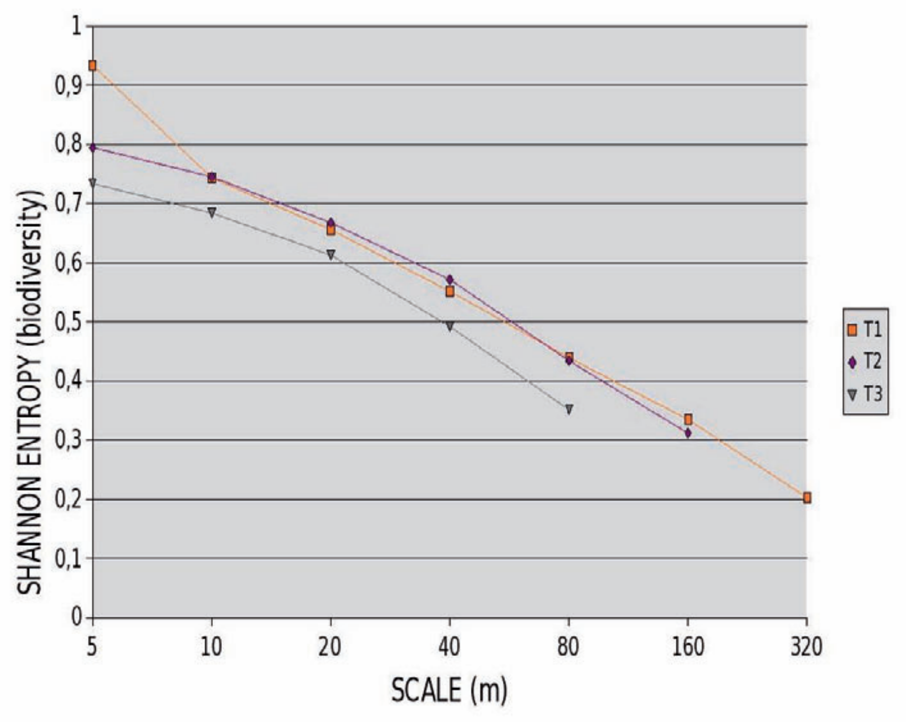


Fig. 10. Relative loss of diversity (%) through scales (m) for biodiversity images with different sizes (T1, T2 & T3)

Changing the size in proportional X and Y dimensions or the shape of the images amounts to the same issue: does the support structure significantly affect biodiversisty assessments? It seems that the answer is no. Indeed, if we look at figures 8 to 11, we can notice different elements. For a given attribute (NDVI *vs* species biodiversity), there are quasi-similar curves in terms of forms and values (compare figures 8 & 10 and then 9 & 11). Some possible edge effects that would influence the measure do not seem to occur. The images are large enough to show the same curves in the graphs. The fact that we find the same relative entropy values means that MAUP and support effects do not change significantly among this subset of images.

Another interesting point is the important difference of the assessed loss of diversity. A comparison between figures 8 & 9 or 10 & 11 shows that the biodiversity loss for the species classification is much more important than the NDVI one. This is due, on the one hand, to the real observed diversity and the kind of data (qualitative *vs* quantitative). The probability to get the same contiguous pixels is higher for species classification than for NDVI. For that attribute, let us recall that any integer pixel value can constitute a single class, allowing a high diversity within large pixel aggregates. On the other hand, it seems that there is a peak of biodiversity loss in the middle scales (20 to 80 m). This shows the specificity of each attribute and spatial structure influence, probably corresponding to the maximum spatial autocorrelation. The scale growth also induces a continuous fall of diversity loss (from about 100 % to 20 %) for species diversity.

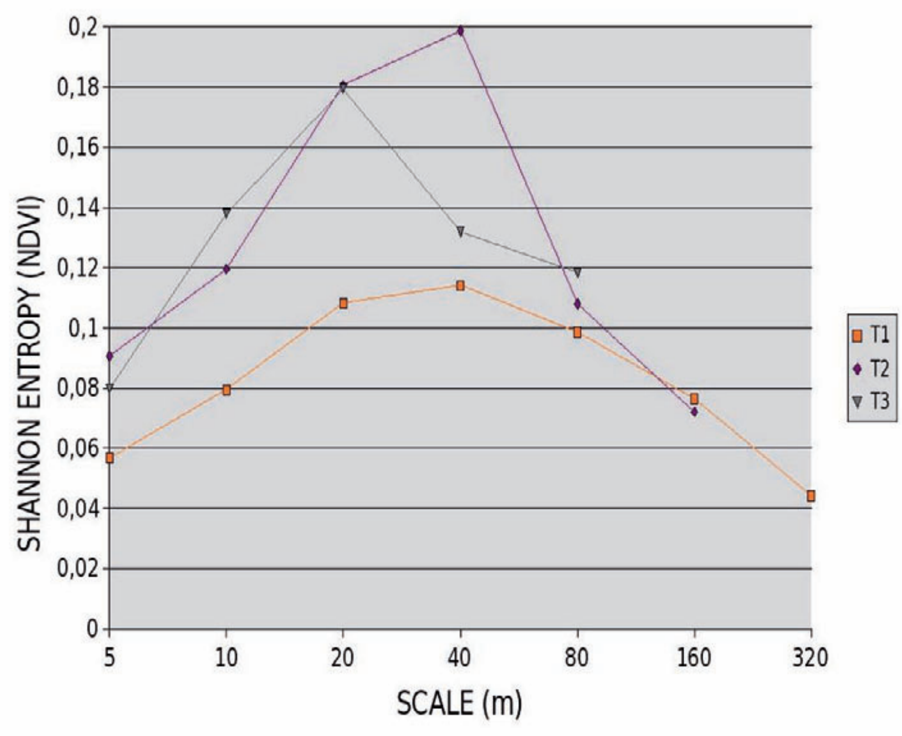


Fig. 11. Relative loss of diversity (%) through scales (m) for NDVI images with different sizes (T1, T2 & T3)

5.3. Do different areas with identical supports have very different biodiversities?

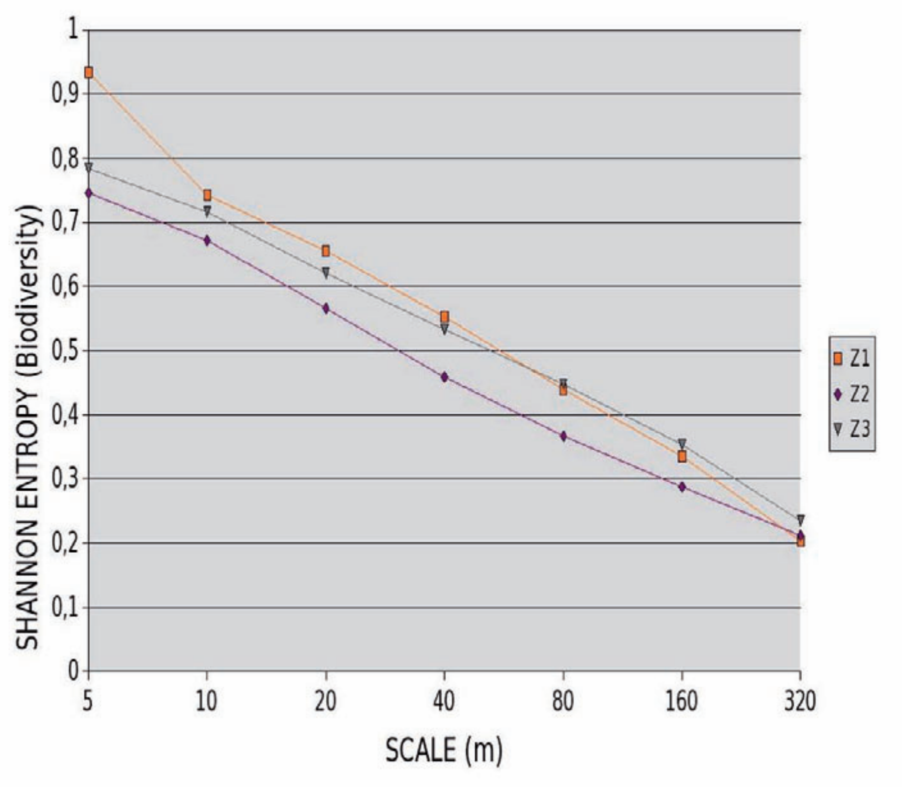


Fig. 12. Relative loss of diversity (%) through scales (m) for biodiversity images on different zones (Z1, Z2 & Z3)

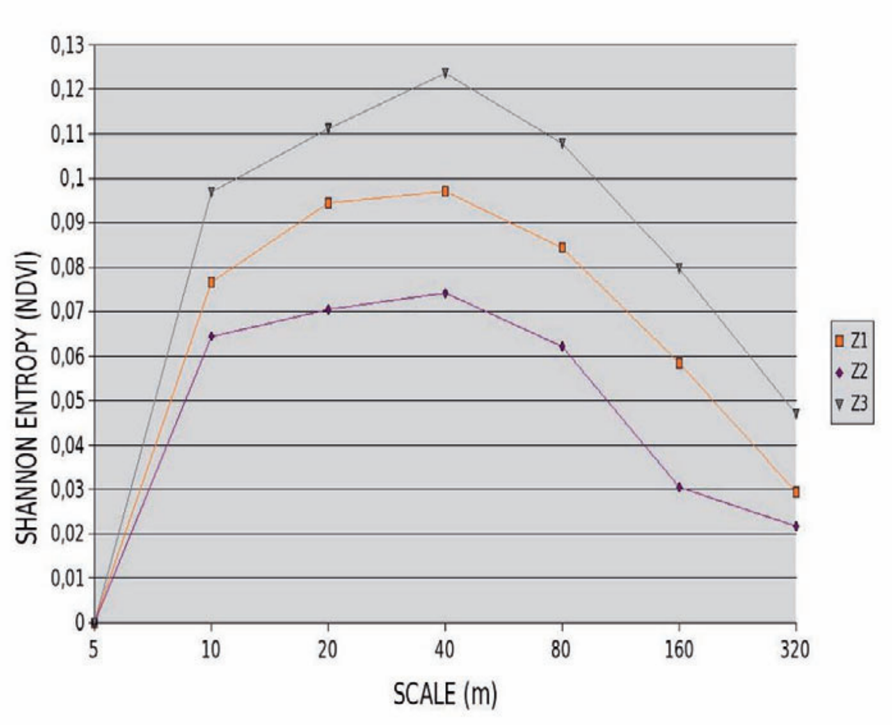


Fig. 13. Relative loss of diversity (%) through scales (m) for NDVI images on different zones (Z1, Z2 & Z3)

We can observe on figure 12 the same results as seen previously when comparing the two attributes (NDVI and species biodiversity): curve shapes are similar and levels of biodiversity are equivalent. There also remains a noticeable effect of statistical robustness (impact of the number of classes on distributions). Contrary to the previous results related to the spatial dimension (essentially the change of support and spatial autocorrelation), the part of the MAUP due to statistical robustness has not be tackled and it can still remains it the biodiversity measures.

Regarding the three images corresponding to rather different areas, with different species and biomass spatial distribution, it is quite surprising not to see many more differences. Indeed, though those three areas show rather different spatial autocorrelation, they keep close levels of relative diversity loss (deviation from 5 to about 15 % of relative biodiversity loss).

Looking at figure 13 and considering only the real biodiversity loss measured, we can notice that the NDVI diversity loss in Z3 image is about the double of Z2 at 40m scale. If we assume that there is no more aggregation effect in this value, we can easily assign a rank of biodiversity loss to those images, loss being directly linked to spatial autocorrelation.

5.4. What is the effect of the initial image resolution on the evaluation of biodiversity?

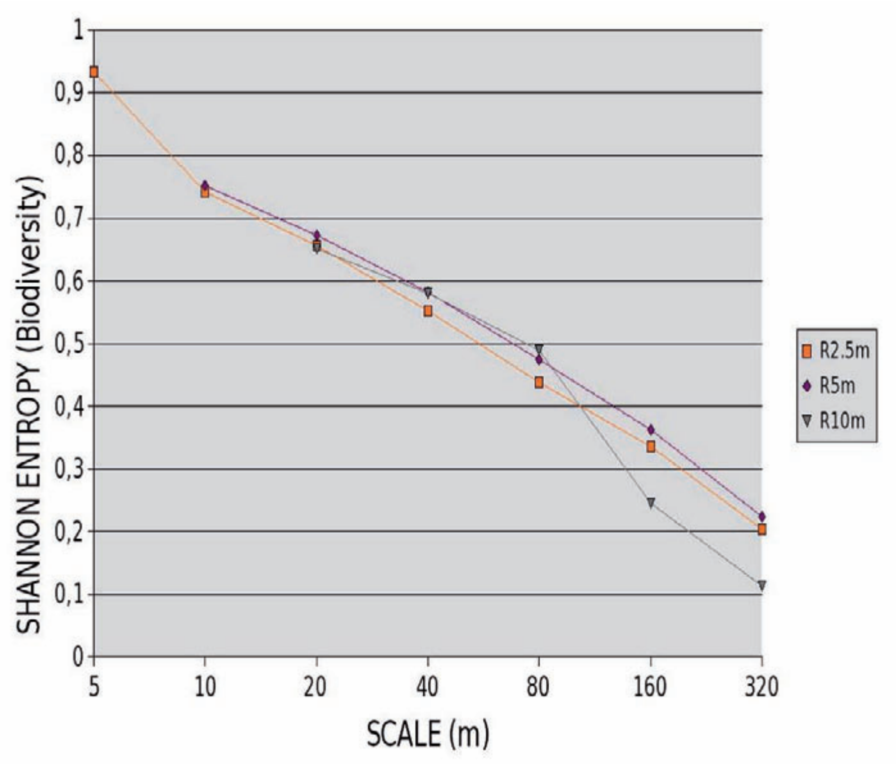


Fig. 14. Relative loss of diversity (%) through scales (m) for biodiversity images of different resolutions (2.5, 5 and 10 meters)

We now propose to study the effect of the initial resolution on entropy, using 2.5, 5 and 10 meters resolutions (figures 14 & 15). This parameter does not seem to particularly affect species biodiversity evaluation. This tends to prove that the image quality makes them quite reliable, even if we can remark some curve deviations at higher scales (160, 320 m).

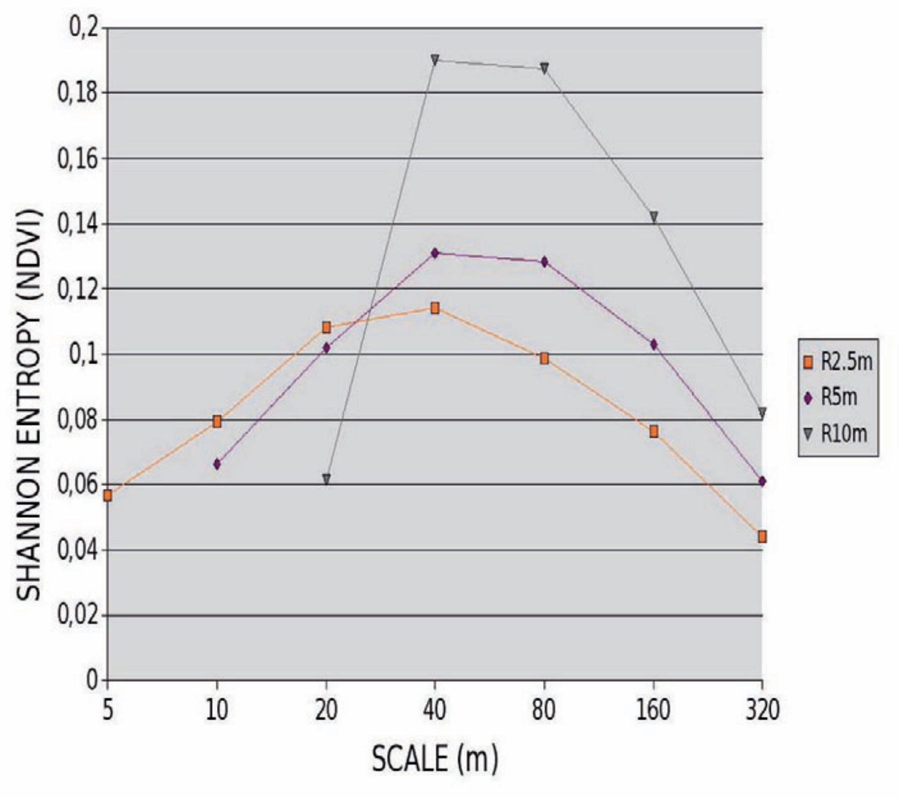


Fig. 15. Relative loss of diversity (%) through scales (m) for biodiversity images of different resolutions

What is most interesting is the fact that there exists a recurrent common maximum value of biodiversity loss for both attributes. The minimal aggregation resolution always corresponds to the highest species biodiversity loss and a middle peak in the NDVI relative diversity curves enhances the scale in which diversity loss is maximal.

6. Conclusion

These results lead us to discuss the concept of 'pertinent scale'. We obtained two contradictory arguments. On the one hand, we noticed that the values of pertinent scales were stable for a given attribute, whatever the size, shape, initial resolution and location of images: 40-80 m for NDVI and 5 m for species classification. On the other hand, the peak that would identify the maximum of measurable biodiversity loss did not appear for species classification.

The highest value of relative diversity loss is somehow the scale where the maximum of spatial autocorrelation can be found. On the opposite, the scale to comprehend biodiversity could be the one corresponding to the lowest diversity loss. This raises the issue of the 'relevant scale': does it correspond to the maximum of diversity loss (and the maximum of spatial autocorrelation) or to the minimum of diversity? Other research handling alternatives measures such as dominance index and Simpson index tend to show that this concept must be used with caution, and that it cannot be yet generalized.

Nevertheless, using different images at different scales and resolutions, we showed how strong the MAUP can be on the assessment of biodiversity, because of the ecological fallacy problem. We succeeded in adjusting our values to focus on the loss of diversity expurgated from the MAUP using a resampling process and a graph comparing biodiversity loss and scales. The 'relative loss of biodiversity' was computed and the recurrence of its design demonstrates that the attribute considered for evaluation is more important than spatial support. Our approach finally provides a multiscalar view of biodiversity, which is very helpful to determine whether or not and in what conditions the diversity measurement can be reliable.

References

Amrhein, C., (1995), Searching for the elusive aggregation effect: evidence from statistical simulations, *Environment and Planning A*, 27, 105-119.

Atlan H., (2006), *L'organisation biologique et la théorie de l'information*, La librairie du XXIème siècle, Seuil:Paris.

Baker W., (1997), The r.le Programs, A set of GRASS programs for the quantitative analysis of landscape structure. Version 2.2, University of Wyoming, USA. http://grass.itc.it/gdp/terrain/r_le_22.html

Baldwin David J. B., Weaver Kevin, Schnekenburger Frank & Perera Ajith H., (2004), Sensitivity of landscape pattern indices to input data characteristics on real landscapes: implications for their use in natural disturbance emulation, *Landscape Ecology*, Vol. 19-3, 255-271.

Clark W.A., Karen L., (1976), The effects of data aggregation in statistical analysis. *Geographical Analysis*, vol. VIII, 429-438.

Cliff, A. D. & Ord J. K. (1973), *Spatial autocorrelation*, Pion:London.

Cressie N., (1993), *Statistics for spatial data*, Wiley:NY.

Dusek T., (2005), The Modifiable Areal Unit Problem in regional economics. The 45th Congress of the European Regional Science Association:Amsterdam.

Frontier S., (1983), L’échantillonnage de la diversité spécifique. In *Stratégie d’échantillonnage en écologie,* Frontier et Masson (eds) Paris (Coll. D’Ecologie).

Farina A., (2000), *Landscape Ecology in action*, Kluwer:London.

Gehlke C.E., Biehl, K., (1934), Certain effects of grouping upon the size of the correlation coefficient in census tract material, *Journal of the American Statistical Association*, 169-170.

Gotway Crawford C. A. and Young L. J. (2004), A spatial view of the ecological inference problem in *Ecological Inference. New Methodological Strategies* Series: Analytical Methods for Social Research, Gary King, Ori Rosen Martin, A. Tanner (Eds).

Jelinski D.E., Wu J., (1996), The modifiable areal unit problem and implications for landscape ecology. *Landscape Ecology*, vol. 11-3, 129-140.

Lovett Gary M., Turner Monica G., Jones Clive G., Weathers Kathleen C. (eds), (2006) *Ecosystem Function in Heterogeneous Landscapes,* Springer.

Marceau D.J., Howarth P.J., Gratton D.J., (1994) Remote sensing and the measurement of geographical entities in a forested environment; part 1, The scale and spatial aggregation problem, *Remote Sensing of environment*, vol. 49-2, 93-104.

Mahfoud I., Josselin D., Fady B., (2007) Sensibilité des indices de diversité à l’agrégation, in Informations géographiques. Structuration, extraction et utilisation, C. Weber & P. Gançarski (Eds). *Revue Internationale de Géomatique*, Hermès, Paris, vol. 17, 3-4, 293-308.

Openshaw S., (1984), The modifiable areal unit problem. *Concepts and Techniques in Modern Geography.*, Number 38, Geo Books:Norwich.

Rastetter E.B., King A.W., Cosby B.J., Hornberger G.M., O’Neill R.V., Hobbie J.E., (1992), Aggregating fine-scale ecological knowledge to model coarser-scale attributes of ecosystems. *Ecological Applications 2*, 55-70.

Reynolds, H. D., (1998), *The modifiable areal unit problem: empirical analysis by statistical simulation*, Thesis, University of Toronto.

Robinson A.H., (1950), Ecological correlation and the behaviour of individuals. *American Sociological Review*, 15, 1-357.

Tucker, C.J. (1979) Red and photographic infrared linear combinations for monitoring vegetation, *Remote Sensing of Environment*, 8-2, 127-150.

Wu J., Gao W., Tueller P.T., (1997), Effects of changing spatial scale on the results of statistical analysis with landscape data: A case study, *Geographic Information Sciences 3*, 30-41.

Wu J., Levin S.A., (1994), A spatial patch dynamic modelling approach to pattern and process in annual grassland, *Ecological Monographs*, 64, 447-467.

Yule, G.U. and Kendall, M.G., (1950), *An introduction to the theory of statistics*, Griffin:London.

Implicit Spatial Information Extraction from Remote Sensing Images

Erick Lopez-Ornelas[1], Guy Flouzat[2]

[1] Universidad Autónoma Metropolitana – Cuajimalpa
Constituyentes 1054, Del. Miguel Hidalgo,
11950 México
email : elopez@correo.cua.uam.mx
[2] Université Paul Sabatier, Toulouse III,
118 Route de Narbonne,31062 Toulouse cedex 4
email: guy.flouzat@ict.fr

Abstract

In this paper we describe the basic functionality of a system dedicated to process spatial information (high-resolution satellite image) and to handle it through (semi-) structured descriptors. These descriptors enable to manage in a unified representation two families of features extracted from the objects identified by image segmentation: the attributes characterizing each object, and the attributes characterizing relationships between objects. Our aim is to focus on the complementarily of two approaches, on one hand concerning the segmentation of spatial information, and on the other hand concerning the knowledge discovery and the modeling.

Keywords: XML, Morphological Processing, spatial information, data modeling.

1 Introduction

The availability of spatial information has been a very significant step in remote sensing. It enables new capabilities in studying a range of non-observable objects until now. However, due to the higher spatial resolution, their processing becomes harder, especially in heterogeneous areas such as the urban areas [15].

The solution we chose, based on a description by graphs and mathematical morphology [30], is a strategy to make an interesting segmentation for the following processing. Indeed, this segmentation must be applied even on images that have an insufficient level of radiometric quality. Moreover, in the segmentation step, different reasoning models must be developed to acquire information on the scene.

Segmentation is the process of grouping an image into units that are homogeneous with respect to one or more characteristics [14]. This method is based on a description of satellite images using an adjacency graph, represented by the Voronoï diagrams, and the application of morphological transformation sequences in order to obtain relevant and significant objects. The result is twofold, with on one hand some specific attributes of the object, and, on the other hand, a set of spatial attributes characterizing relationships between these objects. This is the main interest of the segmentation method that is well adapted to cooperate with knowledge mining.

In this paper, our aim is to illustrate the complementarily of two components of such methodology, one that concerns the segmentation spatial information, and the other one that concerns the knowledge discovery and the modeling. We focus here on the system implemented to associate the segmentation process to the management of image descriptors, gathering the elicited features. Indeed, the “final” segmented spatial information is modeled by a set of instances, which enable to get a higher-level exploitation using directional, metrical or topological queries.

To illustrate interactions between segmentation techniques and underlying models, a unified representation is proposed, based on (semi-) structured descriptors. This modeling task is fundamental, as we must generate relevant descriptions in order to process them without accessing the initial image, only handling the elicited features as metadata, to select relevant objects, for example by querying them [28].

Finally, we show how this method can be applied to extract non-explicit spatial information using the spatial relationships implemented using this process.

2 Spatial Information Processing

For the last twenty years, a variety of image segmentation techniques have been developed. These techniques are classified into six categories: (i) “thresholding”, that separates the histogram of the image into a number of peaks, each corresponding to one or more regions [21]; (ii) “edge approach”, based on the detection of discontinuity on gray level [8]; (iii) “region based approach” attempts to group pixels into homogeneous regions with similar properties, [16]; (iv) “fuzzy approach” provides a mechanism to represent and manipulate uncertainty and ambiguity [22]; (v) “neural network approach” is composed of many computational elements connected by links with variable weights [24]; (vi) other approaches like the “fractal approach” [33], the “use of snakes” [17], “data mining” [29] or “multi-resolution approach” [2], [25] had been used successfully. Other segmentation techniques have been applied for feature extraction from remote sensing images [20], [27] and [23].

2.1 First pre-processing operations

To simplify the segmentation process, the watershed transformation [32] is applied to have a first group of regions, which are represented by a graph. In order to apply morphological processing, these regions are considered as a Voronoï cells. So, the region (Ri) of the image is identified as a virtual Voronoï diagram Vor(P) adapted to the image [31]. This Voronoï diagram is identified in order to exploit the set of neighbors [7]. The Delaunay triangulation is the dual structure of the Voronoï diagram. By dual, we mean to draw a line segment between two Voronoï vertices if their Voronoï polygons have a common arc. DT(p) = (P, A), where P is the set of vertices, and A is the set of arcs.

This process makes it possible to represent the image using a graph (G) by the Delaunay triangulation (DT), creating a correspondence between the radiometric value of each object Ri and the graph.

In order to segment these high resolution images we use a segmentation technique based on morphological processing [26], [34], [4]. This technique is fundamental because it rests on the study of the geometry, the forms, the simplification and the conservation of the principal features by comparing objects with the structuring element [19], which permit it to be applied on graphs [29] and constructing various neighborhood graphs on a set of P objects.

2.2 Morphological Processing

This principle consists in transforming the G(p) value by affecting the nearest radiometric value val(pi) among the p neighbors, and the grouping process is the union of regions (Ri $\cup$ Rj = Rn). The new object Rn is then the result of the fusion of regions. To carry out this transformation, the morphological operations (erosion and dilation) on the graph have to be applied. The dilated graph δ(G(p)) and the eroded graph ε(G(p)) are defined by [31] as

$$\forall p \in P: \delta(G(p)) = \max\{G(pi), pi \in NE(p) \cup \{p\}\} \text{ and}$$
$$\varepsilon(G(p)) = \min\{G(pi), pi \in NE(p) \cup \{p\}\}$$

where N_E(p) is the set of neighbors of p. Other two operations need to be implemented (opening and closing) and are also defined by [31]. The open graph γ(G(p)) and the closed graph ϕ(G(p)) are:

$$\gamma(G(p)) = \delta(\varepsilon(G(p))) \quad \text{and}$$
$$\phi(G(p)) = \varepsilon(\delta(G(p)))$$

The geometrical action of opening and closing transformations, γ(G(p)) and ϕ(G(p)) respectively, enlarges the region size. We can regulate the fusion using the parameters of opening and closing and comparing the size of regions or the length of contact among the connected regions. A loop is then required to alternate the transformations. The graph has to be updated to keep aggregating the different regions always applying the morphological transformations of γ(G(p)) and ϕ(G(p)) until their parameter value is reached.

The result of this process is a segmented image with a set of significant regions. These regions respect spatial and radiometric parameters that permit to obtain a segmented image respecting physical characteristics with a good spatial organization [19].

In figure 1, we show the original image and the segmentation result using a high-resolution satellite data. It is a high resolution panchromatic image delivered by Ikonos satellite from Toulouse, France having 1m resolution. In this segmentation process, we focus mainly on the extraction of houses and we do not focus on streets or green areas.

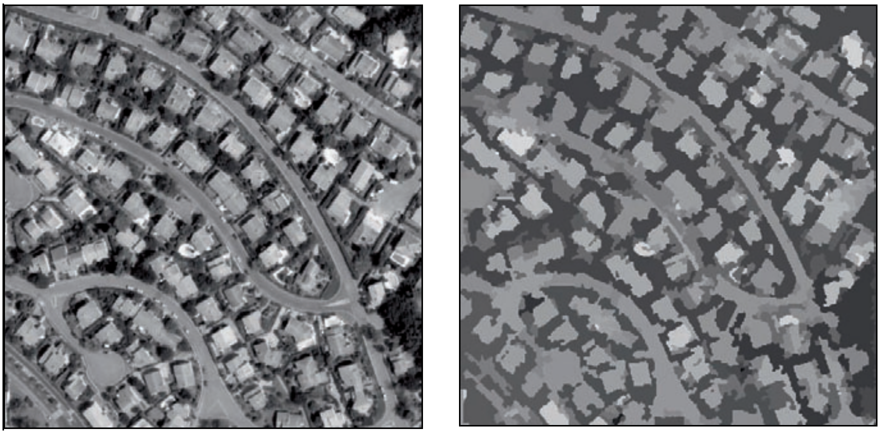


Fig. 1. Original and segmented High-resolution image.

3 Generating Descriptors

The next step of this process consists in calculating the final graph of the image from which one has elicited the set of characteristics {C}. These features {C} called “metadata” [1], [9] characterizing each region are then stored and handled separately from the segmented image. The image will be accessed only to visualize the results.

There are two different features extracted from each region: (i) “intrinsic features”, that are specific to each region (area, perimeter, neighbors, the radiometric value, the minimum bounding rectangle, etc) and (ii) “spatial features”, that describe spatial relations among the regions (topological, metrical and directional). Additional information also can be stored, like the shape of each object or texture [1].

The segmented image I is then considered as a set of layers information $\{L\}=(l_1,\dots,l_n)$. Each layer is associated with a set of characteristics {C}, with their values {v}, which defines the structure of each instance stored according to the layer. Each instance is identified by a unique identifier Oid and a descriptor called the value $v(v=(v_1,\dots, v_n))$. This information can be located via a set of attributes $\alpha = (\alpha_1,\dots, \alpha_n)$. The schema of the layer gives the structure of descriptor α.

3.1 XML-based representation

Any elicited information is translated into XML tags [9]. The spatial position g is given as a result of labeling during the process of final construction of the graph. So, the 3-tuple (Oid,v,g) symbolically describes the instance. This information upon the different objects are gathered and stored into XML descriptors (figure 2).

Indeed, this information is then available with a minimum number of operations. The aim of this eliciting process is to enable post-processing upon the images without referring to the image itself. So we can query these XML descriptors to extract the desired information. In this way, this independence is a great advantage.

```
<?xml version="1.0" encoding="UTF81.0"standalone = "yes"?>
<image>
  <object number = "0" >
            <rad_value> 75 </ rad_value > <area> 637 </area>
            <perimeter-ext> 231 </perimeter-ext>
            <perimeter-int> 224 </perimeter-int>
            <perimeter> 227 </perimeter>
            <card_opening> 637 </card_opening>
            <card_closing> 638 </card_closing>
                 <neighbor number = "8" >
                <topology> inside </topology>
               <angle>330.90</angle><distance>114 </distance>
                <leng_contact> 231 </leng_contact>
            </neighbor>
            <mbr>
                <xmin> 63 </xmin> <xmax> 119 </xmax>
                <ymin> 131 </ymin> <ymax> 160 </ymax>
                <xcenter>91</xcenter><ycenter>146 </ycenter>
            </mbr>
  </object>
  <object number = "8" >
        :
  </object>
</image>
```

Fig. 2. Example of descriptor.

The system stores data in its repository, and supports XPath/XQuery query language, that can be more efficient than multi-table SQL joins with unknown (i.e. null relational) values.

4 Spatial Information Retrieval

Retrieving information is made through a rule expressed by a query [12]. Defining a query relies on the formal notions of:

$$Q = (l, L, C_q)$$

with l the layer of the set of instances, Cq the conditions of the instances in $L = \{l_1, \ldots, l_n\}$.

The equivalent query via the traditional querying language OQL (Object Query Language) would be:

$$\text{Select } \{R_i\} \text{ from } l_1 \ldots l_n \text{ where } C_q.$$

In this way, Q(I) is defined as a result for the Q query applied to the result of the segmentation I. So, Q represents a spatial query where the condition C contains a spatial predicate and the result of Q is based on located instances.

In order to exploit characteristics via the descriptors of the objects, spatial relationships are the most important ones because they describe the relationships between the instances and their characteristics. They are essential to process the queries. The spatial relationships managed can be classified in three categories: topological, directional, and metrical.

Image description and information retrieval is illustrated in the following examples, where the objective is to search the information and organization of the extracted features. Knowledge extraction is applied by querying only the set of descriptors. In the next section we show some examples. This knowledge extraction represents the non explicit information that can be elicited from the spatial information. In order to extract this information some queries have to be applied.

4.1 (Semi-) structured Query Algebra

This selection can be defined using a simple algebra to access only to descriptors (and not to the image itself) even if the corresponding objects are displayed as a result of the query. We explain some query examples based on the segmented image of figure 1.

Query 1a: “Display all objects having a radiometric value between 220 and 250 with an area greater than ‘k’”. This query correspond the extraction of all houses on the segmented image. This query can be expressed as:

$$\{Ri : Ri \in I, 220 < rad\_value(Ri) < 250\} \text{ and}$$
$$\{area(Ri) > k\}$$

The result of this query is shown in Figure 3.

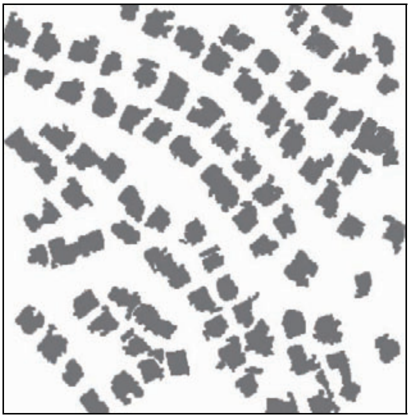


Fig 3. Segmented image and selected objects.

We are now interested in some specific parts of the segmented image. In the next query we combine directional and metrical descriptors with textual ones.

Query 2a: “Display all objects greater than ‘k’ with a distance ‘m’ from the object ‘n’”. This query correspond the extraction of all houses on the segmented image having a distance ‘m’ from the central house ‘n’. This query can be expressed as:

$$\{R_i : R_i \in Neighbor(R_n)\} \text{ and}$$
$$\{dis\tan ce(R_i) < m \wedge area(R_i) > k\}$$

The results are the black colored houses in Figure 4.

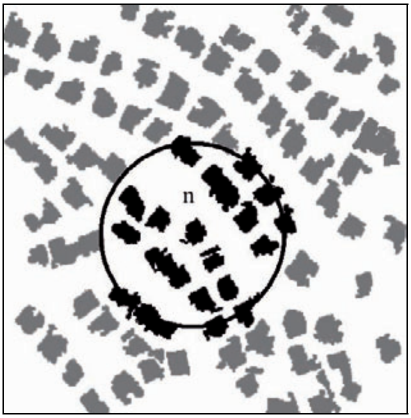


Fig. 4. Segmented image and selected objects with relationships and feature information {C}.

This kind of queries allows extracting the different elements from the segmented image (houses, streets, gardens, etc.). Likewise, this method allows to query each descriptor (intrinsic or spatial) extracted and combine them in order to get an appropriate information retrieval of the spatial information.

4.2 Selecting with XQuery

We illustrate the use of XQuery (or any other XQuery-like language) in order to retrieve information. We can use XPath, which is a simple language that allows navigating through XML structures and retrieving a set of XML nodes [10], [13].

Projection can also be done by using FLWR[1] expression (which looks like SQL), even if path expression is easier and more compact compared to FLWR expression.

This expression is capable of hierarchical "pattern matching" against the tree-structured XML data model and of "restructuring" the result tree. Using this technique, we can express a more specific query.

For example to traduce the Query 2a that uses spatial relationships and feature information of the image, it can be expressed by using two variables representing the same feature:

Query 1b: "Display all objects greater than 'k' with a distance 'm' from the object 'n'"

[1] The acronym FLWR refers to the fact that it consists in one or more *for* and/or *let* clauses, an optional *where* clause, and a *result* one.

```
for $im in document(«Images.xml»)/image/object
 let $ima :=document(«Images.xml»)/image/object/[@id=n]
 let $gima :=grav($ima)
 let $gim :=grav($im)
 let $m :=distance($gima, $gim)
 where ($im/rad_value > 220 and $im/ rad_value < 250) and
 ($im/area > k ) and ($d < m)
 return
      <object id={$im/@id}>
      </object>
```

where grav and dist are functions that calculates the center of gravity and the distance of two objects respectively. The result is shown in Figure 4.

5 Non explicit information

All details that it can be appreciated from an image are very important and they are not always extracted using the different image processing methods. Using this kind of information retrieval and querying, we have now the possibility of the extraction of a set of non explicit spatial information like the arrangement of land use in urban or suburban areas. Even if this information exists in the image, sometimes it is difficult to extract automatically or semi-automatically. This kind of information refers to: i) the density of houses in a selected area to know whether an area (e.g. residential) has a low or a medium density, ii) the verification of distances among houses, in order to satisfy communal criteria, iii) the identification of orientation of houses in order to study the community organization and its evolution by comparing different housing areas, iv) Identify the size of buildings and parcels for a better development of the municipality or to identify non allowed constructions, v) Identify green areas in order to have a better proportion between them and buildings.

With this method we can extract and identified on a simple way some of this information mentioned above. For example, analyzing the selected objects from figure 3, we can easily have the number of houses, the distance between them, the average of neighbors related of each house, house density on the image, etc. This can be modeled with the same principle used before using Voronoï diagrams and Delaunay triangulation (Figure 5). Having this information we can even calculate the "date" of construction of each urban area because this "date" is usually related to the number of houses and the distance among them. This reasoning using distances is very useful but it can also be applied to a study of house orientation or the study of the form of the houses.

This section shows how this morphological processing and the use of descriptors can be very important to extract all different elements from the high resolution image and it shows how some non explicit elements can be also extracted and discovered.

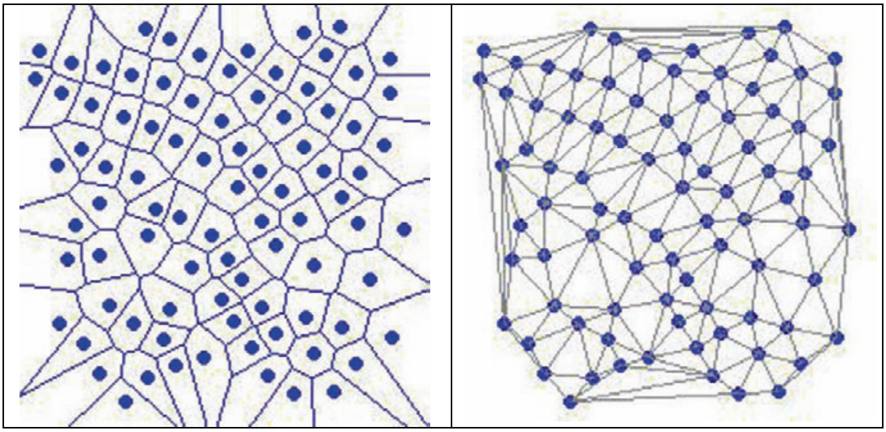


Fig. 5. The use of Voronoï diagram and Delaunay triangulation of the extracted objects from the high-resolution image

6 Conclusion

The aim of this paper is to give the main elements of our approach about key-segments extraction, image content and spatial relationship description and retrieval. The overall data-management architecture relies on the segmentation module, which produces segments and their associated features, considered next as metadata upon them. Once the image input is processed and these metadata elicited, they are gathered into descriptors. The retrieval module queries such a collection to retrieve, display and organize - according to their spatial relationships-, objects relevant to the user request. The result of querying eventually contains the links to images and/or key-segment(s).

We demonstrated that a successful spatial information management system depends on the effectiveness of three most important components: segments extraction, content descriptions and retrieval. Thus, we have been proposing a segmentation approach based on morphological processing, which integrates and chooses segmentation techniques that work more effectively for the current type of images we have to process. Once segments are identified, they must be described using standard descriptions to form descriptors referring to the actual image, so the users can retrieve them and their segments according to their characteristics or relative location.

For this purpose, we chose to extend the first modeling by XML that enables a more flexible and adaptive framework, adding specific annotations and customizing them. For retrieval, we have chosen to implement through XQuery [5], [18], [6], currently the most recent XML query language, and most powerful one compared to the older ones [3], [11].

This method can also be applied to retrieve non explicit information because using the extracted information from the descriptors, different kinds of reasoning can be applied to the high resolution image.

We are thinking of extending this work. First of all, we must test performances on different collections of heterogeneous images. Another problem is to implement an XQuery engine optimizer that supports our complex queries. Furthermore, we need to use our specific domain approach to generalize implementation to enrich the construction and exploitation of the symbolic extracted information (metadata). This will improve the quality of information on land use in the dedicated GIS (Geographical Information Systems).

Finally, we consider as a fundamental future work the potential relevance feedback about the query result in several domains. For example, a similar approach can be developed on digital imagery in the bio-medical domain. Indeed, in such applications, the discovery of spatial information based on medical knowledge can contribute to interesting and important diagnostics.

References

1. Amous, I., Jedidi, A., Sèdes, F. (2002) “A contribution to multimedia document modeling and organizing.”, *8Th International conference on Object Oriented Information Systems, OOIS’02*, Sept. 2002. Springer LNCS n° 2425, pp. 434-444.
2. Baatz, M., and Shape, (2000) “A. Multiresolution segmentation: an optimization approach for high quality multi-scale segmentation”, *Applied Geographical data processing*, pp. 12-23.
3. Baeza-Yates, R., Navarro, G. (2002) “XQL and proximal nodes”, *Journal of the American Society for Information Science and Technology*, Vol 53, no 6, pp.504-514.
4. Bagli, S and Soille, P. (2003). “Morphological automatic extraction of coastline from pan-european landsat tm images”. *In Proceedings of the Fifth International Symposium on GIS and Computer Cartography for Coastal Zone Management*, vol. 3, pp 58–59, Genova.

5. Boag, S., Chamberlin, D., Fernandez, M., Florescu, D., Robie, J., Siméon, J., Stefanescu, M. (2003), XQuery 1.0: An XML Query Language, W3C Working Draft, 12 Nov.
6. Chamberlin, D. (2003) “XQuery: A Query Language for XML”, Proc. SIGMOD Conf., San Diego, p 682, 9-12 June.
7. Chassery, J., and Montanvert, A. (1991) *Geométrie discrète en analyse d’images*, Ed. Hermes.
8. Cheng, W. J. (2001) “Color image segmentation: advances and prospects*”, Pattern Recognition*, 34, pp. 2259-2281.
9. Chrisment, C., Sèdes, F. (2002) “Multimedia Mining, A Highway to Intelligent Multimedia Documents”. *Multimedia Systems and Applications Series*, Kluwer Academic Publisher, V. 22, ISBN 1-4020-7247-3, 245.
10. Clark, J. (1999) XML path language (XPath), http://www.w3.org/TR/xpath
11. Deutsch, A., Fernandez, M., Florescu, D. Levy, A., Suciu, D. (1999) “Querying XML data”, *IEEE Data Engineering Bulletin*, Vol. 22, no. 3, pp. 10-18.
12. Dubois, D., Prade, H., Sèdes, F. (2001) “Fuzzy logic techniques in Multimedia database querying: a preliminary investigation of the potentials”. *IEEE Transactions on Knowledge and Data Engineering*, IEEE Computer Society, V. 13 N. 3, pp. 383-392.
13. Fernandez, M., Marsh, J., Nagy, M.(2002) XQuery1.0 & XPath2.0 Data Model, W3C, November, http://www.w3c.org/TR/query-datamodel/
14. Gonzalez, R. and Woods, R. (2008) *Digital Image Processing*, 3d ed. Addison-Wesley.
15. Hofmann, P (2001). “Detecting Urban Features from Ikonos Data Usingan Object-Oriented Approach”, *RSPS* , Geomatics, Earth Observation and the Information Society.
16. Horaud, R., and Monga, O. (1993) *Vision par ordinateur,* Ed. Hermes.
17. Kass, M.(1998) Snakes: “Active contour models”. *Computer Vision Graphics and Image Processing*, pp. 321-331.
18. Katz, H.(2003) XQuery from the Experts: A Guide to the W3C XML Query Language, Addison-Wesley.
19. Lopez-Ornelas, E., Flouzat, G., Laporterie-Dejean, F. (2003) “Satellite image segmentation using graphs representation and morphological processing”. *SPIE / Remote Sensing Image and signal processing for remote sensing IX*, vol. 5238, Barcelona, Spain.
20. Marangoz, A. M., Oruç, M., Büyüksalih, G., (2004). “Object-oriented Image Analysis and Semantic Network for Extracting the Roads and Buildings From IKONOS Pansharpened Images”, *ISPRS* XXth Congress, Istanbul.
21. Marion, A. (1987) *Introduction aux techniques de traitement d’images*, Ed. Eyrolles.
22. Mohhaddamzadeh, A., and Burbakis, N. (1997) “A fuzzy region growing approach for segmentation color images”. *Pattern Recognition*, 30, pp. 867-881.
23. Nussbaum, S, Niemeyer S, and Canty, M.J (2006) “SEATH–A new toolfor automated feature extraction in the context of object based image analysis.” *Proc. 1st International Conference on Object-based Image Analysis (OBIA 2006),* Salzburg.
24. Ong, S. H. (2002) “Segmentation of color images using a two-stage self-organizing network”. *Image and Vision Computing*, 20, pp. 279-289.
25. Petrovic, A., Divorra Escoda O., Vandergheynst, P. (2003) “Multiresolution Segmentation of Natural Images: From Linear to Non-Linear Scale-Space Representations”. IEEE Transaction on image processing
26. Serra, J. (1982) *Image Analysis and Mathematical Morphology*, Academic Press, London.

27. Shah V., et al., (2005) "Wavelet Features for Information Mining in Remote Sensing Imagery," IEEE Int'l Conf. Geoscience and Remote Sensing Symposium, vol. 8, pp. 5630–5633
28. Smith, J. R and Chang, S-F. (1999) "Integrated spatial and feature image query", *Multimedia Systems* Vol 7, No 2, pp 129-140.
29. Soh, C. (1999) "Segmentation of satellite imagery of natural scenes using data mining". *IEEE Transactions on Geosciences and remote sensing*, 37, pp. 1086-1099.
30. Soille, P (1999) *Morphological Image Analysis*, Springer-Verlag, Berlin.
31. Vincent, L. (1989) "Graphs and mathematical morphology". *Signal Processing*, 16, pp. 365-388.
32. Vincent, L., and Soille, P. (1991) "Watersheds in Digital Spaces: An Efficient Algorithm based on Immersion simulations". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 583-598.
33. Vuduc R. (1997) *Image segmentation using fractal dimension*. Rapport Technique, Cornell University.
34. Zanoguera, F. (2001) *Segmentation interactive d'images fixes et séquences vidéo basée sur des hiérarchies de partitions*. Thèse de Doctorat en Morphologie Mathématique, ENSMP.

The Application of the Concept of Indicative Neighbourhood on Landsat ETM+ Images and Orthophotos Using Circular and Annulus Kernels

Madli Linder, Kalle Remm, Hendrik Proosa

Institute of Ecology and Earth Sciences, University of Tartu,
46 Vanemuise St., Tartu 51014, Estonia
e-mail: madli.linder@ut.ee

Abstract

We calculated the mean and standard deviation from Landsat ETM+ red and panchromatic channels, and from red tone and lightness values of orthophotos within several kernel radii, in order to recognize three different variables: multinomial—forest stand type (deciduous, coniferous, mixed), numerical—forest stand coverage, binomial—the presence/absence of orchid species *Epipactis palustris*. Case-based iterative weighting of observations and their features in the software system *Constud* was used. Goodness-of-fit of predictions was estimated using leave-one-out cross validation. Cohen's kappa index of agreement was applied to nominal variables, and RMSE was used for stand coverage. The novel aspect is the inclusion of additional information from particular neighbourhood zones (*indicative neighbourhood*) using annulus kernels, and combining those with focal circular ones. The characteristics of neighbourhood in conjunction with local image pattern enabled more accurate estimations than the application of a single kernel. The best combinations most often contained a 10…25 m radius focal kernel and an annulus kernel having internal and external radii ranging from 25 to 200 m. The optimal radii applied on the Landsat image were usually larger than those for the orthophotos. The optimal kernel size did not depend on either reflectance band or target variable.

Keywords: indicative neighbourhood, case-based reasoning, Landsat ETM+, orthophotos, kernel size

1 Introduction

Automated classification of remotely sensed imagery enables to process considerable amounts of information. For decades, estimations of forest structure variables and land cover maps have been derived from air- and space-borne data. Remote sensing imagery based species occurrence predictions are also widespread. Until recently, image classification has commonly been performed using pixel-by-pixel approaches, i.e. categorizing each pixel according to its spectral signature and independently from other pixels. The emergence and increasing availability of high and very high resolution remote sensing imagery (satellite images and orthophotos with pixel edge up to less than 1 m) limit the use of conventional per-pixel classifiers. A pixel of a very high resolution image is difficult to classify because it does not carry enough information about the unit it belongs to. Wide range of spectral values of pixels representing certain classes makes it difficult to create sufficiently distinctive signatures for these classes, and this leads to poor classification results. In order to overcome this problem, several methods that incorporate spatial information from neighbourhood have been developed (Yu et al. 2006). Atkinson and Lewis (2000) divide these techniques into two groups: (1) methods providing information about texture and (2) methods used to smooth the image. The first rely on the assumption that texture defined as the spectral pattern and spatial distribution of pixel values in the surroundings of the focal pixel, differs by classes; in the second case, the pixels in the vicinity are presumed to have characteristics similar to the focal one. The spatial extent of the neighbourhood, i.e. the area comprising the pixels used in calculations, is delimited by the shape and size of the kernel (moving window). Kernels may be applied in various ways: different size and shape can be used, calculations may embrace all of the pixels in a kernel, or a sample, and the use of pixels may be delimited by a polygon (Fig. 1).

In order to provide sufficient information for unambiguous classification of a focal pixel, at the same time encompassing as less redundancy and noise as possible and thereby keeping the amount of calculations optimal, it is essential to choose optimal size and shape for the kernel when kernel-based methods are applied. From the point of view of automatic classification, the optimal kernel is the minimal extent of the window that yields the highest accuracy (Hodgson 1998). Franklin et al. (2000) note that inadequate window size is one of the main sources of inaccuracy in kernel-based classifications.

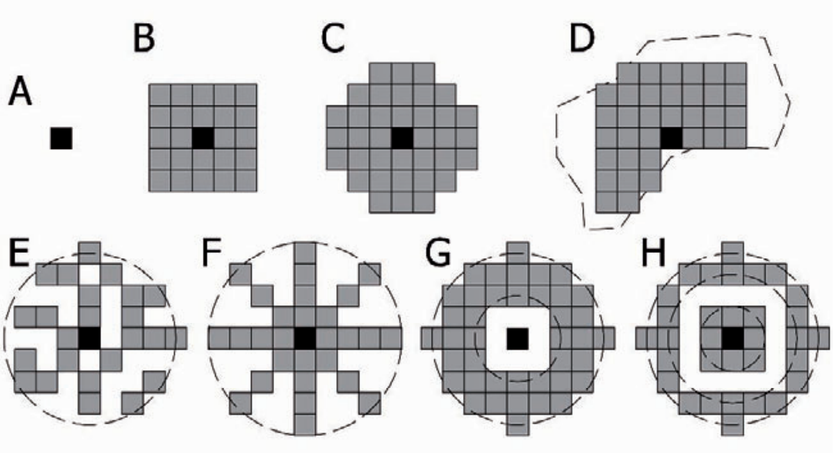


Fig. 1. Options for kernel use. **A** – one focal pixel; **B** – rectangular 5×5 pixel kernel; **C** – round kernel (in this case concurrently octangular); **D** – 7×7 pixel rectangular kernel masked by a polygon; **E** – random 50% sample from a circular kernel (radius 4 pixels from the focus); **F** – star-shaped sample from a circular kernel (radius 4.5 pixels); **G** – annulus kernel (internal radius 2 and external radius 4 pixels); **H** – combination of a circular internal kernel (radius 1.5 pixels) and an annulus kernel (internal radius 3 and external radius 4 pixels)

The subject matter of optimal kernel to enhance remote sensing imagery-based predictions and classifications has been a common topic for decades. The first studies on the relations between kernel size and classification accuracy date back to the 1970s (e.g. Haralick et al. 1973; Hsu 1978). Two fundamentally different approaches to the search for the optimal kernel can be distinguished: (1) the automatic empirical approach and (2) the cognitive approach, derived from the principles of the visual interpretation of remote sensing data (Hodgson 1998).

In the first case, the optimal kernel size is ascertained either experimentally, testing different kernel sizes and choosing the one that gives the best accuracy (Lark 1996; He and Collet 1999; Chica-Olmo and Abarca-Hernandez 2000; Remm 2005), or is derived from textural parameters of the imagery using mainly methods that are based on the textural variety of the image (variograms, covariation, correlograms) (Zawadzki et al. 2005).

An alternative to the fixed window size is the technique of a geographic window, i.e. automatically adjusting the kernel size and/or shape according to the local landscape characteristics (Dillworth et al. 1994; Franklin et al. 1996; Ricotta et al. 2003). The perceptual/cognitive approach, overviewed and experimented by Hodgson (1998), is based on the idea that when a human expert is incapable of classifying remote sensing data, a computer will either be. According to this, the optimal kernel size is the minimal window enabling the human expert to classify the image.

Characteristics of the first order neighbourhood may not be the most informative for estimating or identifying the phenomena in the focal location. Similarly, variables from too far distance have little influence on the properties of target objects at the location centre. Remm and Luud (2003) examined the 100 m wide distance zones up to 25 km from observed moose occurrence/absence locations located at potential moose habitats, and demonstrated that there is an optimal distance zone they called the *indicative neighbourhood*, which provides most information for the recognition of target objects. They found that the difference between the mean relative area of moose habitats of sites where moose pellets were found and where these were not found was greatest at a distance of 1...3 km from the observation site. The difference was slight or nonexistent very close to and too far from the observation sites (Fig. 2). Other ecological phenomena also reveal evidence of indicative neighbourhood. Remm (2005) investigated the correlation between forest stand diversity and landscape pattern at different distances from forest observation sites and found higher correlation at intermediate distances.

In kernel-related remote sensing studies, the common practice is to use single kernels with centres in the focal locations. Examination of neighbourhood independently of the immediate vicinity of the observation location is possible using annulus-shaped kernels; the functionality for applying these exists in many image processing and GIS software packages.

In this study, the concept of indicative neighbourhood was tested on remote sensing data by comparing the predictability of three different target variables using the explanatory variables calculated from medium and high resolution image data within (1) kernels of different sizes having a focus at the observation location, (2) neighbourhood zones at different distances (annulus kernels), (3) the combinations of different size circular focal kernels and annulus kernels. We proposed that instead of applying a single kernel, higher recognition accuracy could be achieved by using features calculated within an immediate neighbourhood in combination with a neighbourhood zone at a particular distance (within the indicative neighbourhood).

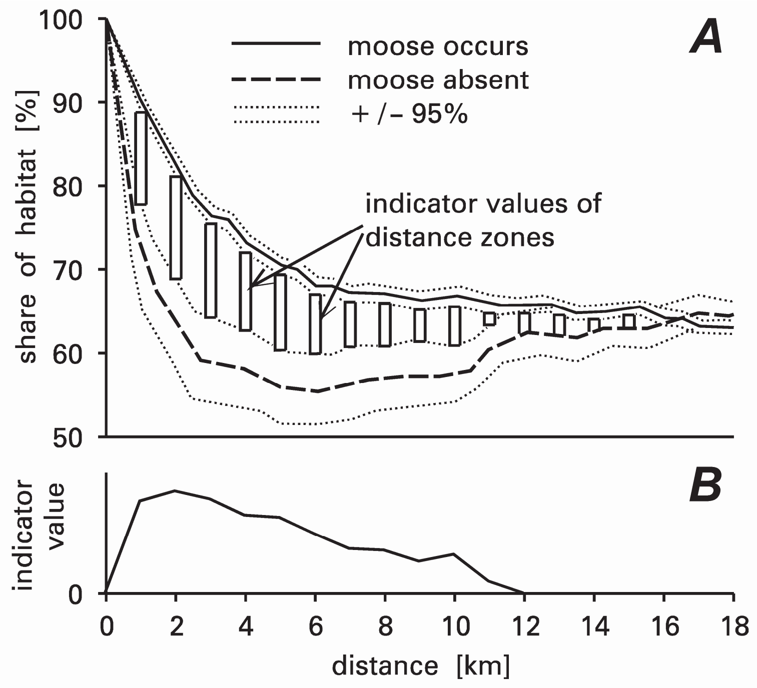


Fig. 2. A – indicator weights of distance zones as the difference between the mean relative area of moose habitats around locations where pellets were found (continuous curve) and around locations where pellets were not found (dashed curve). Dot-lines indicate the 95% confidence limits of the means. **B** – indicator value of the share of moose habitats in surroundings at different distances. Modified from Remm and Luud (2003)

2 Materials

2.1 Study area

Field data were gathered from an area of 700 km^2 located in Southeast Estonia (Fig. 3) characterized by high topographical variability and mosaic land cover. The predominant mixed forests are composed mainly of Norway spruce (*Picea abies*), birch (*Betula pubescens* and *Betula pendula*), Scots pine (*Pinus sylvestris*), aspen (*Populus tremula*), grey alder (*Alnus incana*) and black alder (*Alnus glutinosa*). Many small fens and swamps are rich in orchid species, of which the marsh helleborine (*Epipactis palustris*), studied in this experiment, is among the most frequent.

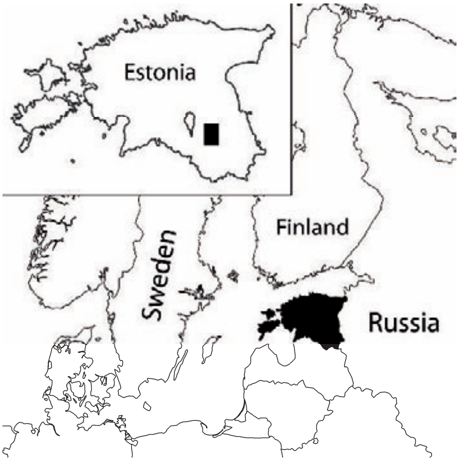


Fig. 3. Location of the study area

2.2 Field data

Field observations of vegetation mapping projects performed in summers from 2001 to 2006 were used to represent three types of variables: (1) numerical—coverage of forest stand, (2) multinomial—forest stand type (coniferous, deciduous, mixed), (3) binomial—the occurrence/absence of orchid species *Epipactis palustris*. The coordinates of the observation locations were recorded in the field using a GPS receiver and later verified on screen using orthophotos and 1:10 000 digital maps. Observation sites covered by clouds or *scan line correction off* stripes on the Landsat ETM+ image, as well as recently altered sites, were excluded. In the case of nominal variables, the sample was compiled to contain equal number of observations in each category. Random sampling was applied when the number of observations representing a particular class exceeded the amount necessary.

Observations of forest stand variables were made in plots where the forest stand structure looked homogeneous within a radius of at least 20 m, as seen from above. In addition, the homogeneity of soils and land use (based on the 1:10 000 soil map and the 1:10 000 base map, respectively) was prerequisite for suitable observation sites. Stand coverage (%) and ten-score stand composition formula were estimated visually as seen from above and considering the midsummer situation. The following stand types were attributed to the observations: (1) coniferous (the share of coniferous species equal to or greater than 8 points), (2) deciduous (the proportion of deciduous species equal to or greater than 8 points), and (3) mixed (the proportion of both deciduous and coniferous species less than 8 points). Altogether, 690 observations of stand type (230 for each class) qualified for the experiment. Stand coverage was not estimated at all observation sites, therefore the sample of coverage data consisted of 275 locations. Coverage in the selected locations ranges from 1 to 98%.

The presence of *E. palustris* was recorded on field excursions performed on foot through all habitat types of the study area. The grass layer was carefully observed, and both the finds and the observation track were recorded. The absence locations were later generated on the observation tracks. The set of 125 occurrence and 125 absence locations corresponding to the following restrictions was compiled: (1) availability of the orthophotos from the year 2006, (2) availability of the used Landsat image in the recorded locations (accounting the *scan line correction off stripes* as image unavailable), and (3) the in-between distance of locations at least 100 m. Too close observations cannot be considered independent due to the spatial autocorrelation of locational features—the habitat may be the same, the same pixels are used in the calculation of spatial indices, etc.

2.3 Remote sensing data

To identify whether there is any difference between remote sensing imagery of different spatial resolution in context of the issues studied, two data sets were used: Landsat ETM+ as representatives of medium resolution imagery and orthophotos as representatives of very high resolution imagery.

Landsat 7 ETM+ multispectral and panchromatic images acquired on June 11, 2006, were used. Original pixels of the Landsat ETM+ image were resampled into the Lambert Conformal Conic projection of the Estonian 1:10 000 base map coordinate system and into the 25 m × 25 m pixel size in the case of ETM+ band 3; panchromatic data layer yielded the pixel size 12.5 m × 12.5 m. Areas covered by clouds and *scan line correction off* stripes on the ETM+ image were treated as the absence of the data layer at the respective location.

Digital colour orthophotos from the year 2006 were originally in the Estonian 1:10 000 base map coordinate system and in RGB format. The local statistics software (description in Remm 2005) was used to separate the layers of red tone and general lightness. The pixel size of the orthophotos was transformed from the original of 0.4 m × 0.4 m to 1 m × 1 m during colour separation.

2.4 Explanatory variables

Two virtually comparable data layers for both the orthophotos and the Landsat image were selected for the experiment: the red tone/channel and the lightness/panchromatic channel. The mean and standard deviation (SD) of pixel values were calculated for each predictable variable around every study location from each data layer within different radii of circular neighbourhood, annulus-shaped neighbourhood and combinations of these (Table 1). A total of 762 explanatory variables were calculated.

3 Methods

The eligibility of different kernels for the estimations of the above-mentioned three variables was tested using case-based machine learning and prediction system *Constud* developed at the Institute of Ecology and Earth Sciences, University of Tartu. *Constud* was used for: (1) the calculation of spatial indices (explanatory variables) from image data, (2) machine learning and the estimation of the goodness-of-fit of predictions.

Case-based approach is similar to the *k*-nearest neighbour methods differentiating from the latter mainly by the use of machine learning. Overviews of case-based methods can be found e.g. in Aha (1998) and Remm (2004). Case-based alias similarity-based spatial predictions are based on the assumption that a phenomenon (e.g. a particular species, vegetation unit or any other study object) occurs in locations similar to those where it has previously been registered, i.e. on the similarity between studied cases and predictable sites.

Similarity-based estimation was chosen since it does not set any restrictions either on the type of relationship between variables or on the distribution of the data. The only presumption is the possibility of estimating the similarity. Furthermore, statistical modelling methods, as the main alternative to the case-based approach, always call for an abstraction—either in the form of a model or a set of constraining rules, and are therefore called rule-based reasoning. A new model should be created or refitted each time new data are added into the system. In case-based reasoning, a generalization in the form of a model is not created, and estimations are derived relatively directly from raw data, from the most similar feature vectors.

Machine learning (ML) in *Constud* is iterative search and weighting of features and observations needed for the most reliable similarity-based predictions of a target variable considering the data available in learning process. In this study, ML iterations consisted of three parts: (1) the weighting of features (in the case of combinations of 2 kernels), (2) the selection of exemplars from the set of observations, and (3) the weighting of exemplars. The observations were weighted one and the features 20 times; the number of ML iterations was 10 in the case of all 762 variables. The best result out of 10 parallel processes was used in the comparisons of prediction fits.

The similarity between observations is initially calculated as partial similarity of single features in *Constud*. In case of a continuous feature (*f*), the difference (*D*) between feature values (T_f and E_f) for an exemplar (*E*) and a training instance (*T*) is calculated as:

$$D = \frac{|T_f - E_f|}{2 \cdot w_E \cdot w_f}, \qquad (1)$$

where w_E—weight of exemplar *E*, w_f—weight of feature *f*. The partial similarity (S_f) between an exemplar and a training instance regarding a feature *f* receive a value *1 – D* if $D < 1$; otherwise $S_f = 0$. The total similarity is calculated as a weighted average of partial similarities. Further details are given in Remm (2004).

The number of exemplars used for predictions (*k*-value in *k-NN* methods) is controlled by the sum of similarity sought for a decision (*smax*) in *Constud*. The value of *smax* is optimized together with the feature weights as has been done also by Remm (2004) and Park et al. (2006).

The goodness-of-fit of the ML predictions was estimated by leave-one-out cross validation, i.e. the predicted value for every observation was calculated using all exemplars, leaving this observation out. The correspondence of estimations to observations was measured by Cohen’s kappa index of agreement in the case of multinomial and binomial variables. The accuracy of the numerical variable was estimated using root mean squared error (RMSE).

The two-sided sign test in Statsoft Statistica 7 was used to estimate the statistical significance of pairwise comparisons.

Table 1. The extents of kernels within which the mean and SD were calculated from the data layers of orthophotos and Landsat ETM+ image. First number – internal radius [meters from focal location], second number – external radius [meters from the focal location]

	Orthophotos		Landsat ETM+	
single circular kernel	0...1*		0...1*	
	0...10		0...25	
	0...20		0...50	
	0...30		0...100	
	0...40		0...200	
	0...50		0...300	
single annulus kernel	5...15		25...50	
	15...25		50...75	
	25...35		75...100	
	35...45		100...150	
	45...55		150...200	
	55...65		200...300	
	focus	vicinity	focus	vicinity
circular & annulus kernel	0...1*	5...15*		
	0...1*	15...25*	0...1*	25...50*
	0...1*	25...35*	0...1*	50...75*
	0...1*	35...45*	0...1*	75...100*
	0...1*	45...55*	0...1*	100...150*
	0...1*	55...65*	0...1*	150...200*
	0...1*	75...85*	0...1*	200...300*
	0...1*	95...105*		
	0...1*	115...125*		
	0...10	15...25		
	0...10	25...35	0...25	50...75
	0...10	35...45	0...25	75...100
	0...10	45...55	0...25	100...150
	0...10	55...65	0...25	150...200
	0...10	75...85	0...25	200...300
	0...10	95...105		
	0...10	115...125		
	0...20	25...35		
	0...20	35...45	0...50	75...100
	0...20	45...55	0...50	100...150
	0...20	55...65	0...50	150...200
	0...20	75...85	0...50	200...300
	0...20	95...105		
	0...20	115...125		
	0...30	35...45		
	0...30	45...55	0...100	100...150
	0...30	55...65	0...100	150...200
	0...30	75...85	0...100	200...300
	0...30	95...105		
	0...30	115...125		

*— SD not calculated

4 Results and discussion

Single kernel did not give the best results (the highest kappa or the lowest RMSE) in case of any investigated variable. In most cases, the best standard circular kernels gave more exact estimations than single annulus kernels. The highest accuracies were always gained when the local features and characteristics from a neighbourhood zone were combined (Table 2).

Table 2. The features and the best fits (kappa or RMSE) of the recognition of target variables using single kernels (annulus kernels are marked with *, others are circular kernels) and combinations of circular and annulus kernels

Feature				Fit	
Variable	Image	Band	Parameter	Single kernel	Combination
Coverage (fit as RMSE)	orthophotos	red	mean	19.29	17.74
			SD	19.44	19.08
		panchromatic	mean	22.22*	20.88
			SD	22.71	21.60
	Landsat ETM+	red	mean	17.16	14.92
			SD	20.98	20.10
		panchromatic	mean	20.23	19.06
			SD	21.55	20.35
Forest type (fit as kappa)	orthophotos	red	mean	0.27	0.42
			SD	0.28	0.38
		panchromatic	mean	0.24	0.35
			SD	0.17*	0.32
	Landsat ETM+	red	mean	0.22	0.35
			SD	0.50	0.60
		panchromatic	mean	0.08	0.23
			SD	0.17	0.29
E. palustris (fit as kappa)	orthophotos	red	mean	0.47	0.63
			SD	0.50	0.59
		panchromatic	mean	0.42	0.58
			SD	0.38*	0.54
	Landsat ETM+	red	mean	0.43	0.64
			SD	0.53	0.61
		panchromatic	mean	0.31*	0.54
			SD	0.28	0.43

The best combinations most often contained a circular kernel with a radius from 10 to 25 m and an annulus kernel that covers a particular zone between 25 and 200 m (Fig. 4). Similarly, Laurent et al. (2005), who investigated the kernel-based bird species habitat mapping options from

15 m resolution Landsat ETM+ images by testing 30...180 m wide radii, found 30 m to be the best kernel radius. Mäkelä and Pekkarinen (2001) suggest that a window larger than 3×3 pixels decreases accuracy due to the edge effect when Landsat TM data are used. The best annulus kernel was closest to the internal one in only two cases out of 24, confirming that the indicative neighbourhood is not the first order neighbourhood in most cases.

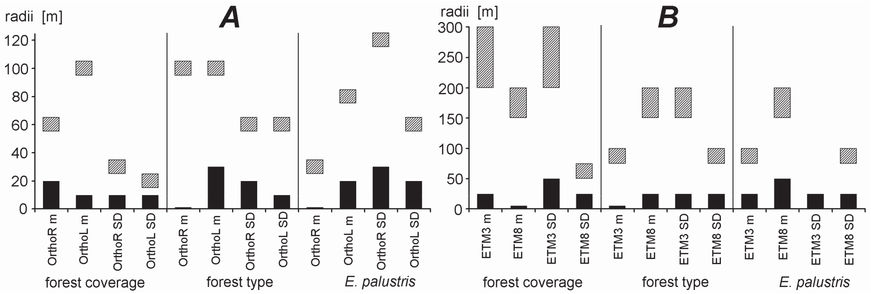


Fig. 4. The radii of circular (black columns) and annulus kernels (hatched columns) in the two-kernel combinations that gave the best prediction accuracies in the case of orthophotos (**A**) and Landsat ETM+ (**B**). Abbreviations: *Ortho* – orthophoto, *R* – red tone on orthophoto, *L* – lightness on orthophoto, *ETM* – Landsat ETM+ image, *3* – band 3 on Landsat ETM+ image, *8* – band 8 on Landsat ETM+ image, *m* – mean, *SD* – standard deviation

The optimal radii of both internal and annulus kernels applied on the Landsat image were usually larger than those for the orthophotos (p = 0.182 and p = 0.027). The best radius for the annulus kernel was closer to the focus in the case of SD than in the case of the mean value of image pixels, but the difference was statistically not significant (p = 0.182).

The optimal kernel size was not found to depend on reflectance band or target variable. It has been suggested that the optimal kernel size varies in accordance with the mean size of objects constituting the predictable classes on the image (Woodcock and Strahler 1987; Coops and Culvenor 2000; Kayitakire et al. 2006). Possibly, in this experiment, the spatial scales of the three target variables were too similar to reveal the influence of the size of objects under recognition. Although a specimen of an orchid is much smaller than a forest stand, the attempt was not to recognize orchids from the images; instead, the separability of habitats and non-habitats of *E. palustris,* that are comparable to the forest stands in size, was tested.

At first approximation, the average radius of the best outputs of single focal kernel options was greater than the radius of the internal kernel in the combinations of two radii. The explanation for the advantage of a larger kernel would be that although the local image pattern can usually serve as the best predictor of a target variable, then sometimes, when the predictable variable is weakly related to the image properties, the characteristics of the vicinity describing the landscape around a location can be better indicators. Thus, the optimal single kernel should be large enough to embrace the indicative neighbourhood to substitute the use of an annulus kernel. Nevertheless, the sign test indicated no significant difference in pairs of the best results of single circular kernel and two-kernel combinations within the same feature. If kernels with a radius greater than 50 m (4 out of 24 features) were removed from the list of radii of the best single kernels, the mean radii of single circular kernels and focal kernels were approximately the same—21.0 and 20.4 m. That is, in most cases the combination of two radii just adds a characteristic of vicinity to the focal parameters—features of the focal circular kernel have substantial indicator value both alone and in combination with the properties of the neighbourhood.

The following features expressed evidence of the indicative neighbourhood in our experiment: forest coverage estimated by the orthophoto red band (most indicative at 60 m), forest type by the SD of lightness of orthophoto image (most indicative at 60 m), forest type by the SD of the red band of the Landsat ETM+ image (most indicative at 175 m), forest type by the SD of lightness of the Landsat ETM+ image (most indicative at 75 m), *E. palustris* by the orthophoto red band (most indicative at 30 or 60...80 m), *E. palustris* by the red band of the Landsat ETM+ image (most indicative at 75 m) (Fig. 5). However, the differences in prediction fits between combinations are not great, and the plateau of higher indicative values is rather broad. Furthermore, the indicative vicinity was not revealed for many features. Therefore many of the best combinations may in essence be a random selection of more or less equal options.

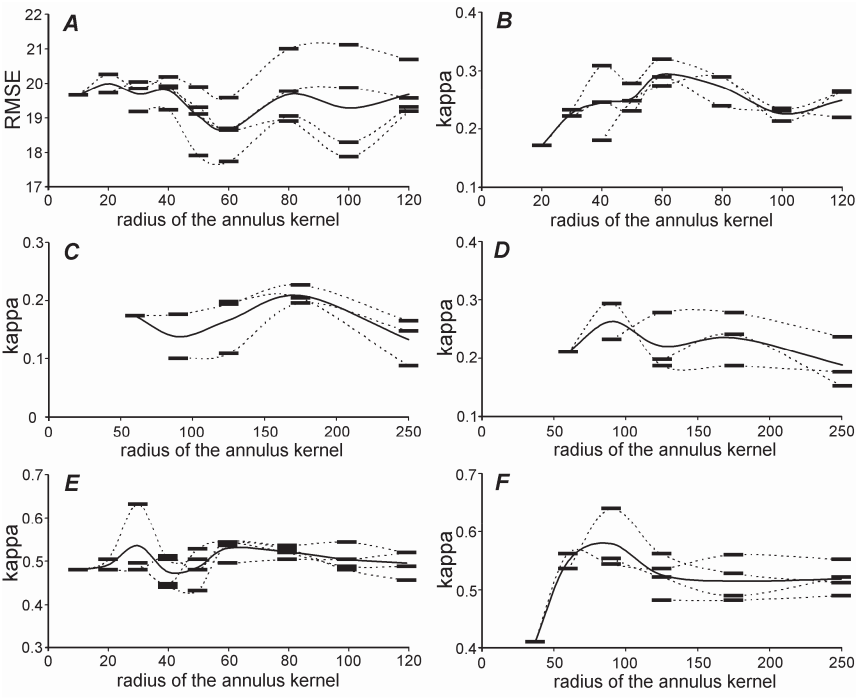


Fig. 5. Prediction fit in the case of different extents of annulus kernels (line-markers), and the mean fit over all tested radii of the internal kernel (bold line). **A** – forest coverage estimated by red tone on orthophoto. **B** – forest type by SD of lightness on orthophoto. **C** – forest type by SD of band 3 on Landsat ETM+ image. **D** – forest type by SD of band 8 on Landsat ETM+ image. **E** – *E. palustris* by red tone on orthophoto. **F** – *E. palustris* by band 3 on Landsat ETM+ image

5 Conclusions

The results of this study demonstrate that the remotely sensed image characteristics of the immediate vicinity in combination with the features from the particular intermediate-distance neighbourhood zone allow more accurate estimations compared to the use of single kernels. Properties of the surroundings at intermediate distance add most predictive value to the characteristics of a location due to the weaker autocorrelation and the certain intrinsic attributes that affect the quality of the study location. This distance—indicative neighbourhood—can be experimentally found. However, albeit we found evidence of indicative neighbourhood in this study, one cannot make far-reaching conclusions about which kernel combinations universally give the best results, because different combinations of internal circular kernel and outer annulus kernel yielded similar prediction fits. Additional comparative tests using different empirical and planned artificial data sets are needed to draw major generalizing conclusions on universal character of the indicative neighbourhood and its optimal extent.

Acknowledgements

The maps and orthophotos were used according to the Estonian Land Board licences (107, 995, 1350). The research was supported by the Estonian Ministry of Education and Research (SF0180052s07) and by the University of Tartu (BF07917). The authors are grateful to Mihkel Oviir and Liina Remm for participation in field work, and to Alexander Harding for linguistic corrections.

References

Aha DW (1998) The omnipresence of case-based reasoning in science and application. Knowledge-Based Systems 11:261–273

Atkinson PM, Lewis P (2000) Geostatistical classification for remote sensing: An introduction. Computers & Geosciences 26:361–371

Chica-Olmo M, Abarca-Hernandez F (2000) Computing geostatistical image texture for remotely sensed data classification. Computers & Geosciences 26:373–383

Coops N, Culvenor D (2000) Utilizing local variance of simulated high spatial resolution imagery to predict spatial pattern of forest stands. Remote Sensing of Environment 71:248–260

Dillworth ME, Whistler JL, Merchant JW (1994) Measuring landscape structure using geographic and geometric windows. Photogrammetric Engineering & Remote Sensing 60:1215–1224

Franklin SE, Hall RJ, Moskal LM, Maudie AJ, Lavigne MB (2000) Incorporating texture into classification of forest species composition from airborne multispectral images. International Journal of Remote Sensing 21(1):61–79

Franklin SE, Wulder, MA, Lavigne MB (1996) Automated derivation of geographic window sizes for use in remote sensing digital image texture analysis. Computers & Geosciences 22(6):665–673

Haralick RM, Shanmugam K, Dinstein I (1973) Textural features for image classification. IEEE Transactions on Systems, Man and Cybernetics SMC-3(6):610–621

He H, Collet C (1999) Combining spectral and textural features for multispectral image classification with Artificial Neural Networks. International Archives of Photogrammetry & Remote Sensing 32(7-4-3 W6), Valladolid, Spain, 3-4 June 1999:175-181

Hodgson ME (1998) What size window for image classification? A cognitive perspective. Photogrammetric Engineering & Remote Sensing 64(8):797–807

Hsu S (1978) Texture-tone analysis for automated land-use mapping. Photogrammetric Engineering & Remote Sensing 44(11):1393–1404

Kayitakire F, Hamel C, Defourny P (2006) Retrieving forest structure variables based on image texture analysis and IKONOS-2 imagery. Remote Sensing of Environment 102:390–401

Lark RM (1996) Geostatistical description of texture on an aerial photograph for discriminating classes of land cover. International Journal of Remote Sensing 17(11):2115–2133

Laurent EJ, Shia H, Gatziolis D, LeBoutonc JP, Walters MB, Liu J (2005) Using the spatial and spectral precision of satellite imagery to predict wildlife occurrence patterns. Remote Sensing of Environment 97:249–262

Mäkelä H, Pekkarinen A (2001) Estimation of timber volume at the sample plot level by means of image segmentation and Landsat TM imagery. Remote Sensing of Environment 77:66–75

Park Y-J, Kim B-C, Chun S-H (2006) New knowledge extraction technique using probability for case-based reasoning: application to medical diagnosis. Expert Systems 23(1):2–20

Remm K (2004) Case-based predictions for species and habitat mapping. Ecological Modelling 177:259–281

Remm K (2005) Correlations between forest stand diversity and landscape pattern in Otepää NP, Estonia. Journal for Nature Conservation 13(2-3):137–145

Remm K, Luud A (2003) Regression and point pattern models of moose distribution in relation to habitat distribution and human influence in Ida-Viru county, Estonia. Journal for Nature Conservation 11:197–211

Ricotta C, Corona P, Marchetti M, Chirici G, Innamorati S (2003) LaDy: software for assessing local landscape diversity profiles of raster land cover maps using geographic windows. Environmental Modelling & Software 18:373–378

Woodcock CE, Strahler AH (1987) The factor of scale in remote sensing. Remote Sensing the Environment 21:311–332

Yu Q, Gong P, Clinton N, Biging G, Kelly M, Schirokauer D (2006) Object-based Detailed Vegetation Classification with Airborne High Spatial Resolution Remote Sensing Imagery. Photogrammetric Engineering & Remote Sensing 72(7):799–811

Zawadzki J, Cieszewski CJ, Zasada M, Lowe RC (2005) Applying geostatistics for investigations of forest ecosystems using remote sensing imagery. Silva Fennica 39(4):599–617

Sensitivity of the C-band SRTM DEM Vertical Accuracy to Terrain Characteristics and Spatial Resolution

Thierry Castel, Pascal Oettli

Centre de Recherches de Climatologie, UMR 5210 CNRS/Université de Bourgogne
6 Bd. Gabriel – 21000 Dijon, France
ENESAD, 26 Bd. Dr Petitjean – 21079 Dijon
Thierry.Castel@u-bourgogne.fr — Pascal.Oettli@u-bourgogne.fr

Summary. This work reports the results of a careful regional analysis of the SRTM DEM (Shuttle Radar Topography Mission – Digital Elevation Model) vertical accuracy as a function of both topography and Land-Use/Land Cover (LULC). Absolute vertical errors appear LULC-dependent, with some values greater than the stated accuracy of the SRTM dataset, mostly over forested areas. The results show that the structure of the errors is well modeled by a cosine power n of the local incidence angle (θ_{loc}). SRTM quality is further assessed using slope and topographical similarity indexes. The results show a lower relative accuracy on slope with a $R^2 = 0.5$ and a moderate agreement (Kappa $\simeq 0.4$) between SRTM- and IGN-derived slopes. The application of a simple cosine squared correction on the 90 m SRTM dataset shows only a slight improvement of the relative accuracy despite a 7 m decrease of the mean absolute elevation error. The accuracy is strongly improved ($R^2 = 0.93$ and Kappa $= 0.75$) for data resampled at a 150 m to 500 m horizontal resolution. These results support the idea that for regional application purposes the topographic correction as well as the spatial resampling of the SRTM dataset are needed.

Key words: DEM, analysis, error structure, modelling, Burgundy

1 Introduction

Digital Elevation Models (DEM) are valuable data for various earth science studies (climatology, hydrology, geomorphology, vegetation cover studies, etc.). The near-global high resolution DEM derived from Shuttle Radar Topography Mission (SRTM - van Zyl (2001); Rabus et al. (2003)) in 2000 is a useful product for such applications. The processed 3-arc-second (90 m) SRTM DEM is widely available over internet (e.g. on the website of Consultative Groups for International Agriculture Research Consortium for Spatial Information, CGIAR-CSI). Even if height accuracy is generally better than the 16 m of the mission specifications (Smith and Sandwell, 2003; Rodriguez et al., 2005; Berry et al., 2007), a close assessment of the accuracy of these data is needed . Indeed, recent studies clearly show the impact of relief

(Falorni et al., 2005; Gorokhovich and Voustianouk, 2006) as well as vegetation (Sun et al., 2003; Bourgine and Baghdadi, 2005) on the strong spatial variability of the vertical SRTM DEM accuracy.

Gorokhovich and Voustianouk (2006) show that the steepest ($> 10°$) slopes facing or away from the radar beam displays the highest vertical errors. Falorni et al. (2005) point out the unreliability of modeling shallow landslides with SRTM data for slopes steeper than 30°. Recent applications of the C-band SRTM DEM demonstrate that results are dramatically DEM-quality dependent (Molotch et al., 2005; Verstraeten, 2006; Racoviteanu et al., 2007). These results suggest that vertical accuracy of the SRTM DEM must be further investigated for various contrasted topographical and Land Use/Land Cover (LULC) situations.

We present the analysis of the vertical accuracy of the C-band SRTM DEM and its relation with the different topographic and LULC characteristics encountered over the region of Burgundy (France). The local incidence angle (θ_{loc}) which synthesizes the complex relative geometry between space-borne systems and terrain is used to account for the various topographic situations (Castel et al., 2001). Their quality (i.e. accuracy) assessment is based on the comparison of slope and aspect indexes computed from the IGN- and SRTM-DEM. These indexes are finally used to characterize the optimal horizontal resolution of the SRTM data to define the "best" compromise between SRTM-DEM quality and horizontal grid size. The dataset is processed for direct comparison within a Geographic Information System (GIS) framework.

2 Site and data sets

2.1 Burgundy area

Burgundy is located in the central part of eastern France. The region has an area of 31730 km^2 with a terrain elevation ranging from 50 m to 900 m. Mean elevation and slope are 284 m and 3.9 degrees, repectively. The eastern part of the area is formed by the Saône river-plain with lower altitudes from 50 m to 100 m. The topography changes markedly in the central part of the region with higher ground (up to 900 m) and steeper slopes (up to 44.7 degrees). This part is dominated by the Morvan mountains. The western part shows a smooth relief with mostly gently to moderately rollings hills. Main land uses are cropland, grassland and forest. Forested areas as well as grasslands are mainly located on steep sloped areas. Crops are found over flat areas and gentle hills. The land cover is dominated by cropland and grassland (65.6%) and by forested areas (30%). This illustrates the fact that Burgundy's lanscape is predominantly rural with urban area covering 3.2% of the land surface.

2.2 Data sets

Land cover data

The land cover and land use is derived from the Corine land cover map 2000 of the French Environment Agency (CLC2000, IFEN Paris, http://www.ifen.fr/clc/). The map has a spatial resolution of 100 m with a geometrical accuracy of 30 m. Land use categories are composed of 44 LULC classes. The classes are grouped from a radar point of view into seven major categories: urban, crop and grassland, deciduous, coniferous, mixedwood lands, shrubs and water. Due to complex interaction mechanisms between C-band microwave and vegetation, a first distinction is made between short/open and tall vegetation. Short and open vegetation are divided into two categories, crop-grassland and shrubs, for which a mixture of backscattering mechanisms (i.e. surface, volume and double-bounce) is assumed.

French National Geographical Institute (IGN) data

Both ground control points (GCPs) and DEM come from the french topographic database, IGN (Table 1, BD TOPO ©IGN). Due to its vertical precision (∼10 cm to 50 cm), the GCP dataset is considered as the ground truth to evaluate the vertical accuracy of the IGN-DEM. Unfortunately, the dataset could not be extended to the whole region. However the large number of GCPs (6100 or 8 points/km^2) permits to accurately estimate the vertical precision of the DEM, even if the GCP elevation data over the area range is limited to 50 m to 400 m. Horizontal gridsize is 50 m with a vertical accuracy that depends on the terrain characteristics.

Table 1. Summary of the properties of the Ground Control Points (GCPs) and Digital Elevation Model (DEM) data sets. Elevation and slope statistics present the mean ± one standard deviation

Data	GCP	IGN	IGN resampled	SRTM
Elevation (m)	192.7 (±16)	283.8 (±110)	283.8 (±110)	284.1 (±112)
Range (m)	50-400	50-900	51-900	39-909
Slope (Deg.)	-	3.9 (±3.8)	3.5 (±3.4)	3.6 (±3.3)
Range (Deg.)	-	0-44.7	0-35.8	0-34.8
Horizontal resolution (m)	8 points/km2	50	90	90
Vertical accuracy (m)	∼ 0.1 to 0.5	3	3	16
Dates	1999	1999	2007	Feb 2000

Shuttle Radar Topographic Mission (SRTM) data

In this study we used the post-processed 3-arc-second SRTM-DEM provided by the CGIAR-CSI (Jarvis et al., 2006). These data are for immediate use and offer a valuable resource for scientific purposes, particularly for regional climate analysis (Oettli and Camberlin, 2005). According to Rodriguez et al. (2005) C-band SAR-derived

DEM has a global absolute vertical accuracy of less than 16 m. However, high spatial variability of vertical accuracy is observed from continental (Berry et al., 2007) to regional (Gorokhovich and Voustianouk, 2006) scales. Hence the user community would gain a lot from additional regional assessments of the accuracy of this product.

3 Methods

3.1 Errors estimation

Towards the computation of the elevation difference, a bilinear interpolation algorithm was used to perform the resampling of the IGN-DEM to a 90 m resolution. Then for direct comparison purposes, the matching of IGN- and SRTM-DEM is performed with a simple linear affine transformation. The offset in the x- and y-direction is less than 30 m permitting a direct pixel-by-pixel computation. Hence the elevation errors are defined as:

$$\Delta E_{x,y} = E_{x,y}^{IGN} - E_{x,y}^{SRTM} \tag{1}$$

Where E denotes the elevation in meters. Figure 1a presents the map of errors computed for the region of Burgundy. We then seek to determine which probability density function better fits the errors distribution. Recent results (Berry et al., 2007; Racoviteanu et al., 2007) clearly show non-Gaussian errors distribution with asymmetry. However to our knowledge, no probability density function (pdf) has ever been proposed to model the error distribution of elevation. The range of the distributional characteristics found within and between topographical situations complicates assessment. For the robust estimation of first order statistics (mean, standard deviation, mode, etc.) a consistent theoretical pdf is needed to attain a specific accuracy and for analysis and modelling purposes.

3.2 Errors modelling

Based on the results presented by Castel et al. (1996, 2001) we choose to model the structure of errors as a function of the local incidence angle (θ_{loc}). The θ_{loc} is defined as the angle between the incident microwave vector $\boldsymbol{\imath}$ and the vector $\mathbf{n}$ normal to the surface given by:

$$\cos(\theta_{loc}) = \frac{-\boldsymbol{\imath} \,.\, \mathbf{n}}{|\boldsymbol{\imath}|\,|\mathbf{n}|} = \cos(\alpha)\,\cos(\theta)\,-\,\sin(\alpha)\,sin(\theta)\,cos(\beta-\phi) \tag{2}$$

The SRTM observation configuration (Rabus et al., 2003) is given by its incidence angle ($\theta = 54.5$ degrees) and viewing azimut angle ($\phi = 147$ degrees). The sloping terrain is characterized by the local slope α and the local aspect β angles. The terrain geometry (i.e. α and β) is computed from the resampled IGN-DEM. We assume that the distance between the two SRTM antenna does not imply geometrical differences. The analysis is conducted on the errors computed over the whole area and by LULC classes.

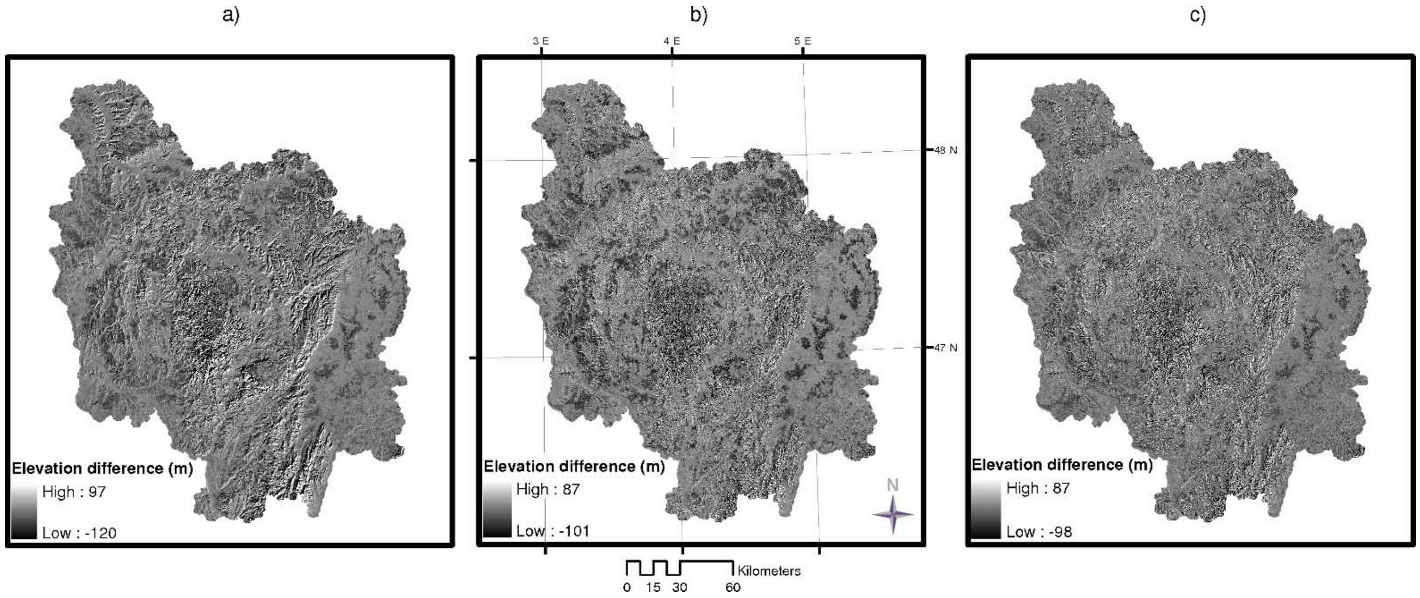


Fig. 1. Map of errors ($\Delta E_{x,y}$) over the whole Burgundy region a) no correction b) global cosine power n law correction c) LULC cosine power n law correction

3.3 Accuracy analysis

The analysis of accuracy is based on both slope and local similarity indexes (Oettli and Camberlin, 2007). The topographical similarity index accounts for both difference of elevation and variable orientation within a neighborhood window. The values range from -1 to 1. A value of -1 corresponds to an anti-similar configuration. A value of 0 may correspond to a flat area or to a transition zone between a ridge and a valley.

Accuracy assessment uses quantitative (i.e. usual R^2 coefficient) and qualitative (i.e. Kappa coefficient) criterions. The Kappa coefficient measures the overall agreement of the error matrix which takes into account all elements in the matrix and not only diagonal elements (Stehman, 1997). The value varies from 0 to 1, although there are no absolute cutoffs for Kappa coefficient. Landis and Koch (1977) quoted by Stehman (1997) suggest various cutoffs such as moderate (0.41-0.6), substantial (0.61-0.8), and almost perfect (0.81-1.0) agreement.

4 Results

4.1 Statistics of error

Results indicate a highly significant correlation ($R^2 = 0.985$ and $p < 0.00001$) between GCP and resampled-IGN elevation. As the mean difference is less than 0.54 m, the resampled-IGN DEM is considered in the following as the ground “truth” and it is used as the reference for direct error computation. A very high correlation ($R^2 = 0.988$) is observed between IGN- and SRTM-DEM elevation with a residual standard error close to 12 m. The errors display however a wide range from -120 m to +97 m (Figure 1a). A close inspection of the error histogram (Figure 2) shows an abrupt climb up and climb down around the mode with pronounced asymmetry. We found that the hyperbolic law (Scott and Haschenburger, 2005) provides a very good model for the distribution of errors whatever the LULC (Figure 2).

Table 2 summarizes mean, standard deviation and mode values derived from the fitted hyperbolic probability density function. Though a slight bias is observed for the whole area, the bias is mainly LULC-dependent. As expected SRTM elevation systematically overestimates IGN-elevation over forested area. This bias is particularly strong for coniferous forests as compared to deciduous or mixed forests. This may be explained by the fact that during the data acquisition (February) leafless deciduous areas permit a deeper penetration of the wave within the canopy (Wegmüller and Werner, 1995). This leads to a shift of the scattering center closer to the ground implying a lower elevation bias.

Crop and non-vegetated areas show lower positive (or negative) mean biases. These are associated with lower standard deviations. Note that for these areas, the mode absolute value is systematically positive and higher than the mean. This clearly denotes the asymmetry of the distribution. The largest standard deviations are observed for the forest with values close to 19 m. Therefore, for forest and to a lesser extent for shrubs and crop, supplementary sources of error are hypothesized. Following recent results of Falorni et al. (2005) and Gorokhovich and Voustianouk (2006)

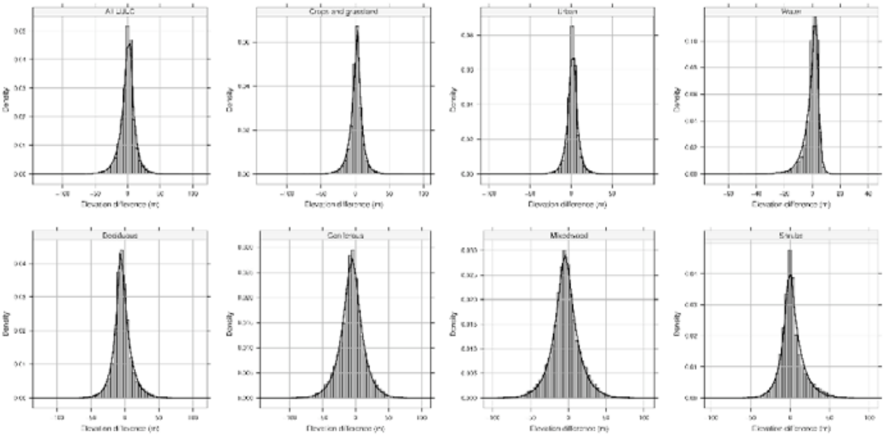


Fig. 2. Histogram showing the distribution of differences between resampled IGN data and the SRTM data. A negative heigth difference represents areas where SRTM data are higher than the IGN data. Black line is the fitted hyperbolic probability density function

Table 2. Summary of the pdf hyperbolic-based computation of the error statistics for the whole region and for the 7 LULC classes

LULC	Mean	Std. Deviation	Mode
All	−0.34	11.45	2.00
Deciduous	−4.00	13.78	−5.83
Coniferous	−6.13	18.79	−5.31
Mixed	−3.93	18.89	−4.45
Shrubs	2.40	14.81	−1.00
Crops	1.28	9.30	2.59
Urban	1.54	6.95	2.00
Water	−0.47	5.21	3.00

we have studied the influence of slope and aspect on SRTM data accuracy. From our knowledge, no generic framework has yet been proposed to model impacts of both slope and aspect.

4.2 Modeling of error structure

Angular behavior of the error is depicted on figure 3 for all LULC together. As θ_{loc} ranges from 37 to 80 degrees, the average error is computed by classes of 5 degrees. For each class, statistics are computed with at least 300 values. θ_{loc} lower than 54.5 indicates a tilted surface facing the radar, a θ_{loc} around 54.5 indicates a flat surface, while θ_{loc} greater than 54.5 indicates a tilted surface opposite to the radar. Results show a very good almost linear relationship between the error and θ_{loc}.

Overestimation (under estimation) are observed for tilted surface facing (opposite to) the radar beam. This partially explains the high dispersion observed on the error map (Figure 1a) and on the histograms (Figure 2).

The results demonstrate that the θ_{loc} is a relevant variable to model the structure of the error. This is confirmed by the error-θ_{loc} relationships observed for each LULC. Except for water areas, a strong angular error trend appears for the LULC scattered over various topographical locations. Hence, waterbodies may be seen as the reference. The corresponding error standard deviation does not exceed 5.5 m while it is systematically larger for the other LULC.

For interpretation purposes, and as supported by previous results (Castel et al., 2001), we postulate that a better adjustment can be obtained by a physical-based cosine power n law (equation 3).

$$\Delta E = a + b * \cos^n(\theta_{loc}) \tag{3}$$

Results show very good adjustments with equation 3 (Figure 3). Estimated values of model parameters are summarized in Table 3.

The power n values range from 0.87 for shrubs to 2.3 for crop and grassland with an average value for the whole area close to 2. For forest, the values of n show a lesser variability (1.65 to 1.7). Hence, n seems to be closely related to the LULC. The

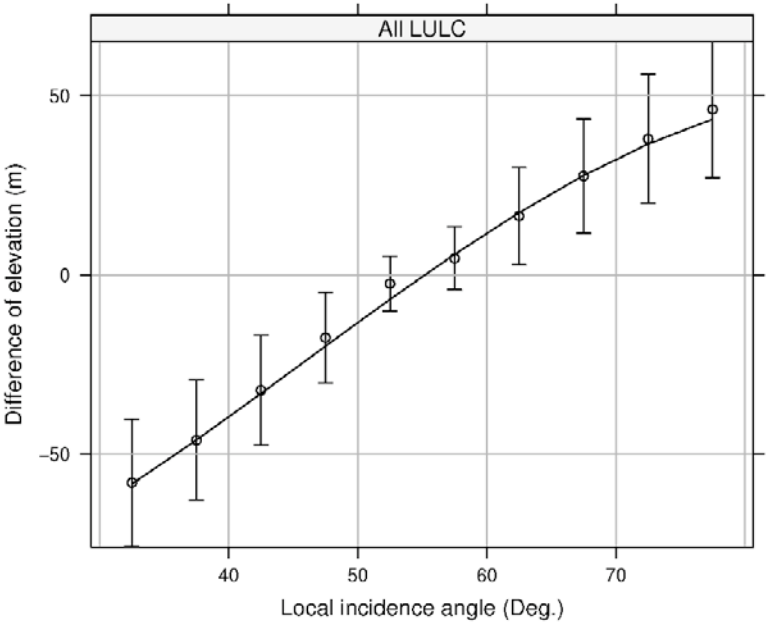


Fig. 3. Angular error behavior (IGN minus SRTM) for the whole area. Dot are computed mean (±1std. dev.) by angular classes of 5 degrees. The black line is the fitted cosine power n model

cosine power n law indicates that the observed error structure is driven by two main mechanisms. First, the cosine emphasizes the variability of the microwave energy over the surface. A tilted surface modulates the energy density which modifies the signal-to-noise ratio, and consequently the coherence magnitude affecting the quality of the interferogram (Ulander, 1996). Second, the power n underlines scattering mechanism changes due to the LULC properties. The results clearly demonstrate that the structure of the error is driven by a joint subtle mixture impact of the topographical and terrain cover characteristics.

One question that arises concerns the model ability to improve the CGIAR-CSI SRTM accuracy. To this aim we applied the model in two levels accuracy.

Table 3. Summary of the fitted regression parameters. Stars indicate rejection of H_0 hypothesis

LULC	a	b	n	$H_0(n = 1)$
All	51.34	−152.06	1.98	***
Deciduous	56.94	−148.42	1.66	***
Coniferous	56.42	−154.50	1.71	***
Mixedwood	57.77	−149.63	1.67	***
Shrubs	107.24	−171.28	0.87	-
Crops and grassland	47.45	−159.79	2.28	***
Urban	56.92	−122.50	1.44	-

4.3 Accuracy assessment

Figure 1 presents the error maps derived from equation 1 computed with uncorrected, corrected and LULC-corrected SRTM data. The maps show a decrease of the error range with a general homogenization. Figure 4 illustrates the error behavior as a function of both LULC and level of correction. The correction based on equation 3 improves the elevation accuracy of CGIAR-CSI SRTM data by reducing the random noise and removing the bias. Despite the application of the correction, coniferous and mixed woodlands show higher noise compared for instance to deciduous forests. Such difference may originate from the volume scattering mechanism that dominates at an incidence angle of 54.5 degrees for dense evergreen vegetation. Complex multiple scattering implies a low signal-to-noise ratio that affects the degree of coherence, and consequently the quality of interferogram phase. To a lesser extent surface and double bounce mechanism may apply to crop lands, grasslands and urban areas but with less effects. Waterbodies that are topography free only have one surface mechanism, and the observed error (around 5 m) is in the range of the random noise given by Rodriguez et al. (2005).

So, the observed error may be seen as the combination of three main sources of variability: scattering mechanism, slope and random noise. Assuming that random noise is constant, the effect of the correction reduces error due to scattering mechanism and slope from 13 m to 6 m. However, IGN-derived slope compared to the

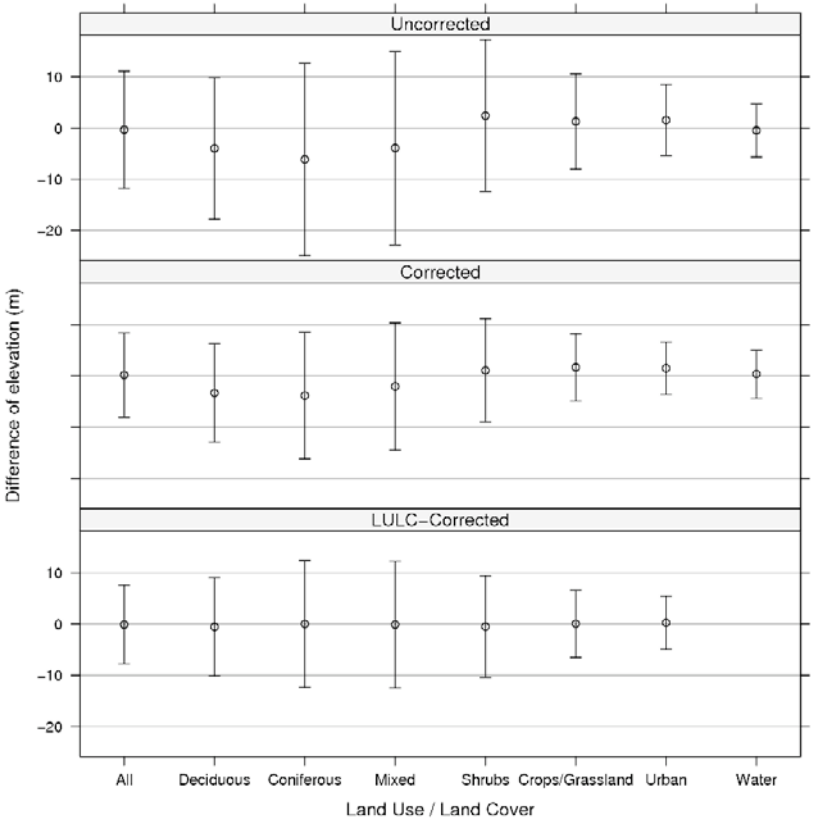


Fig. 4. Elevation accuracy (mean ±1 standard deviation) of the CGIAR-CSI SRTM data as function of LULC and for the three levels of correction.

SRTM-derived slope gives R^2 values of 0.51, 0.59 and 0.59 for uncorrected, corrected and LULC-corrected SRTM data respectively (see Figure 5 for results at a resolution of 90 m). These are much lower correlations than for elevation. Though the correction improves the R^2, these results indicate that the relative vertical accuracy is worse compared to the global absolute elevation accuracy. This is supported by the values of the Kappa coefficient (Figure 5) that indicate a moderate agreement whatever the level of correction. The question is therefore whether 90 m is the suitable resolution towards applications or whether there exists another spatial resolution that optimizes the accuracy-to-resolution ratio.

4.4 Resolution impact on accuracy

Figure 5 shows the results of R^2 and Kappa between the slope derived from IGN and SRTM data, for the three levels of correction and for the two main LULC (crop/grassland and forest). As expected increasing grid size improves the SRTM-data accuracy with a decrease of the slope range.

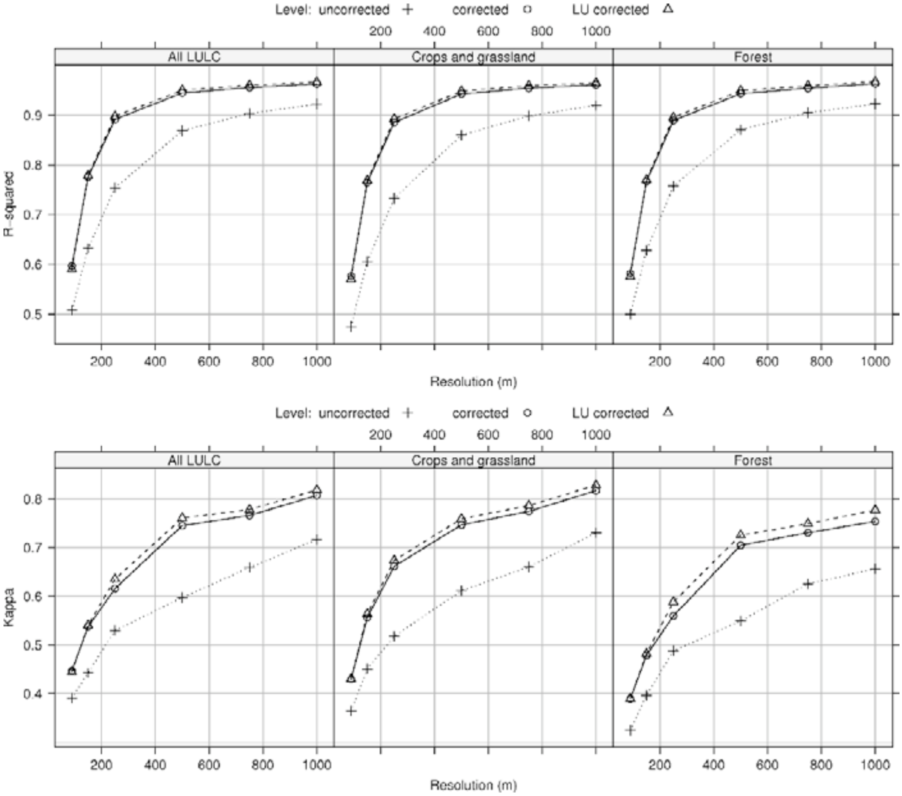


Fig. 5. R^2 (top) and Kappa (bottom) between IGN- and SRTM-derived slope vs. the horizontal grid size (i.e. resolution) for the three levels of correction and for forested and crop/grassland areas. Note that the computation of the Kappa coefficient is based on 10 classes. Forest includes deciduous, coniferous and mixedwood area

This “smoothing” effect follows a non-linear trend with higher accuracy improvement between 90 m and 500 m. After 500 m the gradient of the trend is strongly reduced for all the LULC. Concerning the R^2 a spectacular improvement is observed for the corrected SRTM data in particular between 90 m and 250 m. At 250 m the R^2 value is equal to that of the uncorrected SRTM data at 750 m. Therefore, the above "correction" permits an important gain on the SRTM accuracy. This effect seems independent of the LULC.

The Kappa coefficient generally show the same type of behavior. However, differences can be noted as a function of LULC and to a lesser extent the level of correction. Again, the correction effect is significant and a substantial agreement cutoff is reached at the 250 m resolution. The correction is more effective for croplands and grasslands, compared to forest for which the agreement is systematically lower.

The improvement is not homogeneous and depends on LULC and on the slope value. This is supported by the analysis of the topographical similarity index computed at the resolution of 250 m (Figure 6). A contrasted behavior is observed for the terrain configurations along (i.e. parallel) and facing (i.e. perpendicular) the radar beam. Concerning the parallel case, strong difference is shown between valleys and ridges. Valleys (negative values of the topographical similarity index) have a higher Kappa coefficient than ridges which may be explained by smoother altitudinal gradients and by LULC differences. Valleys have a more concave curvature than ridges implying less topographical impacts. The perpendicular orientation follows a quasi opposite behavior compared to the parallel case. The topographical similarity index is higher for flat and mid-slope areas than for valleys and ridges. Again, the topographical similarity index point out that the accuracy is strongly terrain characteristic-dependent.

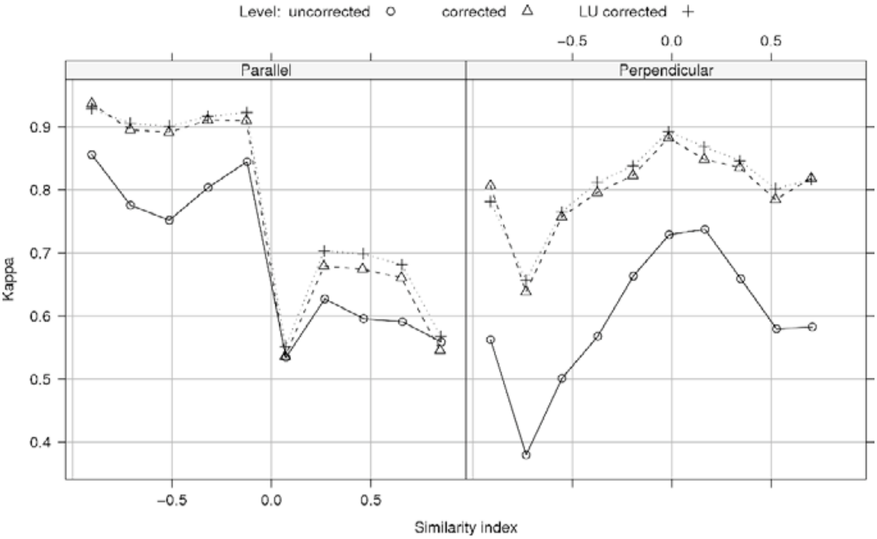


Fig. 6. Kappa coefficient for 10 classes of the topographical similarity index. Computation is made at a 250 m resolution for the three levels of correction.

5 Conclusion

In this study, we have carefully examined the vertical accuracy of the C-band SRTM DEM and its relation with the different topographic and LULC characteristics. After confidence in IGN data was established, the SRTM height was compared with the surface height from IGN. The comparison shows that the whole quality of the SRTM data is very significant ($R^2 > 0.98$). However, at some locations large differences appear with absolute errors that are greater than the stated accuracy of the SRTM dataset. Error patterns are LULC-dependent. For the open and short vegetation areas, accuracy exceeds SRTM specification. On the contrary, forest-covered areas show the lowest accuracy with higher errors bias and non-homogeneous noise (i.e. variance). We found that the structure of the errors are well fitted by a cosine power n law of the local incidence angle. This demonstrates the effects of both topography and LULC on the vertical accuracy of the SRTM data. The results demonstrate that the accuracy gain is optimal for horizontal resolutions between 150 m and 500 m, compare to the uncorrected dataset. The results suggest that a simple angular correction based on a cosine squared is sufficient to significantly improve the accuracy. The analysis of the similarity index for various topographical patterns, reveals that the improvement of the relative accuracy is not homogeneous. The greatest (smallest) improvement is found for valleys (ridges) parallel to the radar beam. This is of interest in order to map the accuracy of the SRTM dataset.

This work supports the idea that for application purposes topographic correction and resampling are needed.

Acknowledgements

The author would like to thanks Pierre Camberlin from CRC and Sarah Strano from the University of Pittsburgh for theirs comments and corrections.

References

Berry PAM, Garlick JD, and Smith RG (2007) Near-global validation of the SRTM DEM using satellite radar altimetry. Remote Sensing Environment, 106:17–27.

Bourgine B and Baghdadi N (2005) Assessment of C-band SRTM DEM in dense equatorial forest zone. C. R. Geoscience, 337:1225–1234.

Castel T, Beaudoin A, Stach N, Souyris JC, and Le Toan T (1996) Sensitivity of polarimetric SIR-C SAR data to forest parameters over hilly terrain : a case study on Austrian pine. In *PIERS 1996*, page 208. Innsbruck, Autriche.

Castel T, Beaudoin A, Stach N, Stussi N, Le Toan T, and Durand P (2001) Sensitivity of spaceborne SAR data to forest parameters over sloping terrain. Theory and experiment. International Journal of Remote Sensing, 22(12):2351–2376.

Falorni G, Teles V, Vivoni ER, Bras RL, and Amaratunga KS (2005) Analysis and characterization of the vertical accuracy of digital elevation models from the Shuttle Radar Topography Mission. Jounal of Geophysical Research, 110(F02005). Doi:10.1029/2003JF000113.

Gorokhovich Y and Voustianouk A (2006) Accuracy assessment of the processed SRTM-based elevation data by CGIAR using field data from USA and Thailand and its relation to the terrain characteristics. Remote Sensing of Environment, 104:409–415.

Jarvis A, Reuter HI, Nelson A, and Guevara E (2006) *Hole-filled SRTM for the globe Version 3*. CGIAR-CSI.

Landis JR and Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics, 33:159–174.

Molotch NP, Colee MT, Bales RC, and Dozier J (2005) Estimating the spatial distribution of snow water equivalent in an alpine basin using binary regression tree models: the impact of digital elevation data and independent variable selection. Hydrol. Process., 19:1459–1479.

Oettli P and Camberlin P (2005) Influence of topography on montly rainfall distribution over East Africa. Clim. Res., 28:199–212.

Oettli P and Camberlin P (2007) Development of a topographical similarity index to analyze spatial distribution of precipitation over Africa. Personal communication.

Rabus B, Eineder M, Roth A, and Bamler R (2003) The shuttle radar topography mission - a new class of digital elevation models acquired by spaceborne radar. ISPRS Journal of Photogrammetry & Remote Sensing, 57:241–262.

Racoviteanu AE, Manley WF, Arnaud Y, and Williams MW (2007) Evaluating digital elevation models for glaciologic applications: An example from Nevado Coropuna, Peruvian Andes. Global and Planetary Change, doi:10.1016/j.gloplacha.2006.11.036.

Rodriguez E, Morris CS, Belz JE, Chapin EC, Martin JM, Daffer W, and Hensley S (2005) An assessment of the SRTM topographic products. Technical Report D-31639, JPL/NASA. 143 pp.

Scott DJ and Haschenburger JK (2005) Using the hyperbolic distribution to estimate the precision of size distribution percentiles of fluvial gravels. Computers & Geosciences, 31(10):1224–1233.

Smith B and Sandwell D (2003) Accuracy and resolution of shuttle radar topography mission data. Geophysical Research Letters, 30(9):20.1–20.4. Doi:10.1029/2002GL016643.

Stehman SV (1997) Selecting and interpreting measures of thematic classification accuracy. Remote Sensing of Environment, 62:77–89.

Sun G, Ranson KJ, Kharuk VI, and Kovacs K (2003) Validation of surface height from shuttle radar topography mission using shuttle laser altimeter. Remote Sensing Environment, 88:401–411.

Ulander LMH (1996) Radiometric slope correction of synthetic-aperture radar images. I.E.E.E. Transactions on Geoscience and Remote Sensing, 34(5):1115–1122.

van Zyl JJ (2001) The Shuttle Radar Topography Mission (SRTM): a breakthrough in remote sensing of topography. Acta Astronautica, 48(5-12):559–565.

Verstraeten G (2006) Regional scale modelling of hillslope sediment delivery with SRTM elevation data. Geomorphology, 81:128–140.

Wegmüller U and Werner C (1995) SAR interferometric signatures of forest. I.E.E.E. Transactions on Geoscience and Remote Sensing, 33(5):1153–1161.

Improving the Reusability of Spatiotemporal Simulation Models: Using MDE to Implement Cellular Automata

Falko Theisselmann[1,2], Doris Dransch[2]

[1] Graduiertenkolleg METRIK, Department of Computer Science, Humboldt-Universität zu Berlin
[2] GeoForschungsZentrum Potsdam (GFZ), Telegrafenberg, Potsdam

Abstract

Numerous modeling and simulation tools, frameworks, and environments support domain experts with the implementation of spatiotemporal simulation models. The implemented models are usually bound to specific tools, because specific modeling languages, simulation engines, or processing platforms have to be used. To improve model reusability, we propose an implementation approach that applies Model Driven Engineering (MDE). In this approach, a simulation model is described on three different levels of abstraction. Starting from an abstract description of a simulation model by the modeler, this model is automatically transformed through all levels into executable code. In contrast to common implementation technologies, the intermediate steps of the transformation are clearly and formally defined by metamodels. For model execution, existing general purpose simulation and spatial data processing frameworks may be used. In this paper, the three-level approach and its application to the modeling of cellular automata are described. Partial metamodels and transformations are presented for two of the three levels. The MDE-approach provides means to enhance model reusability and promotes transparency in simulation modeling. Moreover, a tight integration of simulation and spatial data processing can be realized by synthesizing executable software which is composed of generic spatial data processing and simulation functionality.

Keywords: Spatiotemporal modeling; Model Driven Engineering; Cellular automata

1 Introduction

Spatiotemporal simulation models are widely used to model the spatiotemporal behavior of environmental phenomena. An important characteristic of spatiotemporal simulation modeling is that models are executed in order to observe the spatiotemporal behavior of phenomena. Statements about the original system are derived from these observations. Usually, model execution is realized by the means of software, thus the simulation software is a key artifact in the simulation modeling approach.

The reuse of simulation models could reduce implementation effort and provide the means to share models. By this, the exchange of models between scientists and the transfer of scientific models to application domains, like disaster management or urban planning, can be supported.

To improve the reusability of spatiotemporal simulation models, we propose an implementation approach that applies Model Driven Engineering (MDE) methods and tools. Based on a high-level description of a model, an executable model is automatically generated. This model includes simulation and spatial data processing functionality.

In this paper, our approach is detailed for spatiotemporal modeling with cellular automata (CA). The next section provides an overview of common simulation model implementation approaches with the focus on simulation model reuse and software implementation. This is followed by the introduction of the concept of our three-level approach. After this, two levels of the three-level approach presented in more detail for modeling CA. For modeling CA on one presented level, we developed a generic formalism: the hybrid cellular automaton (HCA), which is completed by means to model access to spatial data. The second presented level is represented by a simulation framework (jDisco) and a library that provides spatial data processing functionality (Geotools). Geotools and jDisco are used to produce executable models. These technologies exemplify how executable models can be synthesized from a formal abstract description that is based on generic functionality. Moreover, simulation and spatial data processing can be integrated using this approach. The paper concludes with remarks and outlook.

2 Implementation Technologies and Approaches to Spatiotemporal Modeling

From an engineering point of view, the implementation of spatiotemporal simulation models requires the realization of simulation and spatial data processing functionality. Fig. 1 shows the tasks that we assume to be processed for a spatiotemporal simulation. Spatial data is used to provide the parameters for the simulation model. For this, it may be necessary to preprocess spatial data. Simulation processing is realized through the iterative calculation of the state of the model. These calculations may require simulation functionality (i.e. execution of transition, synchronization of submodels, data exchange) and spatial data processing. Depending on the modeler's needs, spatial data may need to be stored, processed and visualized before, during, and after a simulation run.

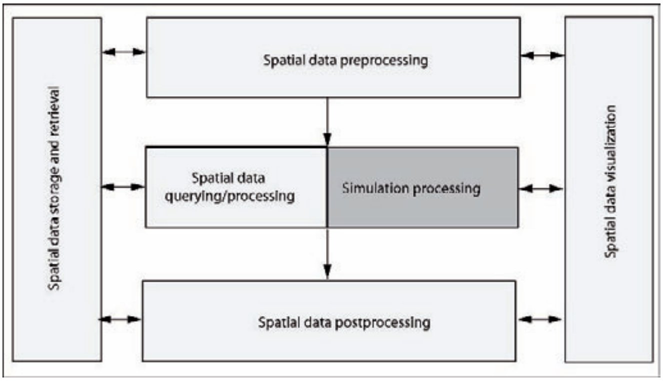

Spatial data storage and retrieval
Spatial data preprocessing
Spatial data querying/processing
Simulation processing
Spatial data postprocessing
Spatial data visualization

Fig. 1. Spatial data processing and spatiotemporal simulation. Spatial data handling (light grey) and simulation functionality (dark grey) has to be integrated

If spatial data processing functionality is needed before and after a simulation run, simulation and spatial data processing is executed serially. Functionalities are executed in parallel, if both functionalities are needed during simulation. Either way, the required functionality may be provided by an integrated modeling and simulation framework, or by distinct, frameworks (e.g. separate GIS and simulation frameworks)[1]. Serial execution and

[1] In the remainder of this paper the term *framework* will be used for any software that provides the functionality to model and simulate dynamic models. This subsumes software that is commonly named simulation library, tool, framework, language, or environment.

the coupling of respective frameworks, is mainly relevant for pre- and postprocessing of spatial data. This may be associated with tasks like georefencing, resampling, and filtering. In this case, data exchange between the frameworks may simply be realized by manual export and import of data. In the case of parallel execution of functionality, the integration of frameworks at runtime is required, where data exchange may be realized via shared memory or network based exchange mechanisms. Runtime integration can be used to read and write spatial data during simulation. This permits to persist intermediate results or to read subsets of spatial data selectively, depending on the state of a model at runtime. In this paper, the integration of simulation and spatial data processing functionality at runtime is relevant.

In the following, a brief overview of a selection of implementation approaches is given, focusing on implementation languages and model reuse. It is common practice to implement CA models using general purpose programming languages, but there are disadvantages, such as limited reusability and high implementation costs of the resulting simulation software. In the past, this led to the development of special purpose tools, frameworks, languages, and environments.

One approach to provide means of modeling and simulation is the development of high level *domain specific modeling languages* (DSL) and respective execution frameworks. As such, SELES (Fall and Fall 2001) and PCRaster (Karssenberg 2002) provide modeling languages to model spatiotemporal processes, based on a cellular discretization of space. These languages contain simulation functionality, including the calculation of state transitions, data input, output, and visualization, predefined functions (i.e. statistic, stochastic, algebraic) and additional simulation functionality, such as batch simulation (SELES 1999, Karssenberg 2002). Within these frameworks, modeling, simulation and spatial data processing functionality is highly integrated. Model reuse is possible within the respective frameworks, but the reuse of models in other frameworks is limited.

The idea of *component based modeling* is the core concept of numerous simulation frameworks which focus on model reuse and integration (Argent 2004). In component based modeling, the system under study is decomposed into connected interacting subsystems which are modeled as components. Although component based frameworks share this key idea, there are differences with respect to their architecture, the modeling concepts, interfaces, and the underlying technologies. Due to this heterogeneity, the reuse of components and models with different frameworks is difficult (Argent 2004).

One way to reuse models across framework boundaries is to use a declarative modeling language as an exchange format for models between frameworks (Argent 2004). However, a generic formalism may be hard to use for domain experts, so that single models may be modeled with the use of a more adequate formalism and consequently be transformed to this generic formalism. With such approach named *multi-paradigm modeling*, Vangheluwe et al (2002) show that by using a common generic and expressive modeling formalism, also the integration of a variety of models is possible.

In an analysis of today's modeling and simulation frameworks for environmental modeling, Evert et al (2005) classify environmental modeling and simulation frameworks as *modeling-level* or *implementation-level* frameworks. Modeling-level frameworks provide domain specific abstractions to domain experts for the definition of models. With implementation-level frameworks, existing models are linked and executed. It is argued that the differences between implementation-level frameworks are relatively unimportant, so that code generators of modeling level frameworks could target a widely accepted implementation-level framework, based on a generic modeling formalism, such as DEVS (Zeigler et al 2000).

Our MDE-approach to spatiotemporal simulation model implementation is conceptually based on this suggestion of Evert et al (2005): Domain specific models, as specified by the modeler, are automatically transformed to executable code with well defined intermediate representations. But, instead of targeting a specific simulation level modeling formalism, we suggest to focus on generic simulation functionality that may be implemented by several frameworks, possibly conforming to different modeling formalisms. This facilitates the storage of models in a more independent way, allowing for more framework independence. Moreover, we integrate spatial data processing based on generic functionality that also may be provided by different frameworks in different application scenarios. Model Driven Engineering provides the tool support for the efficient realization of this approach.

3 A Three-level Model Driven Engineering Approach to Spatiotemporal Modeling

MDE is a software engineering approach to software development. The core idea of MDE is that software is mainly described by models, not by code. Executable code is automatically synthesized from these models (Schmidt 2006). Different models describe the same software on different levels of abstraction, which may be distinct in the amount of implementation detail they encompass. By the provision of domain specific modeling concepts on the most abstract level, this approach promises to hide implementation detail from the domain expert, thus it helps to unburden the modeler from the need of detailed implementation (Muzy et al 2005). Moreover, an MDE-based approach facilitates platform independence, thus this approach enhances the possibility for the reuse of models on different platforms (Schmidt 2006).

The realization of MDE is based on a clear, formal definition of the levels of abstraction and the relationship between them. The different levels of abstraction are formalized by respective modeling languages. Transformations formalize the relationship between these languages. MDE technologies provide the necessary tool support to define modeling languages, model transformations, and code generators, which analyze and synthesize models and code (Schmidt 2006).

Key artifacts within this approach are metamodels, since they are used to define the modeling elements of the modeling languages and relationships between the modeling elements. All models must conform to metamodels, thus metamodels prescribe the set of possible models at each modeling level. For example, a metamodel may prescribe that a model may contain variables and that variables must have a name and a type. The transformations between models are defined on the basis of their metamodels.

The meaning of metamodels and the respective models is twofold. On the one hand the meaning of a model, or a model element, is given by the modeler, and the underlying understanding of the modeling elements and their relations, i.e. a variable may describe a model parameter or a state. On the other hand, as models are processed by the computer, meaning is given by model transformations that finally result in model execution, i.e. a variable is referenced memory that holds a value of a specific type.

We propose an application of MDE for spatiotemporal modeling, which is detailed for spatiotemporal modeling with CA. In this approach, a CA simulation software is explicitly modeled on three distinct levels of abstraction. At each level, the simulation model is represented by a level-specific model. The concepts that are available for modeling on the different levels are formally prescribed by level-specific metamodels.

Fig. 1 shows the concept of the three-level approach to implement of spatiotemporal simulation models. On the highest level, a *domain specific model* is defined by the domain expert using a domain specific modeling language (DSL). This DSL should be particularly tailored to the needs of the modeler and should provide modeling concepts with a clear relation to the problem domain (i.e. hydrological modeling, fire spread modeling). A domain specific model is automatically transformed to a *formalism specific model*.

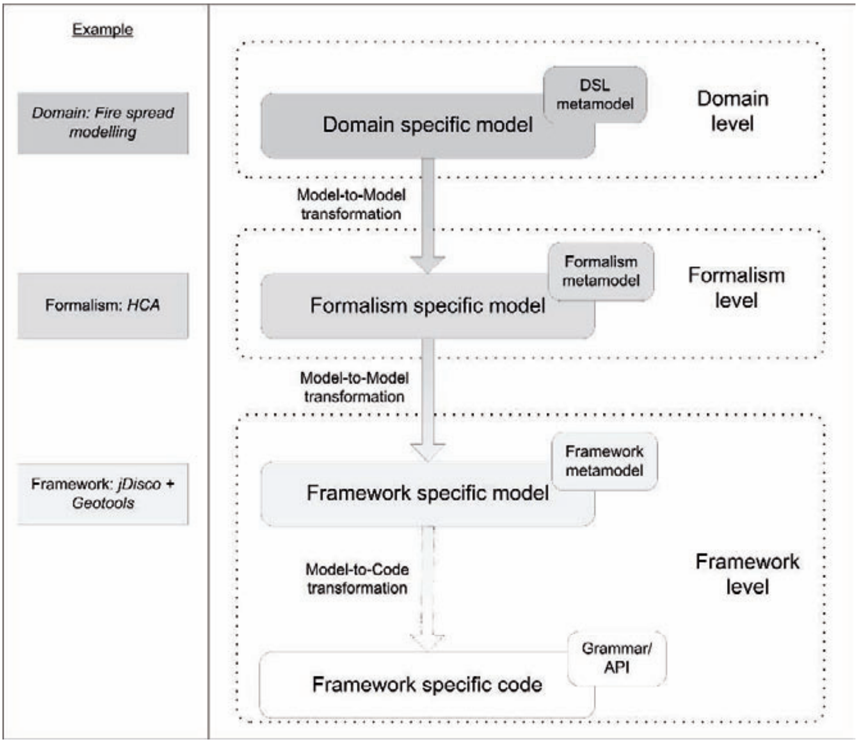


Fig. 2. A three-level, MDE-based implementation approach to simulation modeling

The formalism specific model is modeled using the concepts of a more generic modeling formalism in comparison to the DSL. For modeling on this level, we developed the HCA formalism as a generic modeling language for CA models (see Sec. 4). This level has two main characteristics. On the one hand, the formalism is generic, so that models of different domains may be expressed by the means of this language. On the other hand the formalism level is based on the functionality provided by the executing frameworks, since modeling concepts on the formalism level that can not be realized by frameworks do not make sense in simulation modeling.

A formalism specific model is automatically transformed to a *framework specific model* from which executable code is automatically generated by a code generator. A feature of our approach is to use an existing general purpose simulation framework to execute simulation models. Consequently, the concepts of the modeling language at the framework level are provided by the respective simulation framework and are derived from the API and documentation. Since the approach is based on widely used simulation modeling concepts (see section 4.1), one may find a number of simulation frameworks that realize the required functionality.

The inclusion of spatial data processing functionality is based on the same principle. For this, it is assumed that there are widely accepted concepts related to the modeling of spatial data and spatial data processing. A common ground can be seen in widely used standards i.e. ISO and OGC standards. Software that provides implementation of these standardized concepts is to be combined with simulation frameworks.

In general, two application scenarios for this approach are possible. In the first scenario, several DSLs may be defined for different domains, but modeling uses the same modeling formalism on the formalism-level, i.e. HCA. For each DSL, a transformation to the formalism level is defined individually; an existing transformation from the formalism level to the framework level and code generation can be reused for all DSLs. In the second application scenario, existing models on the formalism level are executed by different frameworks, for example, if the model is reused in another application context, with a different simulation and spatial data processing infrastructure. For this, a new transformation from the formalism level to the framework level and code generation must be defined. The domain specific models, the formalism specific models, and the transformation from the domain level to the formalism level remain unchanged and can be reused in this scenario.

4 A Three-level MDE Approach to Model Cellular Automata

In this section, two levels of our approach, the formalism level and the framework level, are detailed for modeling cellular automata, as an example for the presented generic approach. The main modeling elements on the two levels are presented by the means of (partial) metamodels. An illustration of the relationship between the modeling levels concludes this section. A detailed presentation of the domain level is beyond the scope of this paper, but the two levels are sufficient to illustrate the main characteristics of the approach.

4.1 Common Modeling Concepts on the Formalism and the Framework Level

Basically, we adopt the notion of event-based modeling. A model has a state that changes at times of events. This discrete behavior is extended with the possibility to model continuously changing states. Moreover, a model can be composed of several coupled submodels. The state of a simulation model is modeled by state variables. For modeling state variables, standard data types (i.e. integer, float, string) and standard data structures (i.e. enumeration, array) are used.

Discrete state changes are modeled by discrete transition functions with the means of standard arithmetic expressions that calculate new values that are assigned to state variables. Continuous state changes are modeled by means of ordinary differential equations (ODEs). During a simulation run, ODEs are evaluated by the simulation framework to approximate state changes over time using an integration algorithm.

State changes depend on the value of state variables at certain points in simulation time. This dependency is modeled by conditions. A condition is an expression with which a model's state is evaluated: if the model is in a state that is specified by a condition, the respective condition is true; false otherwise. Within a condition, the model's state variables and input may be evaluated. Conditions are specified by means of standard relational, logical and arithmetic expressions.

Conditions are attached to discrete transition functions and ODEs in order to specify in which state which particular transition is to be executed. In addition, conditions are used to describe conditions on states, which trigger an event (state event). Events interrupt the continuous behavior of a model. Discrete transitions or communication with other models happen at times of events. Events may occur at arbitrary times, thus the time base of models is continuous.

Raster data provides the spatial parameters of a CA model and is used to store the spatial output of a model. Fig. 3 shows how the access to raster datasets residing in the file system can be metamodeled. In the example, a raster dataset is a *RasterDataFile* that can be of type *RasterDataASCIIFile*, *RasterDataWorldImageFile*, or *RasterDataGeoTiffFile*. To access the raster dataset, it is sufficient to provide the type and the location of the dataset (URL).

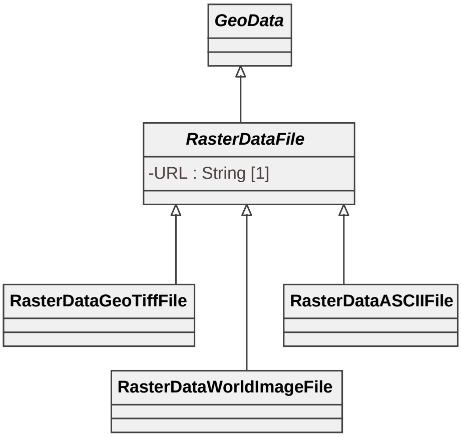


Fig. 3. A partial metamodel for modeling access to raster data (UML notation)

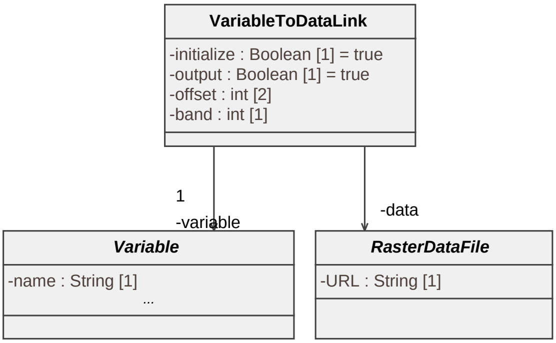


Fig. 4. A partial metamodel showing how the link between model variables and datasets is modeled (UML notation)

For the use of raster data in the simulation model, raster data is linked to a variable of a model. For this, a *VariableToDataLink* is specified, that holds a reference to a variable and the corresponding dataset (Fig. 4).

The *variable* in a *VariableToDataLink* references the *Variable* elements within the metamodels of the modeling formalism with which dynamic behavior is modeled (see Figs 6 and 8 below). The excerpt of the metamodel in Fig. 4 shows, that additional information can be given by means of attributes, e.g. if the data should be used to initialize the variable or if the dataset is used for storing output[2]. This scheme is likewise applied on the formalism and the framework level, which is presented in the following.

4.2 Modeling Cellular Automata on the Formalism Level: The Hybrid Cellular Automaton Formalism (HCA)

For modeling on the formalism level, we extended the classical CA with the means to model continuous behavior of cells in order to be able to express greater variety of CA and to exploit the possibilities of modern simulation frameworks. This enhances expressiveness and may lead to a greater precision of models (for examples see Yacoubi et al 2003, Wainer and Giambiasi 2005).

A HCA consist of a set of structurally equivalent cells. Fig. 5 illustrates the possible behavior of single cells.

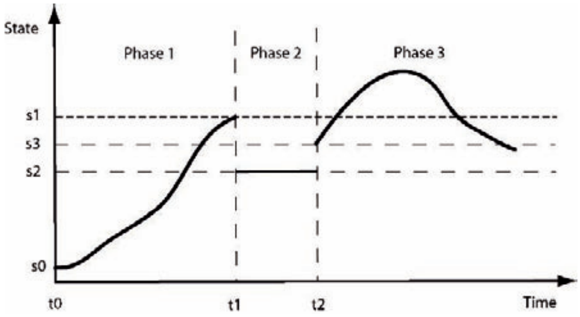


Fig. 5. The hybrid behavior of single cells of the HCA

Starting from state *s0* at time *t0,* the cell's state evolves continuously until time *t1*. At time *t1* a state event occurs as the state reaches the threshold *s1*. The state changes instantaneously to state *s2* and the cell transits into phase 2, where the state remains constant at *s2*. At time *t2* the state changes to state *s3* and evolves continuously in phase 3. The continuous behavior of cells is qualitatively different in the different phases, i.e. it is described by different ODEs. Note that the state change at time *t2* is triggered by an event outside the cell, for example a neighboring cell.

The metamodel in Fig. 6 shows the basic modeling elements of HCA. A cell's state is modeled by variables (*discreteStateVariable*, *continuousStateVariable*). All cells have an identical set of variables. For each phase, an

[2] For simplicity, it is assumed that a dataset has one band.

ODE (*rateOfChangeFunction*) can be specified. The selection of the appropriate ODE for a phase is modeled by a phase selection function (*phaseSelectionFunction*) which contains respective conditions (*phaseCondition*). Discrete state changes are specified with the *discreteTransitionFunction* that holds conditions (*condition*) and single expressions (*expression*) which define the state change. Discrete state changes depend on the state of the cell and the neighboring cells. State events (*stateEvent*) are modeled by means of conditions within *eventState*.

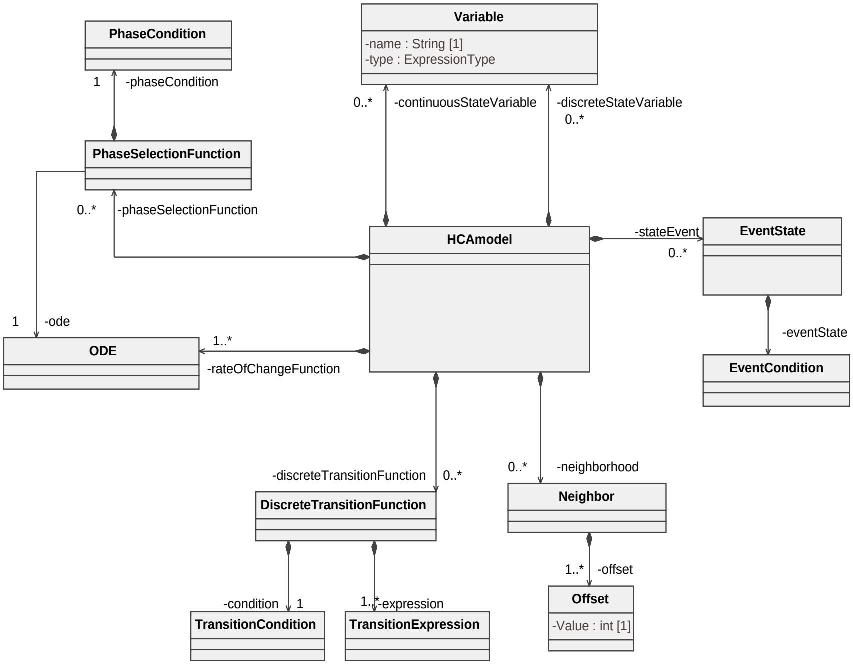


Fig. 6. A partial metamodel for the HCA (UML notation)

Fig. 6 informally illustrates the operation of an HCA model. Simulation time passes only during phases of continuous state change (*rateOfChangeFunction*). This evolution of state is eventually interrupted by an event. An event is followed by the application of the discrete transition function in each cell and the application of the *phaseSelectionFunction*. After this, the state of the cells evolves continuously until the next event or the end of the simulation.

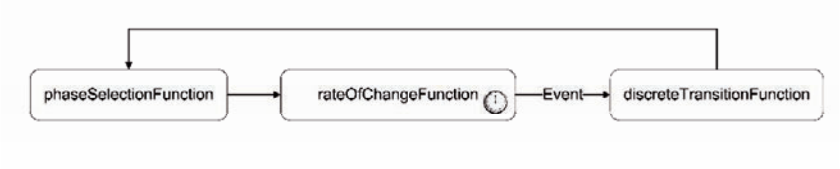


Fig. 7. The execution of HCA cell models

An event occurs when a cell’s state variables have values as specified by a condition (*stateEvent*). The *neighborhood* of a cell is modeled by offsets from the origin of a cell. By definition, dependencies between the neighboring cells, as expressed in the *discreteTransitionFunction*, have to be reflected in the state events of the respective neighboring cells.

4.3 Modeling on the Framework Level: jDisco and Geotools

HCA models are transformed to framework specific models and finally to code (see Fig. 2). On the framework level a HCA is modeled by the means of the simulation framework jDisco and Geotools. Geotools is a library that provides spatial data processing functionality (Custer 2006).

JDisco is a framework that implements the process simulation world view (Helsgaun 2001). In process based modeling, a system is modeled as a collection of interacting processes that compete for common resources. In jDisco, a process can be a continuous (*ContinuousProcess*) or a discrete process (*DiscreteProcess*, see Fig. 8). Conceptually, a *ContinuousProcess* can change the state continuously between events according to ODE (*derivatives*).

Each process has a state which is defined by a set of state variables (*continuousState*, *discreteState*). Processes may hold references to other processes (*partnerProccess*) and other processes’ variables (*referencedVariable*). Via references, values of variables can be read and set, which enables interaction and communication between processes. The behavior of a *DiscreteProcess* is modeled within a lifecycle function (*actions*) which is a sequence of actions (*ActionExpression*). A common action is a discrete change of the value of variables (*StateChangeExpression*).

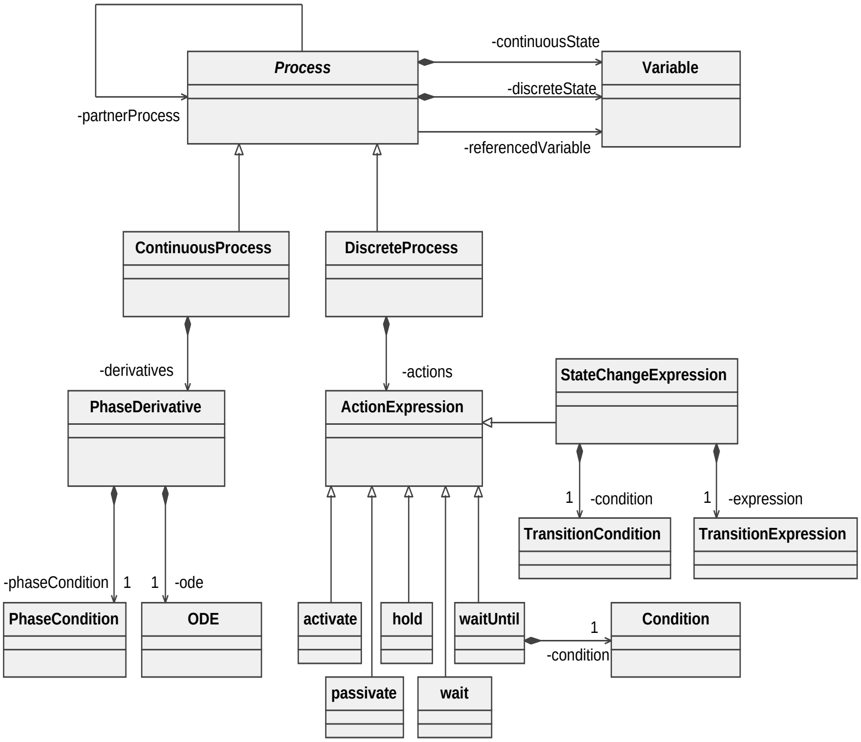


Fig. 8. A partial metamodel for jDisco (UML notation)

To enable interaction, processes must be synchronized. In jDisco, a *DiscreteProcess* can synchronize with a *ContinuousProcess* via the *waitUntil* – action. When performing *waitUntil*, the synchronizing *DiscreteProcess* suspends its lifecycle function until an event condition (*condition*) is true. Within condition, state events are specified. After an event, the lifecycle function (*actions*) continues performing actions.

Geotools is a Java language code library which provides the means to implement standards conformant geospatial applications. The library consists of various modules which can be combined according to the programmer's needs (Custer 2006). Geotools provides classes to access and process spatial data, which is needed during a simulation run. The code to realize raster access to a local raster data can be compiled from the information about the type and the location of the dataset (see the corresponding metamodels in Figs 3 and 4).

4.4 Transformation to the framework level

In the following, the main aspects of how HCA models can be expressed with the means of jDisco and Geotools are presented. Due to the common modeling elements, variables, discrete transition functions, ODE, conditions, and spatial data access can simply be mapped from HCA to the submodels in the framework's formalism (see Section 4.1). The conditions encoded in HCA's *phaseSelectionFunction* are simply mapped onto the respective *phaseCondition* within the continuous transition functions of the jDisco model. HCA-state-events are mapped onto conditions within the condition of a *waitUntil* action in jDisco.

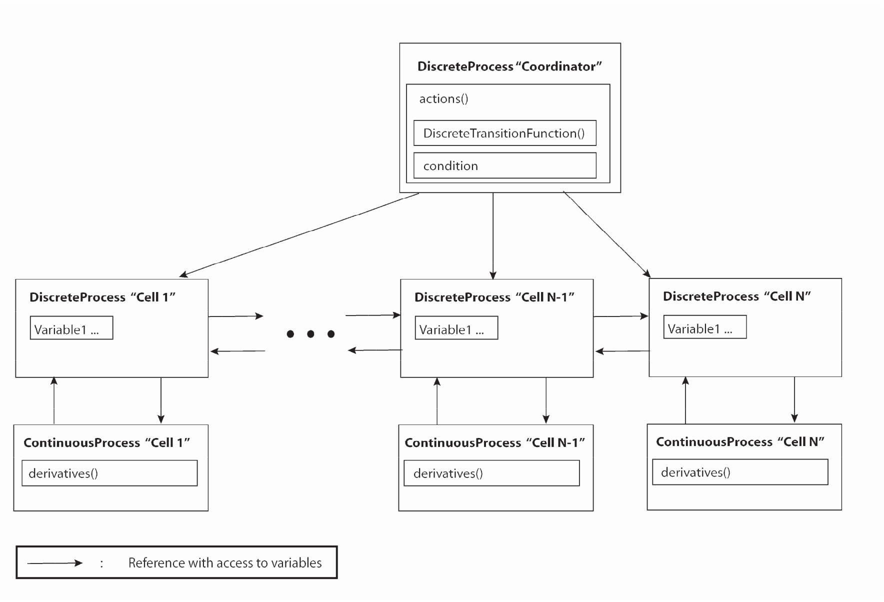
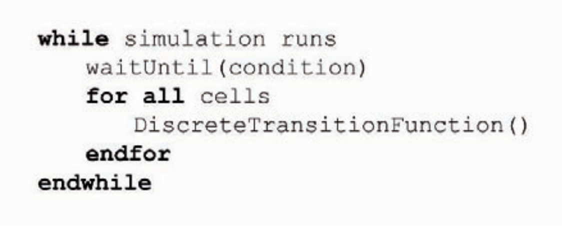

DiscreteProcess "Coordinator"
actions()
DiscreteTransitionFunction()
condition
DiscreteProcess "Cell 1"
Variable1 ...
DiscreteProcess "Cell N-1"
Variable1 ...
DiscreteProcess "Cell N"
Variable1 ...
ContinuousProcess "Cell 1"
derivatives()
ContinuousProcess "Cell N-1"
derivatives()
ContinuousProcess "Cell N"
derivatives()
: Reference with access to variables

Fig. 9. Scheme of a jDisco model of a HCA

In a jDisco-model of HCA, one *DiscreteProcess* and one *ContinuousProcess* is created for each cell of the HCA. In addition, one *DiscreteProcess* is created that acts as a coordinator for all other cell processes. The "coordinator process" holds references to all discrete cell processes (Fig. 9). Each discrete cell process holds the continuous and discrete state variables of the respective cell and references to all neighboring discrete cell processes and their variables. Moreover, the transformation adds a variable to each discrete cell process that holds the position of the cell within the model. The neighborhood is derived from the neighborhood definition of the HCA. Each continuous cell process holds the description of the cell's continuous behavior as ODE.

The discrete coordinator process "controls" the HCA. For this, a condition function holds the condition for the detection state events in all cell processes. Also the execution of the discrete transitions is realized by the coordinator process. This is possible because it holds references to all cells and their variables. The lifecycle function of the coordinator process is simple, as illustrated in Figure 10.

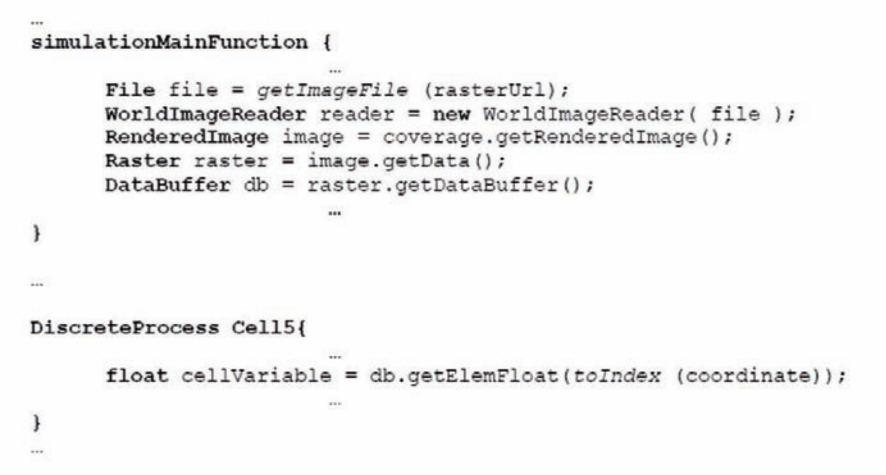


```
while simulation runs
    waitUntil(condition)
    for all cells
        DiscreteTransitionFunction()
    endfor
endwhile
```

Fig. 10. Pseudo-code of the action-function of the coordinator process

In jDisco the coupling of cells is realized through direct access to state variables. Each cell process holds references to the neighboring cell so that state values can be read directly. Since access to neighbors happens only at the execution of *discreteTransitionFunction* after synchronization (*waitUntil*), it is ensured that values of state variables have correct values at the time of access.

The transformation of the spatial data access description to the framework level and code is straightforward. A cell of a raster dataset represents the value of a cell process' variable. The reference to the corresponding raster dataset's cell is realized by mapping a cells' position to corresponding pixel's position, which is the same in the simplest case. Loading a raster file and initializing the value of variables is realized by the instantiation appropriate Geotools-classes, which are chosen by the code generator. The pseudo code in Fig. 11 illustrates this: appropriate classes are initialized at the beginning of the simulation holding a reference to the dataset (*file*) and classes that are provided by Geotools to process the raster dataset's data. Finally, with *DataBuffer*, the data can be directly accessed. The example shows how a variable representing a variable of an instance of a *DiscreteProcess* (*Cell5*) is initialized with a value from the dataset, by calling *db.getElemFloat(…).*



```
...
simulationMainFunction {
                        ...
      File file = getImageFile (rasterUrl);
      WorldImageReader reader = new WorldImageReader( file );
      RenderedImage image = coverage.getRenderedImage();
      Raster raster = image.getData();
      DataBuffer db = raster.getDataBuffer();
                        ...
}

...

DiscreteProcess Cell5{
                        ...
      float cellVariable = db.getElemFloat(toIndex (coordinate));
                        ...
}
...
```

Fig. 11. Pseudo-code showing how access to a raster dataset can be read using Geotools. By using global variables, cells' variables can be initialized by direct access

The right pixel value from the dataset is chosen by using an index that is calculated from the position (*coordinate*) of the cell inside the model by a function that is generated for that (*toIndex()*). The position of the cell is assigned during the model generation process. In a similar way it is possible to integrate the creation and modification of raster datasets.

5 Concluding Remarks and Outlook

Today's simulation and spatial data processing technologies facilitate an MDE-based approach to spatiotemporal simulation modeling, as presented in this paper. A simulation model is initially defined on an abstract level by the modeler and consequently refined by automatic model transformation to finally obtain an executable model. This approach promises to overcome shortcomings of the architecture of today's tools that are related to the reusability of simulation models.

The implementation of models is not bound to a specific simulation framework, but it is required, that generic functionality is implemented by the executing framework. In particular, the framework must provide discrete-event simulation and algorithms to solve ODE. States and discrete state transitions are defined using the means of common general purpose programming languages. The description of spatial data processing is based on concepts, as defined by widely accepted standards in the field.

In the presented example, the simulation and the spatial data processing framework use the same programming language, thus data exchange is straightforward. To integrate more different technologies (i.e. a C++-based simulation framework and a Java-based GIS library), it is necessary to generate “bridges” between technologies, for example wrappers.

The formal description of modeling levels by means of metamodels, not only provides the technical means for implementation, it also serves as the means to understand and communicate models from the viewpoint of the different users and developers of simulation software. Thus, the use of metamodeling supports transparency in model implementation. However, this requires a common understanding of the twofold meaning of metamodels by the users and developers on the different levels of abstraction.

It is an inherent characteristic of the presented approach, that through model transformation generic implementation patterns and recipes are applied to obtain executable software. On the one hand, this gives the possibility to apply best practices. On the other hand it is an obstacle for the inclusion of specific optimizations that may be possible for models with particular characteristics. However, prototypical implementations show that generic optimization is a key challenge, as model execution is demanding and efficient simulation is crucial to make this approach feasible for big cellular automata.

In further work the presented approach is to be elaborated, particularly focusing on the expressivity of HCA, optimization, and the role of spatial data processing.

Acknowledgements

The presented work is supported by Deutsche Forschungsgemeinschaft, Graduiertenkolleg METRIK (GRK 1324/1).

References

Argent, R. (2004). “An overview of model integration for environmental applications--components, frameworks and semantics”, Environmental Modelling & Software, Volume 19 (3), Pages 219-234.

Custer, A. (2006). “Geotools User’s Manual, v.0.1”, URL: http://www.geotools.fr/manual/geotools.xhtml

Evert, F. van, Holzworth, D., Muetzelfeldt, R., Rizzoli, A., and Villa, F. (2005). "Convergence in integrated modelling frameworks", in Zerger, A. and Argent, R.M. (eds) MODSIM 2005 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand, 745-750.

Fall, A. and Fall, J. (2001). "A domain-specific language for models of landscape dynamics", Ecological Modelling, 141(1-3):1–18.

Helsgaun, K. (2001). "jDisco - a java package for combined discrete event and continous simulation". Technical report, Department of Computer Science, Roskilde University.

Karssenberg, D. (2002). "Building dynamic spatial environmental models", PhD thesis, Utrecht University.

Muzy, A. Innocenti, E. Aiello, A. Santucci, J-F. Santoni P-A. and David R. C. Hill. (2005). "Modelling and simulation of ecological propagation processes: application to fire spread", Environmental Modelling & Software, 20 (7), 827-842.

Schmidt D. (2006). "Model-driven Engineering", Computer, 2/2006, 25-31.

SELES. (1999). "SELES v.1.0 Spatially Explicit Landscpe Event Simulator".

Vangheluwe, H. de Lara, J and Mosterman, P. (2002). "An introduction to multi-paradigm modeling and simulation", in Barros, F. and Giambiasi, N. (editors), AIS'2002 Conference, 9–20.

Wainer G. and Giambiasi N. (2005). "Cell-DEVS/GDEVS for Complex Continuous Systems", Simulation, 81, 137-151.

Yacoubi S. El, Jai A. El, Jacewicz P. and Pausas J. G. (2003). "LUCAS: an original tool for landscape modeling", Environmetal Modeling & Software, 18, 429-437.

Zeigler, B. Praehofer, H. and Kim T. (2000). Theory of Modeling and Simulation, Academic Press, San Diego, 2nd edition, 2000.

Support Vector Machines for Spatiotemporal Analysis in Geosensor Networks

Jon Devine[1], Anthony Stefanidis[2]

[1] Dept. of Spatial Information Science and Engineering
University of Maine,
email: jon.devine@umit.maine.edu

[2] Dept. of Earth Systems and Geoinformation Sciences
George Mason University,
email: astefani@gmu.edu

Abstract

Geosensor networks are a growing source of spatiotemporal data. However, the raw data generated by these networks, as simple collections of readings from point locations, allow little analysis to be conducted directly. As such, this research presents support vector machine based methods for the extraction of estimates for the spatial extent of areal events from geosensor data and demonstrates how these results can serve as a basis for spatiotemporal analysis. Support vector machines are a recently developed class of machine learning algorithms that have seen considerable application due to their attractive generalization properties and ability to efficiently handle large datasets. While traditionally applied to classification problems, this research demonstrates how these methods can be applied to geosensor applications where decision boundaries can be interpreted as representations of the boundaries of the spatial extent of events. Once derived, these estimates are shown as capable of serving as input for existing methods for spatiotemporal analysis and enabling description of the evolution of spatiotemporal phenomena in terms of movement and deformation. As coverage of geosensor networks increases, with sensors becoming smaller and cheaper, applications of the techniques described in this research are foreseen in environmental science, public health, and security informatics.

Keywords: Geosensor networks, spatiotemporal analysis, spatiotemporal helix, support vector machines

1 Introduction

Recent advances in sensor technology, mobile computing, and processing have had a substantial impact on the ability to collect data in a wide range of applications. In geoinformatics, these advances have led to the emergence of geosensor networks, wherein the geographic aspect of sensed information is a key component in subsequent analysis (Stefanidis and Nittel 2004). The vast amount of data already being generated by such networks supports diverse domains such as environmental science, public health, and security informatics. As wireless sensor networks become more sophisticated, with smaller and cheaper sensors and improved deployment techniques (e.g., Chen et al. 2008), the volume and complexity of geographic data produced introduces substantial challenges in terms of both data management and analysis.

From a geoinformatics viewpoint, the new form of data generated by geosensor networks implies computational challenges that extend beyond traditional spatiotemporal modeling issues associated with moving objects (e.g., Erwig et al. 1999). These challenges result not only from the extreme resource constraints on energy consumption, data storage, and processing imposed by wireless sensor networks (Karl and Willig 2005; Ganesan et al. 2003), but also from the complexity of spatiotemporal analysis required for the extraction and modeling of events captured by geosensor data.

Both the engineering and geographic literatures have developed responses to these challenges, and a common approach taken from each has been the translation of real-valued quantitative readings to qualitative values. With communication-related power consumption being very expensive in an environment with strict energy constraints, discretization has the benefit of reducing the power costs by condensing the size of data communicated by sensors to only one or two Boolean/qualitative bits (Duckham et al. 2005; Worboys and Duckham 2006; Chintalapudi and Govindan 2003; Nowak and Mitra, 2003; Nowak et al. 2004). In addition, this approach has conceptual benefits, where the separation of a continuously varying field into homogenous regions helps to make the complexity accessible to query and analysis. An important aspect related to this property is the implied existence of salient boundaries (Duckham et al. 2005).

The estimation of such boundaries has become a subject of research and several methods have been developed as means for their derivation. In general, these methods either produce rough, linear, representations (Worboys and Duckham 2006; Nowak and Mitra 2003) or are based on strong assumptions regarding uniform distribution of sensors in space (Nowak et al. 2004). This research introduces a set of techniques based on support vector machines (SVMs) that can produce smooth, highly non-linear, estimates for boundaries in irregularly distributed geosensor networks. Recently developed (Boser et al., 1992), SVMs have already been widely applied in a range of fields and have been shown to have strong generalization properties both theoretically (Christianini and Shawe-Taylor 2000; Devroye et al. 1996; Vapnik 1995) and in practice (e.g., Scholkopf et al. 1997; Yang and Liu 1999; Pontil and Verri 1998). These properties are the motivation for the consideration of SVMs for the estimation of event boundaries from geosensor data where it is thought that their generalization properties will translate to geographical accuracy.

Once derived, it is shown how SVM-derived estimates for the spatial extent of events can be incorporated into an existing spatiotemporal modeling framework. Specifically, this research demonstrates how SVM boundary estimates can be accommodated by the spatiotemporal helix (ST helix). The ST helix is a data structure designed to efficiently store spatiotemporal data and support spatiotemporal query and analysis. By describing the evolution of events in space and time in terms of both movement, defined as change in geographic location, and deformation, which captures any changes in shape (i.e., expansion or contraction in a given direction), the ST helix can represent many of the complexities inherent in the description of event evolution. In addition, the ST helix permits a visual representation of events, which allows for quick, intuitive, interpretation of the evolution of spatiotemporal events (Stefanidis et al. 2003).

To motivate, describe, and then illustrate the methods that are this research, the outline of this paper is as follows: section 2 details the type of geosensor data for which boundaries estimates are derived and reviews current methods for boundary estimation, section 3 introduces the SVM algorithm, section 4 describes the ST helix, section 5 follows an illustrative example, and section 6 is a discussion of conclusions and future work.

2 Geosensor Data and Existing Event Extraction Methods

The techniques outlined in this research are designed for data scenarios where “snapshot” readings regarding environmental conditions are collected at geosensor point locations within a dynamic spatial scalar field (see Duckham et al. 2005 for formal definition) where real-valued observations vary with time over a spatially continuous region. An inherent problem with this type of data is the derivation of accurate areal estimates for the spatial extent of events based upon these point readings. While the architecture of sensors networks has been heavily investigated in the engineering literature (see Karl and Willig, 2005 for an overview) and theoretical constructs and relevant definitions have been established in the geographic literature (Worboys and Duckham 2006; Duckham et al. 2005), few methods have been developed for the actual extraction of areal estimates and how these estimate could be incorporated into spatiotemporal analysis.

The extraction of spatial extents of events has been approached from both an engineering (Chintalapudi and Govindan 2003; Nowak and Mitra 2003; Nowak et al. 2004) and geographic perspectives (e.g., Worboys and Duckham 2006; Duckham et al., 2005). In both cases the theory and methods described involve the discretization of the real-valued observations collected at sensor points into qualitative values, thereby delineating study areas in terms salient discontinuities that can be interpreted as representations of the boundaries delimiting the spatial extent of events.

To illustrate, consider the example of a geosensor network designed to monitor the air for dangerous concentrations of a given chemical. The real-valued concentration levels observed at each sensor can be classified as either hazardous or benign according to an application specific threshold. As such, instead of reporting concentration levels, each geosensor needs only to transmit whether or not the reading at their location is in-event (toxic) or non-event (benign). With communication and storage strict limiting factors in wireless sensor networks, this reduction of real-valued observations to Boolean or quantitative values that can be represented by one or two bits is an important increase in communication efficiency and, therefore, power conservation (Worboys and Duckham 2006; Duckham et al., 2005).

Existing methods for the estimation of spatial extents are based on the determination of the boundary separating homogenous regions discretized values at geosensor point locations include those that result in linear estimates resulting from inverted quadtree (Nowak and Mitra 2003) or triangulation methods (Worboys and Duckham 2006; Duckham et al. 2005). Given that environmental phenomena (e.g., chemical plumes) likely do not have linear boundaries, it is assumed that methods that generate smooth, non-linear, estimates for spatial extents can offer a higher degree of accuracy and could suggest the use of the platelet-based method proposed by Nowak et al. (2004). However, while providing smoother estimates this approach relies on the rather strong assumption of a uniform distribution of sensors which is highly unlikely in most real-world deployment scenarios. The SVM-based methods for the extraction described in the following section are capable of producing complex, highly non-linear, estimates of the spatial extent of events from irregular distributions of sensors. Following their introduction, the resulting areal estimates are shown to be able to be incorporated into a framework for spatiotemporal analysis.

3 Support Vector Machines

Support vector machines (SVMs) are the most well-known and widely applied of the increasingly popular family of kernel methods.[1] The majority of kernel-based applications pertaining to geography have fallen within the realm of remote sensing, where SVMs (e.g., Melgani and Bruzzone 2004; Durbha et al. 2007) and other kernel-based methods such as kernel principal components analysis (e.g., Yang et al. 2006) have been applied to classify features within images.

Defining kernel methods is the use of kernel substitution of inner products to transform input data into a feature space. The motivation behind such transformations is that once in a higher (even infinite) dimension space, relatively simple decision boundaries (e.g., linear hyperplanes) or statistical procedures (e.g., principal components analysis) can be applied. Use of this transformation has become known as the “kernel trick” and it enables non-linear learning in data where direct derivation of decision boundaries in the original input space could be so complex as to be intractable (Christianini and Shawe-Taylor 2000).[2] The introduction of SVMs in this chapter is presented in three parts; the first focuses on the derivation and properties of the decision function, the second on feature spaces and kernel transformation, and the third on the mapping of the decision back to input space where, in the geosensor context of this research, it is interpreted as the spatial extent of an event.

[1] Visit http://www.clopinet.com/isabelle/Projects/SVM/applist.html for an application list.

[2] The kernel trick is not unlike some other “tricks” commonly applied in statistics. For instance, consider the case of regression where the relationship between the dependent and an explanatory variable is non-linear. Instead of relying on complex non-linear regression techniques, the explanatory variable can more simply be transformed (e.g., squared, cubed, natural log) and simple linear regression can be applied.

3.1 Margins and the Maximum Separating Hyperplane

Distinguishing SVMs from other kernel methods is the use of linear maximum margin separating hyperplanes as the decision function in feature space. Linear discriminants have long been a subject of study, with linear discriminant theory dating back to Fisher (1952). An important application of linear discriminants which is useful in terms of the development of SVMs is that of Rosenblatt's perceptron (Rosenblatt, 1956). The perceptron algorithm involves an objective based on the *functional margin* γ where

$$\gamma_i = y_i\left(\langle w \cdot x_i \rangle + b\right) \tag{3.1}$$

with y_i the label assigned to training set point x_i, the weight vector w orthogonal to the linear separating hyperplane, and b the offset. The functional margin will equal one for every correctly classified point. Therefore by maximizing the sum of these values over a training set can produce a separating hyperplane. However, while the perceptron is guaranteed to converge for linearly separable data (Novikoff 1962), maximization involving Eq. 3.1 does not yield a unique hyperplane solution and is sensitive to the choice of starting point (Christianini and Shawe-Taylor 2000). By scaling γ so that

$$\min_{\mathrm{i=1,..n}} \left|\langle w \cdot x_i \rangle + b\right| = 1 \tag{3.2}$$

and normalizing by w so that

$$\gamma = y\left(\langle w \cdot x \rangle + b\right)/\|w\| \tag{3.3}$$

a new quantity known as the *geometric margin* is obtained which is the geometric distance from any correctly classified point the to the hyperplane (see Fig. 1).

With the objective of producing a decision function with the "best" generalization ability, an intuitive choice of decision function would be one that maximizes the distance between the two classes. Indeed, by "splitting the difference" (i.e., maximizing the geometric margin $1/\|w\|$ by minimizing $\|w\|$) these generalization properties have been demonstrated both theoretically in terms of VC generalization bounds (Vapnik 1995; Christianini and Shawe-Taylor 2000; Devroye et al. 1996) and in practice where SVMs have been shown to offer equal or superior generalization results when compared against competing methods (e.g., Pontil and Verri 1998; Scholkopf et al., 1997). Another advantage of optimizing over the geometric margin is that the decision solution is unique (as opposed to functional margin) and characterized by the absence of local minima, which can be a problem with other learning techniques like artificial neural nets.

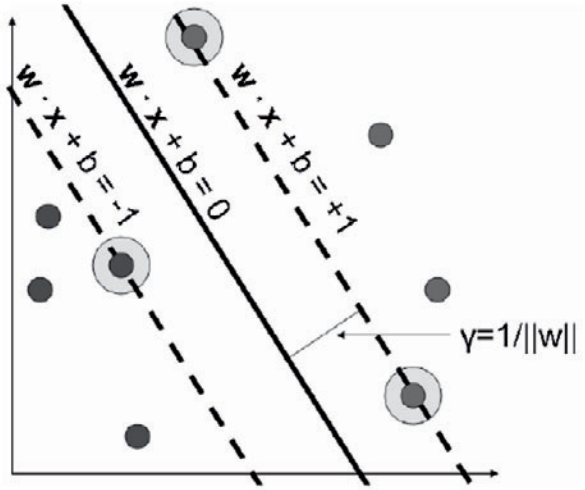


Fig. 1. Maximum (geometric) margin hyperplane. The points highlighted are the support vectors and are the points closest to the hyperplane. Supporting hyperplanes (dashed lines where $w \cdot x + b = \pm 1$) pass through these points and parallel the maximum margin hyperplane.

Having established the basis for the use of the maximum margin as the objective for the determination of an appropriate decision function, the following Lagrangian can be constructed

$$L(w,b,\alpha) = \frac{1}{2}\langle w \cdot w\rangle - \sum_{i=1}^{\lambda} \alpha_i \left[y_i (\langle w \cdot x_i \rangle + b) - 1 \right] \qquad (3.4)$$

where *w*, *b*, *y*, and *x* have the same interpretation as in Eq. 3.1 and the $\alpha_i \geq 0$ are the Lagrangian multipliers for the constraint imposing that all points be correctly classified. Exploiting the existence of a unique minimum and the associated property of stationarity, the first-order derivatives of the parameters imply

$$\frac{\partial L(w,b,\alpha)}{\partial w} = w - \sum_{i=1}^{\lambda} \alpha_i y_i x_i = 0 \qquad (3.5)$$

$$\frac{\partial L(w,b,\alpha)}{\partial b} = -\sum_{i=1}^{\lambda} y_i \alpha_i = 0\,. \tag{3.6}$$

These equations can be used to derive the Wolfe dual of the original Lagrangian expressed solely in terms of the data points, labels, and Lagrange multipliers so that

$$\begin{aligned} &\min_{\alpha} \ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j x_i x_j - \sum_{i=1}^{n} \alpha_i \\ &\text{s.t.} \ \sum_{i=1}^{n} y_i \alpha_i = 0 \qquad\qquad . \\ &\qquad \alpha_{\mathrm{i}} \geq 0 \quad i = 1,\ldots,n \end{aligned} \tag{3.7}$$

which is equivalent to the following formulation written in terms of a dot product of the input data as

$$\begin{aligned} &\min_{\alpha} \ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j \left\langle x_i \cdot x_j \right\rangle - \sum_{i=1}^{n} \alpha_i \\ &\text{s.t.} \ \sum_{i=1}^{n} y_i \alpha_i = 0 \qquad\qquad . \\ &\qquad \alpha_{\mathrm{i}} \geq 0 \quad i = 1,\ldots,n \end{aligned} \tag{3.8}$$

When solved for in conjunction with the Kuhn-Tucker (KT) complementarity conditions

$$\alpha_i^* \left[y_i \left(\left\langle w^* \cdot x_i \right\rangle + b^* \right) - 1 \right] = 0 \qquad i = 1\ldots,n \tag{3.9}$$

these formulations yield the solution for the maximum margin hyperplane. The KT conditions imply that for only those points closest points to the hyperplane are the α_i^* non-zero. These are the support vectors, and by looking at the expressions for both the objective and the constraints in Eq. 3.7-3.8, one can see that the solution is dependent only on these points. The fact that the solution is dependent only on a subset of the input data is a feature of SVMs that contributes to their ability to scale to large datasets (Scholkopf and Smola 2002).

3.2 Non-Linearity and Kernel Substitution

Up to this point all formulations have described the derivation of a maximum margin hyperplane for data that are linearly separable in input space. Key to non-linear learning with SVMs is the mapping of the input data *x* into a higher dimensional nonlinear feature space via a transformation $\Phi(x)$. As such, the optimization problem appearing in Eq. 3.8 would instead appear as

$$\begin{aligned} &\min_{\alpha} \ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j \Phi(x_i)\cdot\Phi(x_j) - \sum_{i=1}^{n}\alpha_i \\ &\text{s.t.} \ \sum_{i=1}^{n} y_i \alpha_i = 0 \\ &\qquad \alpha_{\text{i}} \geq 0 \quad i = 1,\ldots,n \end{aligned} \tag{3.10}$$

Explicit calculation of these transformations could be computationally expensive, and even impossible in case of infinite dimensional transformations such as the Gaussian (Burges 1998). However, via the “kernel trick” (Aizermann et al. 1964) it is possible to compute the inner product $\Phi(x_i)\cdot\Phi(x_j)$ implicitly and directly from the input data with a kernel function *K* where

$$K(x_i, x_j) = \Phi(x_i)\cdot\Phi(x_j). \tag{3.11}$$

Characterizing which functions can be considered kernels is Mercer’s Theorem which states that symmetric functions *K* are kernel functions if and only if the Gram matrix $\mathrm{K_G}$, written

$$\mathrm{K_G} = \left(K(x_i, x_j)\right)_{i=j=1}^{n} \tag{3.12}$$

is positive semi-definite which ensures convexity in the optimization problem and therefore implies a unique hyperplane solution (Christianini and Shawe-Taylor 2000). See Table 1 for a list of commonly used kernel func**tions.**

Table 1. Commonly Used Kernels

Kernel	**Functional Form**
Polynomial	$(x_i \cdot x_j + 1)^d$

Table 1. (Continued)

Kernel	Functional Form
Gaussian	$\exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma}\right)$
Neural Network[a]	$\tanh\left(\kappa\left(x_i, x_j\right) + \Theta\right)$

In Table 1 The parameters $\kappa > 0$ and $\Theta \in \Re$ are the gain and horizontal shift. While this kernel formulation has been shown effective, it also presents some challenges from a mathematical perspective and requires careful application (Abe, 2005).

Substitution of a kernel representation for a feature space inner product results in yet another formulation for the optimization problem so that

$$\begin{aligned} &\min_{\alpha} \quad \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{i=1}^{n}\alpha_i \\ &\text{s.t.} \quad \sum_{i=1}^{n} y_i \alpha_i = 0 \qquad . \\ &\qquad \alpha_{\mathrm{i}} \geq 0 \quad i = 1,\ldots,n \end{aligned} \tag{3.13}$$

This is the optimization problem for the SVM algorithm. The effect of the kernel substitution in Eq. 3.13 introduces non-linearity into the classification problem and, by implicitly defining the feature space directly from the input data, does so in a manner that does not require significantly more computation than the derivation of the maximum margin classifier in input space (Eqs. 3.7-3.8). In addition, the attractive generalization properties described in the previous section for the maximum margin hyperplane still hold (Scholkopf and Smola 2002).

3.3 Mapping the Hyperplane Solution to Input Space

Once derived in feature space, the maximum margin separating hyperplane can be mapped back from feature space to input space. This process involves deriving the optimal parameter values for *w** and *b** (Eqs. 3.2-3.4) from the optimized Lagrange multipliers *α** obtained from Eq. 3.13. The expression for *w** can be simply derived from the constraint from Eq. 3.5 which shows that

$$w^* = \sum_{i=1}^{\lambda} \alpha_i^* y_i x_i \ . \tag{3.14}$$

Meanwhile, *b** may be obtained through the use of the KT conditions (Eq. 3.9) along with the optimized value for w as

$$b^* = y_i - \sum_{i=1}^{\ell} y_i \alpha_i^* K(x_i, x_j) \tag{3.15}$$

over all points with $\alpha_i > 0$ (from the λ support vectors). With these optimized values, the decision function

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{m} y_i \alpha_i^* K(z, x_i) + b^* \right). \tag{3.16}$$

is obtained and is used to suggest labels for test points *z* (Scholkopf and Smola 2002). With the initial input data representing geosensor locations and Boolean event predicate results and by inputting a mesh of test points *z* over the area considered as being monitored by the geosensor network, the results of the decision function can be used as a means of interpolating the spatial extent of events from these point readings.[3] The effect of the test function (Eq. 3.16) is the generation of a raster which gives positive results for locations hypothesized to be in-event and negative results for locations estimated to be non-event.

In addition, and more importantly for the purposes of this research, the event boundary can be estimated by finding those points for which the decision function is equal to zero. Such an interpretation is the basis for this research where it is thought that the attractive generalization properties and robustness of SVM-derived decision boundaries suggest accurate representation of the spatial extent of events. A focus of on-going research is the testing of these methods against ground-truths in order to verify the accuracy of these estimates. In the following section, an overview of the ST helix is presented to provide a background for the demonstration of how SVM-generated estimates can be incorporated into existing spatiotemporal data structures for query and analysis.

4. The Spatiotemporal Helix

The spatiotemporal helix is a framework for summarizing the evolution of spatiotemporal phenomena. Designed to allow efficient querying of data and to support intuitive visual representations of event evolution, other strengths of the ST helix include its ability to facilitate complex description of event evolution in terms of both the event’s trajectory and its deformation.

[3] The spatial extent of the meshgrid should be dictated by application specific assumptions concerning the effective coverage of a geosensor network.

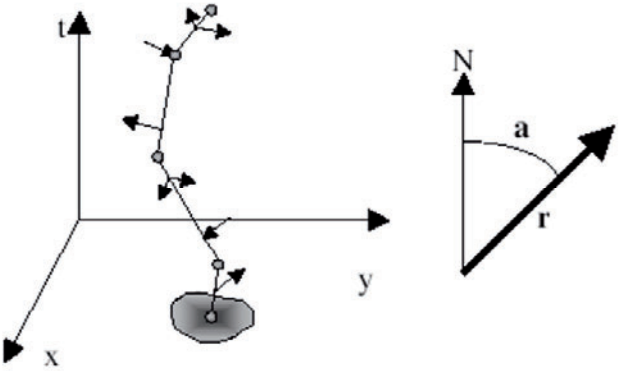


Fig. 2. A spatiotemporal helix. The gray spheres depict nodes and define the helix's spine. Prongs are represented by arrows. Outward facing arrows denote expansion and inward facing arrows indicate contraction. Arrow length reflects the magnitude of deformation while their angle reflects the azimuth range over which the deformation occurred.

Structurally, the spatiotemporal helix is composed of two parts (see Fig. 2). The first is a central spine, which depicts an entity's trajectory and is defined by a series of nodes, $s_i = \{x, y, t, q\}$ where *x, y,* and *t* are the node's spatiotemporal coordinates and *q* is a qualifier classifying the node's dominant type of movement (acceleration or deceleration). Prongs, $p_i = \{t, r, a_1, a_2\}$, protrude from the spine and describe an object's deformation in terms of the time coordinate *t*, the magnitude of outline change *r* (with positive values for *r* indicating expansion and negative values indicating contraction), and a_1 and a_2 denoting the azimuth range where the deformation took place. To minimize redundant information and reduce data storage requirements, only significant changes in the event's velocity and outline are recorded as nodes and prongs (Stefanidis et al. 2003; Agouris and Stefanidis 2003).

In order to determine which changes in velocity and shape are significant, and consequently which nodes and prongs constitute a ST helix, self-organizing map (SOM) and differential snakes techniques are applied. The derivation of appropriate nodes is complex in that significant change in trajectory implies consideration not only of the distance traveled over time, but also of the direction. For this reason, a sophisticated approach involving a geometric adaptation of self-organizing maps is used which assigns more nodes to time periods of intense change and few nodes to periods of stability (Partsinevelos et al. 2001).

To develop the prongs an adaptation of deformable contour models, differential snakes, is used. This method considers changes in an events shape as a function of differences in the distance from an event's center of mass to points on its boundary from time t to time $t + dt$. The percentage of change in these distances is successively compared against a user defined threshold to identify where significant changes have occurred (Agouris et al. 2001). These significant changes and their sign, negative change implying contraction and positive change indicating expansion, are recorded as prongs (Agouris and Stefanidis 2003).

With selection of appropriate parameter values for determining significance, a concise signature of the evolution of event occurring over the time period frame *t1* to *t2* can be captured by a ST helix and written $Helix_{t1,t2}^{objid} = \{node_1,...node_n; prong_1,...prong_m\}$ (Stefanidis et al. 2003). This signature is the basis for the development of similarity metrics outlined by Croitoru, Agouris et al. (2005) which demonstrated the ability of the node and prong data stored in ST helix as capable of differentiating the evolution of 25 different hurricanes and facilitated discussion regarding the similarity of their evolution. These results suggest that the ST helix could also be used for spatiotemporal analysis and allow for similar comparison of the evolution of events captured by geosensor data.

5. Simulation

Foreseen applications for the methods outlined in this paper include those for which the locations of sensors are known (i.e., geosensor positions are recorded as they are installed or they are tracked as they move), and in which individual sensor readings are communicated to "base station" nodes (Nittel et al. 2003). As is standard policy in geosensor networks, for energy conservation reasons, once a baseline set of readings has been established only those nodes detecting change relay their information along to these base stations. These larger, less energy constrained nodes, could then relay the collected information to a central location for processing or conduct the processing themselves.

With trends in geosensor research pushing towards decentralization and with improving processing capabilities of sensors, the SVM methods described in this work could also be applied in-network where the network is able to adapt its sensing efforts to the evolution of events. Such network designs are known as *active* and have the advantage, from a power conservation standpoint, of being able to direct resources to areas of interest. For example, SVM boundary estimates could be used to "wake up" sensors in power save mode that are in proximity to the boundary estimate to increase accuracy.

Regardless of the behavioral structure of the network, this research is designed for situations where the distribution of sensors is sparse relative to the desired level of spatial detail for estimates of the spatial extent of events. The purpose of this chapter is primarily demonstrative, to outline how SVMs can be used to translate geosensor readings (i.e., point data) into spatiotemporal analysis using the ST helix framework. Discussion highlights several objectives for future research, notably the expansion of ST helix techniques for the description of multiple polygon events and the testing against ground truths to validate SVM estimates and tune parameter values. The basic steps described in this chapter appear in the schematic in Fig. 3.

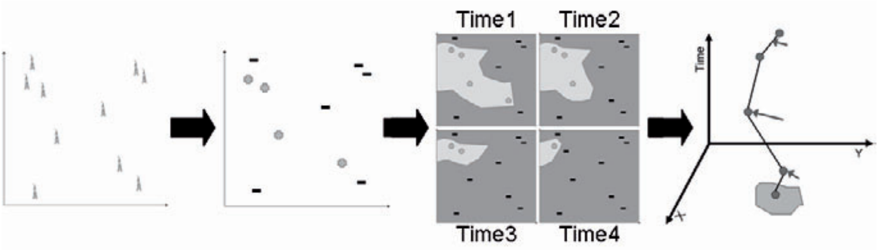


Fig. 3. Schematic of the SVM-based procedure for the modeling of spatiotemporal events. The first frame represents a geosensor network. In the second frame, readings at a given time instance for each of the sensors are labeled as being in-event (+'s) or non-event (-'s). The third frame depicts SVMs results. These areal estimates can then be incorporated into the ST helix framework for summarization and analysis.

For each temporal instance of the simulation, the spatial location of each sensor along with a label corresponding to in-event/non-event status is inputted into a SVM.[4] When implementing SVMs a number of parameters are involved. The first of these involves the selection of a functional form for the kernel for the substitution of the inner product (see Equation 3.13). A number of different kernel types have emerged in applications, the most popular of these appear in Table 1 (Chapter 3).

Due to the use of Gaussian kernels in other geographic applications, (e.g., kernel density estimation), this kernel was selected for this demonstration. As with the other kernels, selection of the Gaussian functional

[4] Other dimensions can be added for more complex data situations. For instance, if the sensors were allocated for oceanographic applications a dimension for depth dimension could also be included.

form implies the assignment of parameter values. Examples of such parameters are the degree, *d*, for the polynomial kernel and the bandwidth parameter, *σ*, for the Gaussian. The determination of values for these parameters has an effect on results (see Fig. 4) and a major criticism of SVMs and other kernel-based methods, stems from the weight of the choice of parameter values. However, the flexibility by the choice of parameters can also be seen as an advantage in that application specifications could be used to inform value selection. For example, in the context of geosensors, bandwidth values could be reflective of the average distance between sensor nodes in a network or a reflection of the sensing radius appropriate to the application.

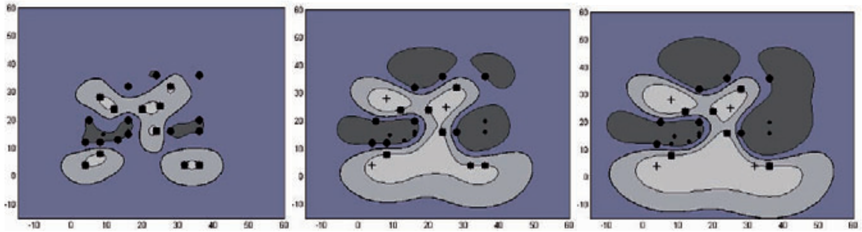


Fig. 4. Hypothetical geosensor network at the same temporal instance and SVM estimates with varying bandwidths. The three frames show SVM estimates for the spatial extent of events with bandwidths of σ = 10, 50, 70.

In any given timeframe the results from the SVM used in this demonstration delimit four regions (Fig. 4-5). Two of these regions, the lightest and the darkest, are bound by support vectors and correspond to the supporting hyperplanes (in feature space) in Fig. 1. The support vectors are differentiated from interior points and appear either as large black squares (in-event support vectors) or as large black dots (non-event support vectors). Interior points appear either as crosses (in-event case) or small dots (non-event case). The two intermediate gray regions (neither the darkest nor the lightest shades) are separated by the decision boundary (i.e., where Eq. 3.19 equals zero). To model event evolution, either the in-event side of the decision boundary (second lightest shade) or the in-event support vector bounded regions (lightest shade) could be modeled. A more complete representation could model both.

Given that SVMs can produce complex estimates for the spatial extent of events, including those composed of multiple polygons, the differential snakes approach is modified (i.e., changes are examined at both the component polygon and aggregated composite event levels) so that information describing changes in the outline of both component polygons and the composite event can be captured. The fact that SVM estimates can produce these multi-resolution estimates (i.e., component polygon and composite event) opens up analytical possibilities for spatiotemporal *occurrents* such as merging/splitting and appearance/disappearance of event components (see Duckham et al., 2005 for formal definitions).

More challenging is the adaptation of ST helix methods to describe trajectory. At the root of this issue is the establishment of a framework which can differentiate between occurrents that produce similar changes in the number of component polygons (i.e., splitting/appearing and merging/disappearing). Distinction is important because it informs how trajectories should be allocated (e.g., with a decrease in the number of polygons should the trajectory of a component polygon stop or should it merge) and, correspondingly has an impact on results from spatiotemporal analysis. While methods for discriminating among occurrents are beyond the scope of this paper, a crude framework based on the count, centers of mass, and area of component polygons. When merging or splitting is suggested, the corresponding helix representations also merge/split. When appearance is suggested a new trajectory is established, with disappearance an existing trajectory stops.

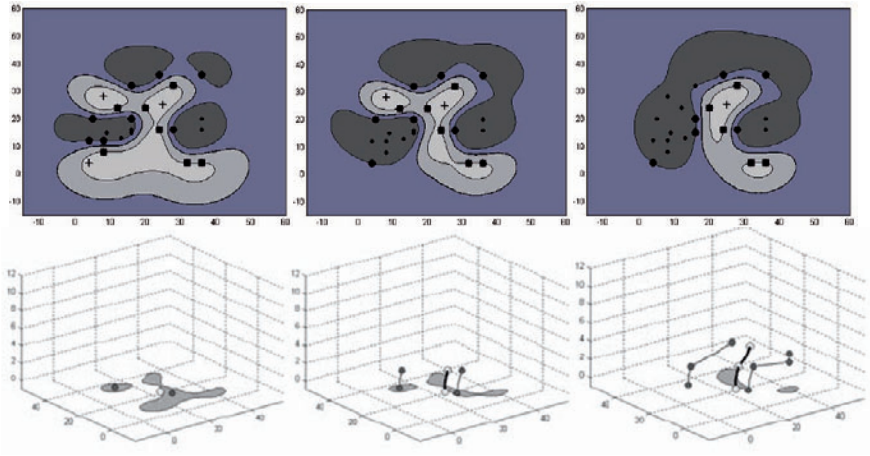


Fig. 5. Temporal sequence of frames from the simulated sensor network data set and traces of their trajectories (no prongs) for each frame (σ=50). The light gray regions on the floor of the 3-dimensional graphs on the second row reference the support vector bounded in-event regions. The dark nodes trace the trajectories of the component polygons while the light nodes trace the trajectory of the composite event.

With this informal model, SVM generated boundary estimates corresponding to different time instance were used to create the ST helix representations in Fig. 5 and 6. Fig. 5 shows trajectories incorporating all temporal instances while Fig. 6 depicts “thinned” trajectories using the SOM techniques described in the Chapter 4. Fig. 6 also depicts prongs representing deformation at both for the component polygon and the composite event level.

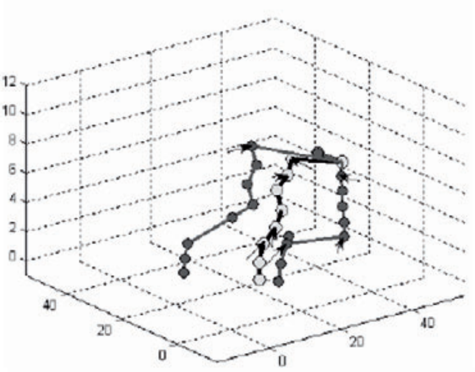


Fig. 6. Spatiotemporal helix summarizing the simulated event.

6. Conclusions and Future Work

The research presented in this paper outlined a set of methods for spatiotemporal analysis of areal events as described by point data. A novel approach implementing SVMs was introduced as a means of estimating the spatial extent of events from simulated scalar field geosensor data. These estimates were incorporated into existing spatiotemporal helix based techniques for data management and analysis. Results demonstrated that complex events can be extracted from the type of data (i.e., dynamic spatial scalar fields) generated by geosensor networks with no bias a priori regarding the shape or number of component polygons. Analysis based on the spatiotemporal helix showed that event evolution can be described in terms movement (i.e., change in geographic location), deformation, and explored the potential for description of merging/splitting and appearance/disappearance of component polygons. It is thought that the techniques demonstrated in this work offer significant potential for advancement in a range of fields, especially as the deployment of geosensors becomes more widespread as technology makes sensors smaller and cheaper.

While promise for this advancement exists, the results demonstrated in this research are derived from a relatively simple simulated data set.

Additional research should be conducted using geosenor test beds in the presence of real-world events with known spatial extents (i.e., a ground truth) in order to compare results. Such analyses could be particularly useful in refining the informal model for the description of occurrences outlined in the previous section. Experiments such as these would also be valuable not only in providing an empirical measure of the strength of these methods but in investigating appropriate parameter values for SVMs.

In addition, future work should exploit the ability of SVMs to handle high dimensional data and examine more complicated scenarios. One such an example could be that of certain oceanographic applications involving depth. Another, perhaps more widely applicable dimension which could be examined is that of time. This research, being of exploratory nature and with the objective of presenting an interpretation of SVM generated results for geographic applications, considered time in a discrete "snapshot" fashion. However, with their ability to handle high-dimensional data, SVMs can directly accommodate the near continuous monitoring of environment that geosensor networks allow. Future work should examine how this more precise temporal data could be accommodated and meaningfully analyzed with spatiotemporal frameworks such as the ST helix.

Acknowledgements

The authors acknowledge the helpful comments provided by three anonymous reviewers. This material is based upon work supported by the National Science Foundation under Grant No. 0429644 through the Science and Engineering Informatics and Information Integration Program (SEIII), and the National Geospatial-Intelligence Agency through a NURI grant.

References

Abe S (2005) Support Vector Machines for Pattern Classification, Springer

Agouris P, Stefanidis A (2003) Efficient Summarization of Spatiotemporal Events. Communications of the ACM 46**:** 65-66

Agouris P, Stefanidis A, Gyftakis S (2001) Differential Snakes for Change Detection in Road Segments. Photogrammetric Engineering & Remote Sensing 67**:** 1391-1399

Aizermann M, Braverman E, Rozonoer L (1964) Theoretical Foundations of the Potential Function Method in Pattern Recognition. Automation and Remote Control 25**:** 821-837

Boser BE, Guyon IM, Vapnik V (1992) A Training Algorithm for Optimal Margin Classifiers. In: Haussler, D. (Ed.) 5th Annual ACM Workshop on Computational Learning Theory. ACM Press

Burges CJC (1998) A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery 2**:** 121-167

Chen Y, Chuah C, Zhao Q (2008) Network Configuration for Optimal Utilization Efficiency of Wireless Sensor Networks. Ad Hoc Networks 6**:** 92-107

Chintalapudi KK, Govindan R (2003) Localized Edge Detection in Sensor Fields. Ad Hoc Networks 1**:** 273-291

Christianini N, Shawe-Taylor J (2000) An Introduction to Support Vector Machines. Cambridge University Press

Croitoru A, Agouris P, Stefanidis A (2005) Rotation, Translation, and Scale Invariant 3D Trajectory Matching by Pose Normalization. In: Shahabi, C. & Boucelma, O. (Eds.) ACM-GIS'05. ACM Press, Bremen

Devroye L, Gyorfi L, Lugosi G (1996) A Probabilistic Theory of Pattern Recognition, Springer

Duckham M, Nittel S, Worboys M (2005) Monitoring Dynamic Spatial Fields Using Responsive Geosensor Networks. ACM International Workshop on Geographic Information Systems. ACM Press, Bremen

Durbha SS, King RL, Youman NH (2007) Support Vector Machines Regression for Retrieval of Leaf Area Index from Multirange Imaging Spectroradiometer. Remote Sensing of Environment 107**:** 348-361

Erwig, M., Gueting, R. H., Schneider, M. & Vazirgiannis, M. (1999) Spatio-Temporal Data Types: An Approach to Modeling and Querying Moving Objects in Databases. Geoinformatica 3**:** 143-148

Fisher, R. (1952) Contributions to Mathematical Statistics. Wiley, New York

Ganesan, D., Estrain, D. & Heidermann, J. (2003) Dimensions: Why do we Need a New Data Handling Architecture for Sensor Networks? ACM SIGCOMM Computer Communication Review 33**:** 143-148

Karl, H. & Willig, A. (2005) Protocals and Architectures for Wireless Sensor Networks. Wiley, West Sussex, England

Melgani, F. & Bruzzone, L. (2004) Classification of Hyperspectral Remote Sensing Images with Support Vector Machines. IEEE Transactions on Geoscience and Remote Sensing 42**:** 1778-1790

Nittel, S., Duckham, M. & Kulik, L. (2003) Geographic Information Science. In: Egenhofer, M. & Mark, D. M. (Eds.) Second International Conference, GIScience 2003. Springer

Novikoff, A. B. (1962) On Convergence Proofs on Perceptrons. Symposium on the Mathematical Theory of Automata. Polytechnic Institute of Brooklyn

Nowak, R. & Mitra, U. (2003) Boundary Estimation in Sensor Networks: Theory and Methods. In: Guibas, L. & Zhao, F. (Eds.) Second International Workshop on Information Processing in Sensor Networks. Springer, Palo Alto

Nowak, R., Mitra, U. & Willet, R. (2004) Estimating Inhomogenous Fields Using Sensor Networks. IEEE Journal on Selected Areas in Communications 22**:** 999-1007

Partsinevelos, P., Stefanidis, A. & Agouris, P. (2001) Automated Spatiotemporal Scaling for Video Generalization. IEEE International Conference on Image Processing. Thessaloniki, Greece

Rosenblatt, F. (1956) The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. Psychological Review 65**:** 386-408

Scholhopf, B. & Smola, A. (2002) *Learning with Kernels,* MIT Press, Cambridge, MA

Stefanidis, A., Eickhorst, K., Agouris, P. & Partsinevelos, P. (2003) Modeling and Comparing Change Using Spatiotemporal Helixes. In: Hoel, E. & Rigaux, P. (Eds.) ACM-GIS'03. ACM Press, New Orleans

Stefanidis, A. & Nittel, S. (2004) GeoSensor Networks CRC Press

Vapnik, V. (1995) The Nature of Statistical Learning Theory. Wiley, New York

Worboys, M. & Duckham, M. (2006) Monitoring Qualitative Spatiotemporal Change for Geosensor Networks. International Journal of Geographic Information Science 20**:** 1087-1108

Yang, R., Tan, J. & Kafatos, M. (2006) A Pattern Selection Algorithm in Kernel PCA Applications. First International Conference on Software and Data Technologies. Setubal, Portugal

Toward a Method to Generally Describe Physical Spatial Processes

Barbara Hofer, Andrew U. Frank

Department of Geoinformation and Cartography, Vienna University of
Technology, Gusshausstrasse 27-29 E127, 1040 Vienna, Austria
e-mail: {hofer;frank}@geoinfo.tuwien.ac.at

Abstract:

Spatial processes are the focus of geography and should play a prominent role in geographic information systems (GIS). However, current GIS focus on the static description of properties in space and do not systematically support processes. A general method to describe spatial processes is a prerequisite to including processes in GIS software. This paper outlines an attempt to a general and application independent method to describe processes, limited currently to physical spatial processes. The methodology is based on first modeling a process with a deterministic model. The deterministic models employed here divide the region of interest into blocks and define the influence of the process on each block. The resulting model equations are then related to partial differential equations (PDEs), which are an alternative method for describing processes. Thereby, the qualitative characteristics of processes are identified. A method for describing processes has to be capable of covering the identified characteristics of the processes. As an example the process of diffusion of a contaminant in water is analyzed. The results of this study suggest that this approach allows identifying commonalities among spatial physical processes. These insights can lead to a set of types of processes on which a method to describe spatial processes can be based in the long run.

Keywords: spatial physical processes, deterministic models, partial differential equations

1 Introduction

Most currently available models of space in geographic information systems (GIS) focus on the representation of the earth in a static way; there is, however, an increasing need to systematically support change, dynamics, and processes in GIS, representations of data, visualization schemes, etc. (MacEachren 1995; Frank 1998; Blok 2000; Yuan 2001; Miller and Wentz 2003; Worboys 2005; Goodchild et al. 2007).

Spatial processes are processes taking place in space and may depend on location in space. They show different natures and are studied in different disciplines like ecology, geography, geocomputation, and physics. Examples are the spread of forest fires (Yuan 2001), the growth of cities (Batty et al. 1999), the migration of species (Seppelt 2005), and the flow of water (Mitasova and Mitas 2002). Terminology across disciplines varies. Different disciplines describe the process of interest in the application, but no commonality between disciplines is achieved.

Physical spatial processes are governed by physical laws like mass conservation. In addition, they are continuous processes and are dominated by local influences. They are considered spatial, if they fall into the temporal and spatial frequency interval typical for geography.

The long term goal of the work reported here is to provide the outline of a domain and application independent method to generally describe spatial processes, limited to physical spatial processes. Such a method is a prerequisite to including processes in GIS software and to extending our current concepts of space.

For describing physical spatial processes on a general level, we need to identify the qualitative characteristics, which explain the behavior of the process over time. Our methodology is based on modeling a spatial process with two different models, namely deterministic block models and partial differential equation (PDE) models. These two types of models are alternative ways to describe processes, having different advantages. Block models of processes are useful for conceptualizing processes and for simulating processes computationally. Models of processes based on PDEs are useful for identifying generic properties of a process or a family of processes. The theory of linear PDEs discerns three main types of processes that are described by different types of equations: wave-like, diffusion-like, and steady-state processes.

Deterministic models formulated as difference equations can be related to PDEs. Thereby we establish a link between the two models and have a description of a process from both points of view. This procedure allows us to gather information about qualitative characteristics of a process, which have to be included in the description of a process. This methodology is applied here, as a practical example, to the process of the diffusion of a contaminant in water.

The results of our research show that general insights on a formal and qualitative level can be gained. Applying the approach repeatedly on a list of spatial physical processes will allow the identification of commonalities among processes. This is an important step towards a set of tools for describing spatial physical processes.

This article is divided into seven sections. Following the introduction, a brief review of the literature related to spatial processes and GIS is given. Subsequent to a definition of spatial physical processes in section 3, two models for these processes are introduced: deterministic models and PDEs. A specific example of modeling a process is given in section 5 and the characteristics of the example process discussed in section 6. The section on conclusions and future work is the final section of the paper.

2 Spatial processes and geographic information systems

Numerous attempts to describe spatial processes exist. In the sequel of the quantitative revolution in geography a focus on detailed treatment of processes in geography became feasible. Abler, Adams, and Gould (1977, p.60) define spatial processes as "…mechanisms which produce the spatial structures of distributions". For them the task of geography is to answer the question: "why are spatial distributions structured the way they are?" (Abler et al. 1977, p.56). Getis and Boots (1978) and Cliff and Ord (1981) worked in this direction. They were interested in understanding the connection between a process and the resulting form of patterns on a map. They intended to connect the static, observable state of geographic space with the process that shaped the geographic reality, linking the snapshot with the dynamics.

The work on spatial processes in the field of geographic information systems and science is extensive and can be driven by very different objectives. The following listing briefly mentions various related achievements and research contributions:

- Development of software packages like Map Algebra (Tomlin 1990) and PCRaster (Van Deursen 1995) for analyzing and simulating spatial phenomena.
- Development or extension of data models for handling the dynamics or particularities like continuity of spatial phenomena (Kemp 1992; Reitsma and Albrecht 2005; Worboys 2005; Goodchild et al. 2007).
- Analysis of analytical GIS operations and investigation of the links between processes manipulating GIS data and processes in reality (Albrecht 1998).
- Investigation of the linkage of modeling tools and GIS (Kemp 1992; Van Deursen 1995; Abel et al. 1997; Hornsby and Egenhofer 1997; Bivand and Lucas 2000).
- Investigation of a single process like diffusion (Hornsby 1996) or geographic movement (Tobler 1981).
- Investigation of network geography and the representation of network related process in GIS (Batty 2005).
- Modeling of geographic phenomena with existing respectively prototype GI systems (Yuan 2001; Mitasova and Mitas 2002).

Despite all the efforts to analyzing and classifying spatial processes, these have not been widely accepted yet. Part of the confusion, making discussion of processes so difficult, is the sheer variability. The scope of the discussion is overwhelming and grouping in arbitrary many ways possible. The paper addresses this issue by aiming at a domain and application independent method of analyzing physical spatial processes. The novel contribution of this work is the use of PDEs and deterministic models in a qualitative study of spatial processes.

3. What are physical spatial processes

Generally speaking, spatial processes happen in space and may depend on location in space. Getis and Boots (1978, p.1) define spatial processes as "…tendencies for elements to come together in space (agglomeration) or to spread in space (diffusion)". These definitions indicate that nearly every spatial phenomenon is a process and discussing processes seems to be discussing everything.

In order to avoid this trap, the approach here concentrates first on physical processes. This links to the tiered ontology Frank (2001) has used in other contexts successfully: physical processes cover a very large part of geographic processes, but not all of them. If this restriction is useful it must lead to conceptual clarity and extending some of the insight beyond the limitation possible.

Ontologically we separate the physical reality as the part of reality which is governed by physical laws from the tier of our reality which is socially constructed and governed by social (legal) rules. Ontologists assume that physical processes have all their effects continuously in space and their influences restricted to the neighborhood. Therefore, physical processes are strictly local and do not depend on global knowledge. Frank (2001) pointed out that physical processes are describable by (partial) differential equations.

This argument is used here in the reverse direction: the physical processes studied here are exactly those describable by differential equations; this restriction enforces the focus on strictly local processes. This restriction seems to be acceptable in geography; Tobler's first law of geography says: "everything is related to everything else, but near things are more related than distant things" (Tobler 1970, p.236).

Processes are considered geographic if their frequency in time or space falls into the frequency interval typical for geography. Geography focuses on spatial objects of size between 0.1m and 40.000km and processes where change is noticeable in minutes to 10.000 years. Typical examples for geographic physical processes include: soil erosion, migration, groundwater flow, stream flow, sediment transport, forest fires, floods, saltwater intrusion, surface runoff, flux of pesticides.

Excluded as non physical are processes which are not controlled by physical causation, but by information causation (Frank 2007). If a computerized or human information processing unit at one place is the cause of a physical spatial process at a possibly distant other place, we speak of information causation. Information causation is not limited to neighborhood: a decision by a single person in a "center of power" can be transmitted for and have devastating effects at a very distant location. Such information caused processes are excluded from the current discussion.

4. Two models of spatial physical processes

Two fully general and equivalent models for the description of spatial physical processes are presented in this section: deterministic models based on blocks and partial differential equation (PDE) models. The block model describes the process and its characteristics with respect to blocks of finite size. These models are useful for conceptualizations of processes and communication about processes but also a useful approach to computational simulation. The theoretical analysis of physical spatial processes is helped by the focus on making the blocks smaller and smaller till they become infinitely small; this leads to continuous models of processes, which are here represented by partial differential equations. The two models are equivalent and every process can be described in either of them and the translation between the two models is always possible.

4.1. Deterministic Models Based on Blocks

An important principle of physical systems is the conservation of some quantity like mass or energy. Fundamental conservation laws state, that the amounts of a quantity going in, going out, and being created or destroyed in a region, have to correspond to the amount of change in a certain region (Logan 2004). Besides these laws of conservation that describe the storage of a quantity, the transfer of a quantity is described by transport laws or flow laws (Thomas and Huggett 1980). Thomas and Huggett (1980, p.64) define a deterministic model as “a storage equation in which the input and output rates are defined by suitable transport laws.” The following explanations of deterministic models are based on (Thomas and Huggett 1980).

A physical spatial process occurs in space and we can cut out a small piece of space, a block or an element, and describe the change in the relevant parameters describing the process. We thereby define the spatial domain as a set of blocks (Fig. 1(a)). Blocks can be combined in various ways, depending on the process; for studying, e.g., water flow in a river, blocks may be arranged in a line (Fig. 1(b)).

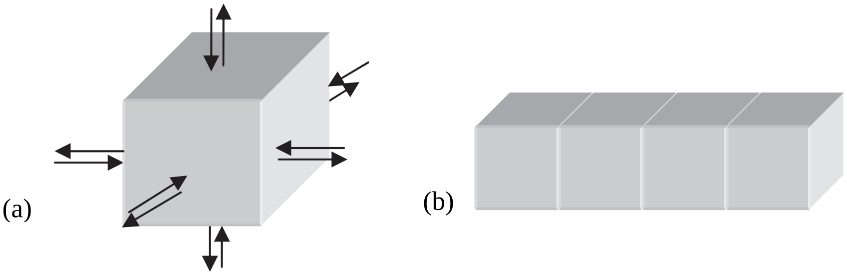


Fig. 1. (a) a block as spatial unit, (b) a sequential alignment of blocks for studying for example water flow in a river.

After selecting the process of interest and defining the spatial domain, storage equations are established for every block in the spatial domain. As said above, the storage equations state the change of a quantity Δq in the block; this change is determined by the difference of the flow in f_i and the flow out f_o of the block in a given time interval Δt (see Eq. 4.1.1).

$$\Delta q = \left(f_i - f_o\right)\Delta t \tag{4.1.1}$$

For defining the input and output terms of the quantity, we need transport laws. These laws are derived from physical characteristics of the processes. Conservation laws apply again, which means that the outflow of a block through one face must be equal to the inflow in the neighboring block. Important transport laws are Fick's first law of diffusion, Fourier's law of heat transport, and Darcy's law of water flow. Fick's law, for example, states that the negative gradient of the concentration of the quantity (ΔC) times the diffusion coefficient D of the quantity, is proportional to the quantity flow rate f. Eq. 4.1.2 states the gradient of the concentration of a quantity in x direction with $\frac{\Delta C}{\Delta x}$.

$$f = D\left(-\frac{\Delta C}{\Delta x}\right) \tag{4.1.2}$$

The transport law applies to all blocks except those at the boundary of the region of interest. Special flow conditions known as boundary conditions are defined for blocks at the region's boundary.

The equations that are formed by applying this modeling technique describe the storage change in discrete time intervals t1, t2, t3 etc. Therefore, they are difference equations. In a difference equation the change of a quantity over time can be expressed by the relation between successive values of the quantity. For running the models, initial conditions for the storages at the start, boundary conditions and parameter values have to be given. The difference equations can then be solved, the results evaluated and the model adjusted.

4.2. Differential Equations to Model Processes

A differential equation is an equation where variables and derivations from variables are brought into a relation. The general solution to a differential equation is a function or a family of functions describing some aspect of a process. Ordinary differential equations (ODEs) depend on one independent variable and contain derivatives with respect to this variable only.

Spatial processes are described by partial differential equations (PDEs), because they depend on more than one independent variable like space and time, or several spatial dimensions. PDEs allow modeling the change of a variable of interest that depends on more than one independent variable (Logan 2004); the derivatives in PDEs are partial in the independent variables.

Partial differential equations (PDEs) have long been used for modeling, analyzing, and simulating continuous physical phenomena as well as spatial phenomena (Tobler 1981; Giudici 2002; Mitasova and Mitas 2002). PDEs are widely applicable, because they show how processes evolve.

In this paper, PDEs are used for describing generic or theoretical information about spatial processes. Theoretical characteristics of PDEs are therefore more important here than computational issues related to PDEs. The focus is on basic, linear PDEs of at most third order. In the theory on PDEs, three main types of processes are differentiated: wave-like, diffusion-like, and steady-state processes. Different equations are used for describing these types of processes. The following specifications of the types of processes are based on (Logan 2004).

The types of equations used for modeling *wave-like processes*, are hyperbolic PDEs. These equations are evolution equations and model how a process evolves over time. One example for a wave-like process is advection or convection. The advection process describes the bulk movement of particles in some transporting medium like water or air. A boat floating downstream in a river is an example for an advection process.

Diffusion-like processes are modeled with parabolic equations, which are evolution equations like the hyperbolic equations. Diffusion describes the random motion of particles, which generally diffuse from regions with a higher to regions with a lower concentration of particles. The example of a contaminant diffusing in water is discussed in section 5 of this paper.

In the case of a *steady-state process*, we deal with an elliptic equation. These types of equations do not contain a time variable and are therefore independent of time. They are known as equilibrium equations that model processes like the steady-state flow in fields where a balance between input and output in the systems exists. An example for a steady-state process is the flow of groundwater in a certain region with fixed boundary conditions.

An important difference between wave-like and diffusion-like processes is how the quantity of interest is affected over time. Wave-like processes preserve the quantity, whereas diffusion-like processes tend to smear out the initial configuration of the quantity. Wave-like and diffusion-like phenomena are two important types of phenomena that occur in different disciplines. Combinations of these two types of motions are also possible (Table 1).

An important part of the methodology in this work is relating difference equations to PDEs and thereby deriving theoretical insights about a process. For this purpose, a list of linear PDEs was compiled based on (Hohage 2004; Logan 2004; Markowich 2007). The recurrence of equations in the different sources suggests that it gives an overview of basic linear PDEs, although the list may not be complete.

Table 1. Linear PDEs assigned to the three main types of processes

Types of phenomena	Type of equation	PDEs
Wave-like phenomena	Hyperbolic	
		Wave equation Advection equation Advection-decay equation McKendrick or von Foerster equation Continuity equation Boltzmann equation
Diffusion-like phenomena	Parabolic	
		Heat equation Diffusion equation Diffusion-decay equation/ Diffusion-growth equation Advection-diffusion equation Advection-diffusion-decay equation Continuity equation
Steady-state phenomena	Elliptic	
		Poisson’s equation Laplace’s equation Helmholtz equation

5. Example: diffusion of a contaminant in water

The two types of mathematical models introduced in the previous section, are now applied to the specific example of the diffusion of a contaminant in water. Section 5.1 gives the deterministic block model of the process as difference equations. This block model provides a conceptualization of the process. In section 5.2 the difference equations are related to the corresponding partial differential equation, which sheds more light on generic properties of the process. The insights about the example process are discussed in section 6.

5.1. A Block Model of the Example Process

The derivation of the conceptual deterministic model of the diffusion of a contaminant in water is based on examples discussed in (Thomas and Huggett 1980). We assume that the diffusion of a contaminant follows the law of mass conservation. Fick's law of diffusion defines the rate at which the contaminant diffuses along the contaminant concentration gradient.

The spatial domain in which the process takes place is an enclosed and stationary water body like a basin. We divide the water body into a set of blocks that are placed one next to another and also one above and below another. The following storage equation is formulated for a block surrounded by other blocks at all of its six faces. The contaminant can enter the block of interest from any of its six faces. There is a certain amount of the contaminant in the water body, which is conserved under the law of mass conservation, and no sources or sinks of contaminants exist in this example. The symbols used in the following equations are:

C … concentration of the contaminant
D … contaminant diffusion coefficient
A … area of a face of the block
V … volume of a block
f … flow rate due to diffusion
f_i … contaminant inflow in a block
f_o … contaminant outflow of a block
Δ … a difference
Δq … change in contaminant storage in a block in a time interval
Δt … time interval
Δx … distance interval in x direction
Δy … distance interval in y direction
Δz … distance interval in z direction

The change of the contaminant storage in a block of water over a time interval is specified by the following equation (Eq. 5.1.1):

$$\Delta q = \left(f_i - f_o \right) A \, \Delta t \tag{5.1.1}$$

The input and output of the contaminant are due to diffusion, which is the "movement of [the contaminant] along the concentration gradient between two blocks" (Thomas and Huggett 1980, p.119). Fick's law defines the flow rate in the case of diffusion as "proportional to the negative gradient of [contaminant concentration] through the face of the block" (Thomas and Huggett 1980, p.119). Eq. 5.1.2 states the flow rate in all directions of a block.

$$f = (f_i - f_o) = D\left(-\frac{\Delta C}{\Delta x}\right) + D\left(-\frac{\Delta C}{\Delta y}\right) + D\left(-\frac{\Delta C}{\Delta z}\right) \tag{5.1.2}$$

A second way of identifying the changes in the storage of the contaminant is multiplying the change in the concentration of the contaminant by the volume of the block. The change in the contaminant concentration corresponds to the difference in the contaminant concentration at the beginning and at the end of a time interval (Eq. 5.1.3).

$$\Delta q = \Delta C\ V \tag{5.1.3}$$

We equate the two equations describing the change in the storage of the contaminant (Eq. 5.1.1, Eq. 51.3), simplify them and get:

$$\frac{\Delta C}{\Delta t} = -\left[\frac{\Delta}{\Delta x}\left(-\frac{D\,\Delta C}{\Delta x}\right) + \frac{\Delta}{\Delta y}\left(-\frac{D\,\Delta C}{\Delta y}\right) + \frac{\Delta}{\Delta z}\left(-\frac{D\,\Delta C}{\Delta z}\right)\right] \tag{5.1.4}$$

Eq. 5.1.4 can be rewritten in the following way:

$$\frac{\Delta C}{\Delta t} = D\left(\frac{\Delta^2 C}{\Delta x^2} + \frac{\Delta^2 C}{\Delta y^2} + \frac{\Delta^2 C}{\Delta z^2}\right) \tag{5.1.5}$$

This difference equation (Eq. 5.1.5) describes the diffusion of a contaminant in a water basin. The conceptual model is complete with the derivation of the difference equation. For an actual simulation of the problem, parameters, initial conditions and boundary conditions would have to be specified.

5.2. Relating the Block Model to a PDE

In section 5.1 we derived a deterministic model of the process of diffusion of a contaminant in a water basin. This model is based on discrete temporal and spatial units, with the spatial units being blocks. If we imagine we make these units smaller and smaller until they are infinitely small, we can understand the difference equation as a continuous differential equation. This idea is used here for relating the deterministic model of a process to a PDE; the set of PDEs available was listed in Table 1. Usually, this procedure is used in the reverse, and continuous PDEs are approximated with difference equations to find their solution. Seen from a conceptual, rather than a mathematically precise point of view, the difference equation (Eq. 5.1.5) corresponds to the following partial differential equation (Eq. 5.2.1):

$$\frac{\partial C}{\partial t} = D\left(\frac{\partial^2 C}{\partial x^2} + \frac{\partial^2 C}{\partial y^2} + \frac{\partial^2 C}{\partial z^2}\right) \tag{5.2.1}$$

This partial differential equation is known as the diffusion equation without sources in three dimensions. This equation is a second order, linear PDE. The meaning of the PDE is of course equivalent to the meaning of the difference equation derived in the previous section, but independent of the block structure assumed in section 5.1. The state variable *C(x,y,z,t)* depends on time and three spatial dimensions. It defines the concentration of the contaminant in space over time. The right hand side of the equation defines the flow of the quantity at a certain moment in time. The next section discusses the insights we gain through applying the presented methodology on the example process.

6 Qualitative insights about the example process

As a physical spatial process, the process adheres to physical principles of mass conservation and Fick's law for determining the rate of diffusion. Particles always diffuse from areas with higher to those with lower concentrations of particles; if there is no difference in the concentration of particles between two blocks, no flow takes place. Expressed differently, diffusion takes place down the concentration gradient of the contaminant in water.

The modeling of the example process with difference equations, allows us to easily determine the components that make up a process. In our example we have:

- the concentration of the contaminant, which depends on time and space and is defined by the flow of the contaminant over time.
- the flow of the contaminant, which is specified by the contaminant diffusion coefficient times the negative contaminant concentration gradient across a face.

The related partial differential equation is the diffusion equation, which is a parabolic equation showing the behavior of a process over time. The behavior of the process we can expect is that the original contaminant concentration spreads throughout the cells in a random manner and gets lower in specific cells the more blocks are affected by the process. An exemplary two-dimensional representation of the behavior of this process is shown in Fig. 2.

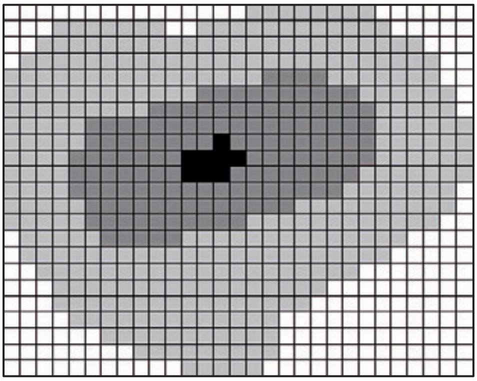


Fig. 2. Exemplary two-dimensional representation of a substance diffusing in space; the darker the color the higher the concentration of the substance.

7 Conclusions and future work

In this paper we presented our approach to gaining insight in spatial physical processes on a qualitative level. We employ simple deterministic models based on blocks for modeling processes and getting conceptualizations of processes. The resulting model equations are then related to types of partial differential equations (Table 1).

The methodology was presented exemplarily for the process of contaminant diffusion in water and general properties of this process were identified. In a next step, this methodology will be applied repeatedly to different spatial physical processes; thereby a catalog of components for describing qualitative characteristics of processes will be created. This catalog will allow the domain and application independent description of physical spatial processes on a qualitative level. One characteristic of the components in the catalog has to be that they can be composed in order to form more complex processes out of basic components.

Having in mind the extension of models of space and GIS with general functionality for handling processes, such a catalog could serve the following purpose in the long term: it could serve as a toolbox for describing a process qualitatively and for generating a concept for modeling a process. Such a concept of a process model would contain the required data sets, parameters and equations describing the general behavior of a process. In the case of the example of the diffusion of a contaminant in water, the required data set is the distribution of the contaminant, which serves as initial condition. For solving the diffusion equation, the diffusion parameter and boundary conditions have to be given.

Acknowledgements

Barbara Hofer is a recipient of a DOC-fFORTE-fellowship of the Austrian Academy of Sciences at the Department of Geoinformation and Cartography of the Vienna University of Technology.

References

Abel DJ, Taylor K, Kuo D (1997) Integrating modelling systems for environmental management information systems. SIGMOD Record. 26: 5-10

Abler R, Adams JS, Gould P (1977) Spatial Organization: the Geographer's View of the World. Prentice-Hall International, London

Albrecht J (1998) Universal Analytical GIS Operations - A task-oriented systematisation of data-structure-independent GIS functionality. In: Craglia M, Onsrud H (eds.) Geographic Information Research - Trans-Atlantic Perspectives. Taylor & Francis, London. 577-591

Batty M (2005) Network Geography: Relations, Interactions, Scaling and Spatial Processes in GIS. In: Unwin D, Fisher P (eds.) Re-Presenting GIS. Wiley, Chichester, UK.

Batty M, Xie Y, Sun Z (1999) Modeling urban dynamics through GIS-based cellular automata. Computers, Environment and Urban Systems 23(3): 205-233

Bivand RS, Lucas AE (2000) Integrating Models and Geographical Information Systems. In: Openshaw S, Abrahart RJ (eds.) GeoComputation. Taylor & Francis, London. 331-363

Blok C (2000) Monitoring Change: Characteristics of Dynamic Geo-spatial Phenomena for Visual Exploration. In: Freksa C, Brauer W, Habel C, Wender KF (eds.) Spatial Cognition II, Integrating Abstract Theories, Empirical Studies, Formal Methods, and Practical Applications. Springer-Verlag, Berlin Heidelberg. 1849: 16-30

Cliff AD, Ord JK (1981) Spatial Processes - Models & Applications. Pion, London

Frank AU (1998) GIS for politics. CD-ROM Proceedings of the GIS Planet 1998 Conference, Lisbon, Portugal

Frank AU (2001) Tiers of ontology and consistency constraints in geographic information systems. International Journal of Geographical Information Science (IJGIS) 75(5): 667-678

Frank AU (2007) Material vs. information causation-An ontological clarification for the information society. Wittgenstein Symposium in August 2007, Kirchberg, Austria: 5-11

Getis A, Boots B (1978) Models of Spatial Processes - An approach to the study of point, line and area patterns. Cambridge University Press, Cambridge

Giudici M (2002) Development, Calibration, and Validation of Physical Models. In: Clarke K, Parks B, Crane M (eds.) Geographic Information Systems and Environmental Modeling. Prentice Hall, Upper Saddle River, N.J.: 100-121

Goodchild MF, Yuan M, Cova TJ (2007) Towards a general theory of geographic representation in GIS. International Journal of Geographical Information Science (IJGIS) 21(3): 239-260

Hohage T (2004) Skriptum - Partielle Differentialgleichungen. Göttingen, Universität.

Hornsby K (1996) Spatial diffusion: conceptualizations and formalizations. NCGIS Specialist Meeting on Formal Models of Commonsense Geographic Worlds, San Marcos, Texas

Hornsby K, Egenhofer MJ (1997) Qualitative Representation of Change. In: Hirtle SC, Frank AU (eds.) Spatial Information Theory - A Theoretical Basis for GIS (International Conference COSIT'97). Springer-Verlag, Berlin-Heidelberg. 1329: 15-33

Kemp KK (1992) Environmental Modeling with GIS: A Strategy for Dealing with Spatial Continuity. Ph.D. thesis, University of California, Santa Barbara.

Logan JD (2004) Applied Partial Differential Equations. Springer-Verlag, New York

MacEachren AM (1995) How Maps Work - Representation, Visualization and Design. Guilford Press, New York

Markowich PA (2007) Applied Partial Differential Equations: A Visual Approach. Springer Verlag, Berlin Heidelberg

Miller HJ, Wentz EA (2003) Representation and spatial analysis in geographic information systems. Annals of the Association of American Geographers 93(3): 574-594

Mitasova H, Mitas L (2002) Modeling Physical Systems. In: Clarke K, Parks B, Crane M (eds.) Geographic Information Systems and Environmental Modeling. Prentice Hall, Upper Saddle River, N.J.: 189-210

Reitsma F, Albrecht J (2005) Implementing a new data model for simulation processes. International Journal of Geographical Information Science (IJGIS) 19(10): 1073-1090

Seppelt R (2005) Simulating invasions in fragmented habitats: theoretical considerations, a simple example and some general implications. Ecological Complexity 2(3): 219-231

Thomas RW, Huggett RJ (1980) Modelling in Geography - A Mathematical Approach. Harper & Row, London

Tobler W (1970) A computer movie simulating urban growth in the Detroit region. Economic Geography 46(2): 234-240

Tobler WR (1981) A Model of geographical movement. Geographical Analysis 13(1): 1-20

Tomlin CD (1990) Geographic Information Systems and Cartographic Modeling. Prentice Hall, New York

Van Deursen WPA (1995) Geographical Information Systens and Dynamic Models. Ph.D. thesis, NGS Publication, Faculty of Spatial Sciences. University of Utrecht, Utrecht.

Worboys MF (2005) Event-oriented approaches to geographic phenomena. International Journal of Geographical Information Science (IJGIS) 19(1): 1-28

Yuan M (2001) Representing complex geographic phenomena in GIS. Cartography and Geographic Information Science 28(2): 83-96

A Data Model for Multi-scale Topographical Data

J.E. Stoter[1], J.M. Morales[1], R.L.G. Lemmens[1], B.M. Meijers[2], P.J.M van Oosterom[2], C.W. Quak[2], H.T. Uitermark[3], L. van den Brink[4]

[1] ITC, P.O. Box 6, 7500 AA Enschede, The Netherlands.
e-mail: {stoter|morales|lemmens}@itc.nl
[2] Delft University of Technology, P.O. Box 5030, 2600 GA Delft, The Netherlands.
e-mail: {b.m.meijers|c.w.quak|p.j.m.vanoosterom}@tudelft.nl
[3] Kadaster, P.O. Box 9046, 7300 GH Apeldoorn, The Netherlands.
e-mail: harry.uitermark@kadaster.nl
[4] Dynasol BV, Buffelstraat 124a, 3064 AD Rotterdam, The Netherlands.
e-mail: linda@dynasol.nl

Abstract

The lack of fully automated generalisation forces National Mapping Agencies to maintain topographical data sets at different map scales. For consistency between map scales, but also for supporting (future) automated generalisation processes, information on similarities and differences of the separate data sets should be identified and formalised. This includes information on valid data content at the different scales (‘scale state’), but as important is the semantics of multi-scale and generalisation aspects (‘scale event’). As ‘scale state’ and ‘scale event’ are strongly related (‘different sides of the same coin’) it is important to integrate these in a single model. This paper presents a semantically-rich data model for an integrated topographical database, facilitating (semi-)automated generalisation. UML (including OCL) is used to formalise the model. The scope of the model is outlined and the model is presented based on an analysis of several alternatives for modelling multi-scale and generalisation aspects. The model is evaluated by instantiating the model and applying it to test data.

Keywords: multi-scale, spatial data modelling, generalisation, knowledge formalisation

1. Introduction

Integration of topographical data sets at different map scales is an important requirement for implementing automatic generation of (updates in) smaller scale data sets from (updates in) larger scale data sets. For implementation of this integration within a DBMS as well as of the implementation of (semi-)automated generalisation processes, a data model is needed which makes all required information and knowledge explicit in a formalised way. This comprises firstly information on data content covering all scales (as in information model design for a single data set): object classes, attributes, attribute values, constraints for valid data content and relationships between object classes within one map scale (‘map scale state’). For supporting generalisation additional semantics on multi-scale aspects (‘scale event’) is required, such as how object classes and instances behave at map scale transitions and relationships between object classes and instances at different map scales. This paper presents a semantically-rich integrated data model for multi-scale topographical data sets, maintained by the Kadaster (Dutch Land Registry Office) facilitating (semi)automated generalisation between topographical data sets at different map scales. The designed multi-scale data model is called Information Model TOPography (IMTOP). The Unified Modelling Language (UML), including the Object Constraint Language (OCL), is used to formalise the model.

In Section 2 previous initiatives on multi-scale data modelling (2.1) as well as the three basic spatial data models (2.2) are presented. Section 3 defines the scope of IMTOP and presents the requirements for IMTOP. Section 4 describes the various steps that have been taken to design the multi-scale data model. The model is evaluated in Section 5 by applying the model to test cases. The paper ends with conclusions in Section 6.

2. Previous approaches for multi-scale and single data models

2.1 Data modelling approaches for multi-scale data

A multi-scale data model is a specific type of a multi-representation data model. The issue of multi-representation was introduced in a research program of the National Center for Geographic Information and Analysis (NCGIA 1989; Buttenfield and Delotto 1989). Since then many researchers have focused on this issue. The Multiple Representation Management System (MRMS) of (Friis-Christensen and Jensen 2003) provides MR-methods such as ‘checkConsistency’ and ‘restoreConsistency’, as well as triggers to execute those methods in case of updates and insertions, modelled with UML and OCL and implemented on top of Oracle. The Modeling of Application Data with Spatio-temporal features (MADS) of (Parent et al. 2006) is based on stamps. One or several stamps can be assigned to object classes, relationships, attributes, values, etc. to indicate for which map scale the object class, etc. is relevant. ‘Perceptory’ (Bédard et al. 2004) is a plugin extending existing UML editors with spatio-temporal icons allowing modelling of multi-representation concepts in methods of the object classes. There is no independent description of multi-representation concepts, as in MADS. Jones et al. (1996) propose a conceptual model for a multi-representation database as a single database that is capable of storing spatial objects with multiple geometries. This approach does not take into account the complexity of the relationships that can exist in multi-representation (and multi-scale) data sets. The work of Devogele et al. (1996) models map scale transitions, but only between pairs of objects; it does not consider a complete topographical database. The work of Kilpelainen (1997) focuses on the link between object instances when there is an exact dependence among the object classes (e.g. building as complex polygon, building as simple polygon, building as point, building as part of a building area). What is new in the research presented in this paper is that the data model formalises all knowledge required for both integration and for automated generalisation of topographical data sets.

2.2 Three basic approaches for spatial data models

When looking at spatial modelling in the past, three main approaches can be distinguished: 1) geometry/topology-first approach, 2) object-first approach, and 3) a hybrid approach.Geometry is the main entrance for object classes in the geometry-first approach, often structured in a topological structure (e.g. a linear network, or a partition of space). Attributes are added to these geometries in order to classify the objects.

The object-first approach models the object classes first with added geometry attributes. Every object class can have its own set of thematic attributes which may vary for the different object classes. Every object class has its own geometric description independent of any other object. The model does not explicitly contain topological relationships, which are very important for generalisation; e.g. what are the neighbours of this instance (candidates for aggregation)? is the network connectivity damaged when this road segment is removed? etc.

The third approach, the hybrid approach, treats geometry and the object class equally. It combines the strengths of both approaches: the thematic attributes are specifically designed for every object class, but the model also enables shared geometry and use of embedded structures. The spatial domain is a full partition and the result is described using tables for nodes, edges, and faces (and solids in 3D). The instances are modelled in the same way as in the object-first approach with the exception that objects do not have their own independent geometry attributes, but refer to primitives in the geometry/topology part of the model (node, edge, face,…). This is the approach as described in the 'formal data structure' (FDS) theory of Molenaar (1989) and quite recently implemented in products such as 1Spatial's (formerly LaserScan) Radius Topology, and Oracle's spatial topology (first introduced in version 10g). It cannot be claimed that one model is 'better' than another model. This depends on the application context and use. If one specifies a number of important characteristic of the application domain and typical use, then it is possible to state which approach is preferred (Stoter et al. 2007).

3. Scope of IMTOP

3.1 Previous initiatives on multi-scale models as input for IMTOP

What the Information Model TOPography (IMTOP) adds to past research is that the model specifically focuses on data sets that cover the same reality using a similar set of object classes. It models how classes, as well as their instances, change at map scale transitions. In that sense IMTOP does not model different (=multi) representations of real world objects. Instead

IMTOP models one collection of object classes together with semantically-rich information on scale transitions. Another specific aspect of IMTOP is that complete topographical coverage at every scale needs to be modelled (there are no gaps), which is more inclusive than defining inheritance relationships between object classes as in most multi-representation approaches. An object cannot be eliminated without being merged with another object because of the topological structure at every scale. For example, in TOP10NL (1:10k base map of The Netherlands) faces of the topological structure consist of road polygons, water polygons and land use. In TOP50NL (the 1:50k map), and smaller scales (TOP100NL, TOP250NL, and TOP500NL), road polygons are collapsed, and therefore faces in topological structure are formed by water polygons and land use.

For IMTOP the hybrid approach (Section 2.2) was identified as optimal approach. Some criteria justify the object-first approach with functionality also supported in the hybrid approach, but not in the geometry-first approach: bridge over water should be allowed; administrative area can overlap topographical objects; multi-geometry of objects should be possible, e.g. both center lines and polygons for roads. On the other hand topological structure as available in the geometry-first and hybrid-approach, but not in the object-first approach is needed for automated generalisation. Therefore IMTOP is based on the hybrid approach and topological primitives are used to model geometries. The model will adhere to ISO and OGC standards as much as possible.

3.2 Integration of landscape model and cartographic model

For every map scale, Kadaster supplies two products that should be covered by IMTOP: a data set for GIS analyses and a digital map, which is a cartographic version of the data set. It is not trivial to answer the question whether a so-called Digital Landscape Model that does not take into account any symbolisation (DLM) as well as a so-called Digital Cartographic Model (DCM) should be available (and modelled) at every scale. Current TOPxxvector data sets as supplied by Kadaster integrate DLM and DCM aspects: the geometries in the vector data sets take already into account the way they will appear on the map. For example, a motorway in TOP50vector will be portrayed with a line-symbol of width 1.5 mm, which is 75 meter in reality. To avoid overlap of the motorway symbol with other instances such as buildings, instances are displaced and simplified in current TOPxxvector products. Creating the map is so to speak a simple button push, which adds symbology to the geometries (see Figure 1).

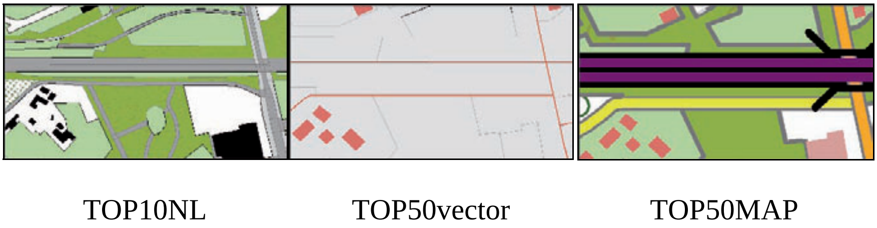


Fig. 1. TOP10NL, TOP50vector and TOP50MAP in current production process of Kadaster

From a theoretical point of view it seems straightforward to distinguish between database and cartographic representation, since inaccuracies because of symbolisation are avoided in the database. However, for IMTOP it was decided to integrate the landscape and cartographic model. The database product is therefore a vector representation of the map. This approach is also applied by for example KMS, Denmark (West-Nielsen and Meyer 2007). There are several arguments which favour this approach:

- Generalisation leads to loosing accuracies, whether this is for the database or the map. The inaccuracies of current vector products are no problem for GIS analyses at small scales. If more accurate data is needed, one can use TOP10NL where symbolisation does not yield major graphical conflicts and therefore does not yield inaccuracies.
- It was tried for IMTOP to separate between DLM and DCM, in contrast to current practice. This showed that for many transitions it is not easy to identify where they fit, e.g. elimination of small buildings.
- A multi-scale topographical database requires keeping data models as well as databases at all scales consistent. Separation between model and cartographic representations requires twice as many data models and databases to keep consistent.

3.3 Basic requirements for IMTOP

Main objective of IMTOP is supporting semi-automated generalisation of medium to small map scales. The criteria that follow from this requirement comprise the possibilities:

1. To have a model describing topographical data in the scale range 1:10k to 1:1,000k at specific scales (e.g. 1:10k, 1:50k).
2. To extract a UML class diagram for a specific map scale.

3. To produce a GML application schema (.xsd) for a specific map scale from the UML diagrams generated in step 2.
4. To produce a GML application schema (.xsd) for multi-scale topographical data.
5. To generate the DBMS structure for a specific map scale from every schema generated in step 3.
6. To generate the DBMS structure for multi-scale topographical data, including scale transition information for instances and object classes.
7. To populate the multi-scale database and link instances at different map scales.
8. To use all information (including supporting structures, map scale and transition information) directly in the generalisation process, with minor human intervention.

The model was tested on these requirements to see to what extent IMTOP is suitable for multi-scale data modelling taking generalisation aspects into account. Results are reported in Section 5.

4. A data model for multi-scale topographical data

IMTOP is a result of several steps:

1. Designing data models at separate map scales in UML including refined semantics expressed in OCL.
2. Integration of data models at different scales to model scale transitions of object classes that are apply to a complete set of an object class, e.g. conversion of geometry types (collapse or combine), or reclassification, for example ‘streets’ and ‘local roads’ are eliminated at scale 1:250k, and smaller.
3. Extend the model with semantics on transitions that apply only to specific instances.

These steps are described in respectively Sections 4.1, 4.2 and 4.3.

4.1 Modelling object classes, attributes, and relationships at separate scales

An analysis of current product specifications for TOPxxvector products (Kadaster, 2002) showed which object classes, attributes, attribute values, and geometry types (which could be multiple geometries) should be modelled in IMTOP. Separate UML diagrams were designed per map scale to cover this information. For generalisation, the diagrams at the separate scales should also express how the specific data as generalisation output should look like. This information is currently available in generalisation specifications, software code or even in human minds; only at the human knowledge level, since the information is meant for cartographers to be used in interactive generalisation processes. From current generalisation specifications (Kadaster, 2006) generalisation-related constraints within and between different object classes were deduced, and added to the diagrams in OCL. In a case study it was evaluated to what extent this information can serve the automated generalisation process (see Section 5).

The constraints that were defined in this process are called *invariants* in OCL terminology and express a valid state of the data set at a certain map scale. This is the ‘scale state’ aspect of the model. Based on these constraints, a data set at a particular level of detail can be validated (e.g. all buildings have a minimum size at scale 1:50k).

Specifying valid contents of the model is the known role of constraints in data modelling. Constraints are used more and more to include additional semantics in data models (Louwsma et al., 2006; Oosterom, 2006). In IMTOP, constraints addressing semantics on transition processes were added in a next phase: this is the ‘scale event’ aspect of the model. For that phase, constraints are expressed as part of the change from one level of detail to the next, i.e. as *pre conditions* and *post conditions* of the transition process, for respectively selecting appropriate large scale input (i.e. instances) and for checking resulting smaller scale output which might trigger another event. The result of the scale transition can be stored via associations between larger and smaller scale object classes and instances. Therefore, post conditions of scale transitions are related to the invariants (constraints at the resulting scale). See Section 4.3 for pre and post condition constraints.

In this section, constraints of the ‘invariant’ type are presented, thus defining which data content is allowed, only addressing aspects within one scale. The conceptual modelling of these constraints using OCL is illustrated in Figure 2, which depicts an excerpt of the model for the classes ‘Building’ and ‘Road’.

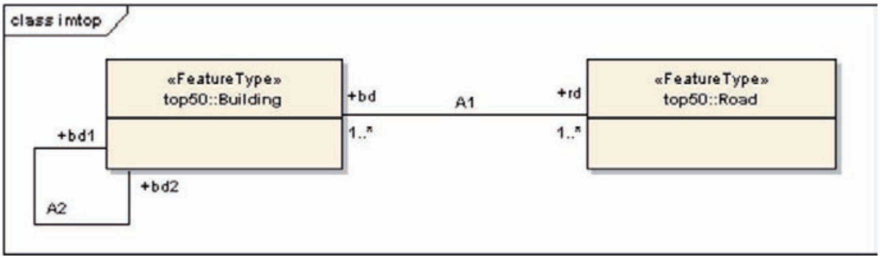


Fig. 2. UML class diagram of object classes ‘Building’ and ‘Road’

We distinguish four types of spatial relationships constraints related to the partial model in figure 2:

- Type 1: constraints on a single instance from a single class.
- Type 2: constraints between two instances belonging to the same class, e.g. between two buildings.
- Type 3: constraints between two instances belonging to different classes, e.g. between building and road.
- Type 4: constraints on a group of instances, e.g. group of buildings.

In the UML class diagrams, constraints of types 2, 3 and 4 are modelled as constraints which navigate through associations. In the example of Figure 2 an association A1 is added between Building class and Road class for TOP50NL, with the following constraint in OCL (of Type 3) defining that roads and buildings must be disjoint:

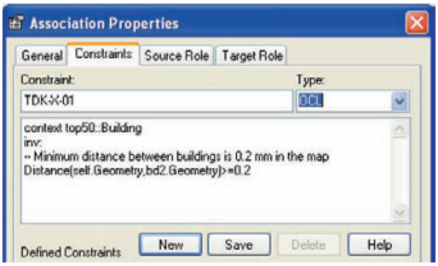
```
context top50::Building
inv:
  -- Building and Road should be disjoint
  Disjoint(self.Geometry,rd.Geometry)
```

Note that the predicate ‘Disjoint’ can be evaluated because of the topological structure available in the hybrid-approach. A constraint of Type 2 is defined through association A2, identifying a minimum distance between two symbolised buildings:

```
context top50::Building
inv:
  -- Minimum distance between buildings is 0.2 mm in
the map
  Distance(self.Geometry,bd2.Geometry)>=0.2
```

An example of a Type 4 constraint is the constraint that building instances on area of land use type ‘other’ should never exceed 10% of the area coverage. The action to be taken if objects do not adhere to the constraint (e.g. ‘displace’) are not modelled in this part of the model, as this is an issue of the transition process between the different map scales (see Section 4.3). Functions such as ‘Disjoint’, and ‘Distance’ as used in these examples are

assumed to have standardised implementations in software. Standards from ISO and the Open GeoSpatial Consortium (OGC) are used as much as possible.
As stated before, these example invariant constraints will also be expressed as post conditions of scale transitions. Figure 3 shows how constraints appear in the UML modelling software used in this research (Enterprise Architect, 2008).



Constraint: TDK-X-01

Type: OCL

```
context top50::Building
inv:
-- Minimum distance between buildings is 0.2 mm in the map
Distance(self.Geometry,bd2.Geometry)>=0.2
```

Fig. 3. Appearance of OCL constraints in the UML modelling software

4.2 Integration of data models at different map scales

The integration of the separate data models should make explicit what happens to object classes at scale transitions. For IMTOP these scale transitions are identified from product and generalisation specifications (Kadaster, 2002; Kadaster 2006). An analysis of the data models as generated in step 1, showed that going to smaller scales does not only lead to reducing information, but sometimes adding information. An example is the attribute value ‘roundabout’ for TypeOfInfrastructure, which is not registered at map scale 1:10k (since roads are represented by polygons) but is needed at map scale 1:50k, where roads are collapsed and roundabouts are represented by points. There are basically three alternatives for the integration which are discussed below.

Constraints for scale dependent attributes and attribute values

The appearance of the object class at a specific map scale is defined by constraints to allow or disallow attributes and attribute values at specific scales. For example ‘if 1:50k then geometry type of ‘secondary roads’ is line’. The disadvantage of this approach is that the model will not be easy to read as a lot of OCL expressions have to be inspected and these OCL constraints are used both for modelling valid data content as well as for modelling scale dependent information. It is also not easy to automatically derive a model per scale.

Inheritance and derived attributes for scale dependent information

For every class that occurs in topography an abstract superclass is modelled containing attributes that are valid at the starting scale (TOP10NL in our case). A subclass is modelled as specialisation for TOP10NL, whereas all similar object classes at the other scales are modelled as a derived class from the previous scale. An example for the 'Road' object class is shown in Figure 4 for TOP10NL, TOP50NL and TOP100NL. Road classes at scales 1:50k and 1:100K contain derived attributes (indicated with '/'). The derivation rules can be modelled in OCL (for example: 'derive: derivedFromTOP50NL.typePavement' for typePavement in TOP100NL). Apart from derived attributes, the classes contain two other types of attributes: a) attributes that are introduced at this scale (e.g. 'exit' for Roads in TOP50NL) and b) attributes that disappear at this specific scale, indicated with multiplicity of 0 (e.g. 'geometrySurface' for Roads in TOP50NL and TOP100NL and 'exit' in TOP100NL). The inheritance approach is only used for object classes, and not for enumeration since inheritance is only appropriate for object classes according to ISO 19103. Advantages of this approach are that the model is easy to read and it is easy to get back to a model per map scale by just showing the relevant classes for that map scale only.

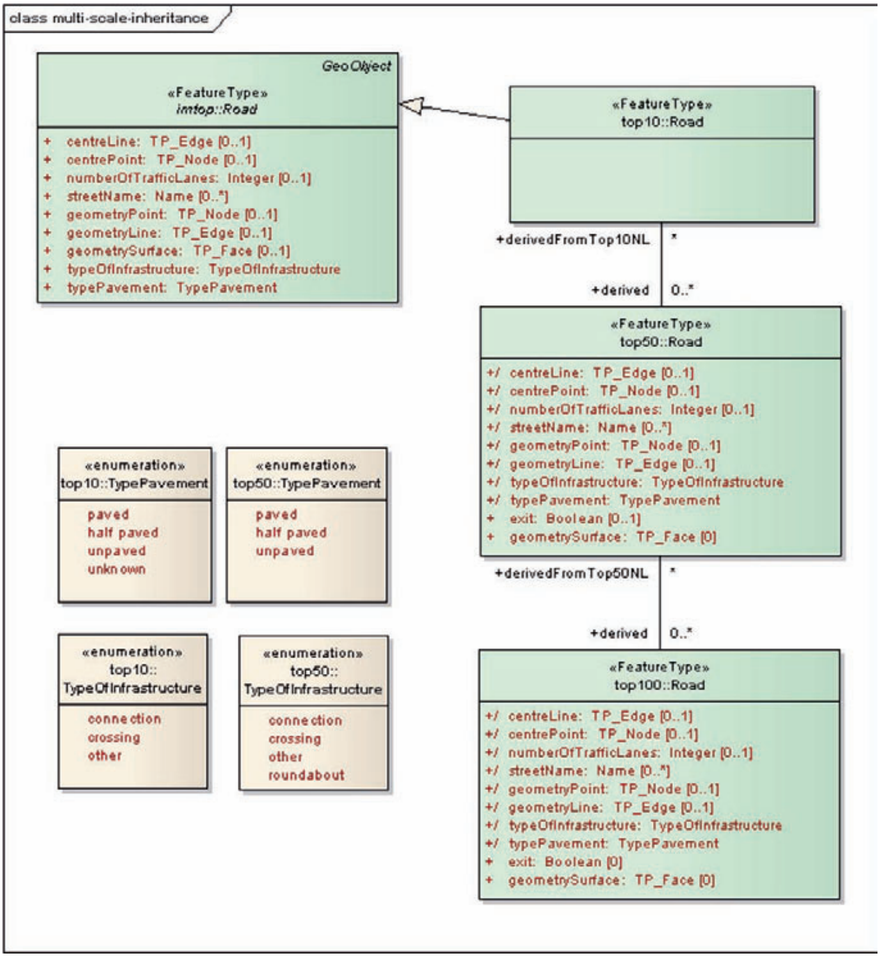
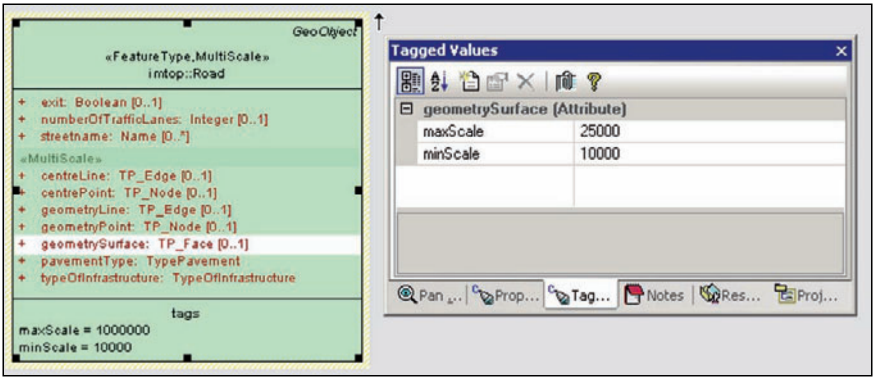

class multi-scale-inheritance
GeoObject
«FeatureType»
imtop::Road
+ centreLine: TP_Edge [0..1]
+ centrePoint: TP_Node [0..1]
+ numberOfTrafficLanes: Integer [0..1]
+ streetName: Name [0..*]
+ geometryPoint: TP_Node [0..1]
+ geometryLine: TP_Edge [0..1]
+ geometrySurface: TP_Face [0..1]
+ typeOfInfrastructure: TypeOfInfrastructure
+ typePavement: TypePavement
«FeatureType»
top10::Road
+derivedFromTop10NL
*
+derived
0..*
«FeatureType»
top50::Road
+/ centreLine: TP_Edge [0..1]
+/ centrePoint: TP_Node [0..1]
+/ numberOfTrafficLanes: Integer [0..1]
+/ streetName: Name [0..*]
+/ geometryPoint: TP_Node [0..1]
+/ geometryLine: TP_Edge [0..1]
+/ typeOfInfrastructure: TypeOfInfrastructure
+/ typePavement: TypePavement
+ exit: Boolean [0..1]
+ geometrySurface: TP_Face [0]
+derivedFromTop50NL
*
+derived
0..*
«FeatureType»
top100::Road
+/ centreLine: TP_Edge [0..1]
+/ centrePoint: TP_Node [0..1]
+/ numberOfTrafficLanes: Integer [0..1]
+/ streetName: Name [0..*]
+/ geometryPoint: TP_Node [0..1]
+/ geometryLine: TP_Edge [0..1]
+/ typeOfInfrastructure: TypeOfInfrastructure
+/ typePavement: TypePavement
+ exit: Boolean [0]
+ geometrySurface: TP_Face [0]
«enumeration»
top10::TypePavement
paved
half paved
unpaved
unknown
«enumeration»
top50::TypePavement
paved
half paved
unpaved
«enumeration»
top10::
TypeOfInfrastructure
connection
crossing
other
«enumeration»
top50::
TypeOfInfrastructure
connection
crossing
other
roundabout

Fig. 4. Using inheritance, derived attributes and multiplicity on attributes to define what is valid on every scale

Stereotypes for multi-scale semantics

In the third approach the underlying meta-model of UML was extended with multi-scale aspects. UML stereotypes and tagged values that can be used to model additional semantics in UML are applied to extend the UML model. The stereotype <<MultiScale>> is used to indicate that the given class will have different representations at different map scales (similar to stamps in MADS, see Section 2.1). Attributes and attribute values of MultiScale objects that are labelled with the <<MultiScale>> stereotype, get a 'minScale' and 'maxScale' tag to indicate on which map scales they are valid. In the example of Figure 5 all spatial attributes are stereotyped with the <<MultiScale>> stereotype and the correct minScale and maxScale tags are added. Disadvantage is that the tags are not visible in the class diagram itself, as can be seen in Figure 6. In addition, the model is very compact and it is therefore not easy to read: every object class (e.g. 'Road') is modelled as a single object class for all map scales and multi-scale aspects are only visible when analysing the specific attributes and attribute values. Finally it is not easy to automatically generate separate UML models from this model.



```
GeoObject
«FeatureType,MultiScale»
imtop::Road
+ exit: Boolean [0..1]
+ numberOfTrafficLanes: Integer [0..1]
+ streetname: Name [0..*]
«MultiScale»
+ centreLine: TP_Edge [0..1]
+ centrePoint: TP_Node [0..1]
+ geometryLine: TP_Edge [0..1]
+ geometryPoint: TP_Node [0..1]
+ geometrySurface: TP_Face [0..1]
+ pavementType: TypePavement
+ typeOfInfrastructure: TypeOfInfrastructure
tags
maxScale = 1000000
minScale = 10000
```

Tagged Values

geometrySurface (Attribute)	
maxScale	25000
minScale	10000

Fig. 5. Example of using a MultiScale stereotype to model map scale dependent information. The tags are only available in the GUI of Enterprise Architect

Based on the considerations outlined above the second approach was selected for IMTOP. Constraints are used to address valid data content whereas scale-related information is modelled with inheritance, which makes the model transparent. It is also possible to extract a data model for a specific map scale as will be seen from the tests in Section 5.

4.3 Modelling map scale transitions

With consistent sets of object classes throughout the various map scales, the model is completed with transition relationships, which further formalise knowledge on the generalisation process. Transition relationships represent associations between instances and object classes at different map scales. Semantically, these relationships represent transition paths for specific instances from one map scale to another. This as additional information to the information on transitions applied to all instances of an object class (see Section 4.2). To capture the transitions that can differ per instance, the model is extended with transition models for volatile transition classes and processes. Figure 6 shows a simplified example of a transition model for the transition path to generate *TOP50NL built-up area* instances from *TOP10NL* instances.

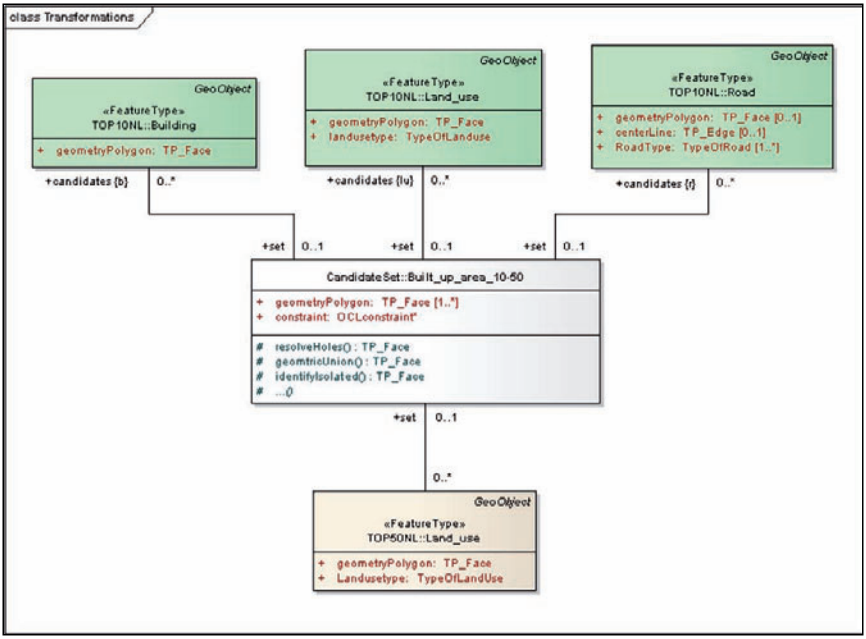

class Transformations
GeoObject
«FeatureType»
TOP10NL::Building
+ geometryPolygon: TP_Face
GeoObject
«FeatureType»
TOP10NL::Land_use
+ geometryPolygon: TP_Face
+ landusetype: TypeOfLanduse
GeoObject
«FeatureType»
TOP10NL::Road
+ geometryPolygon: TP_Face [0..1]
+ centerLine: TP_Edge [0..1]
+ RoadType: TypeOfRoad [1..*]
+candidates {b}
0..*
+candidates {lu}
0..*
+candidates {r}
0..*
+set
0..1
+set
0..1
+set
0..1
CandidateSet::Built_up_area_10-50
+ geometryPolygon: TP_Face [1..*]
+ constraint: OCLconstraint*
# resolveHoles() : TP_Face
# geomtricUnion() : TP_Face
# identifyIsolated() : TP_Face
# ...()
+set
0..1
0..*
GeoObject
«FeatureType»
TOP50NL::Land_use
+ geometryPolygon: TP_Face
+ Landusetype: TypeOfLandUse

Fig. 6. Model of transitions for generating built-up area in TOP50NL from TOP10NL buildings.

A transition process specifies firstly the selection of instances based on *pre conditions* and secondly actions that should be applied to those instances at map scale transition based on *post conditions* (which can be similar as the pre conditions used for selecting instances). Central concept in the process is the so-called “Candidate Set” class which is a container for instance sets derived from source classes that are potential members of a target class. The class has an attribute called *constraint*. This attribute is used to define pre conditions in the form of OCL constraints to identify instances from an object class at a source map scale that should go to the Candidate Set class. At scale transitions the constraints are evaluated to populate the Candidate Set class. Each instance in the Candidate Set class is a collection of objects. Each candidate set is determined by its own constraints. Consequently new constraints for other sets can easily be added afterwards. This is why pre condition constraints are defined as attributes of the Candidate

Set class and not on the class itself. Post conditions are defined on the Candidate Set class to specify how the instances in this class should be treated in the generalisation process. As was indicated in Section 4.2, these post conditions are similar to the invariant constraints defined for the separate scales. The transition models of IMTOP cover in the first place the currently available generalisation specifications (see also Section 4.1). In a case study it was evaluated to what extent this information specified in IMTOP can serve the automated generalisation process (see Section 5). In the future the transition models can be extended with new, machine-based knowledge on generalisation. The steps used for the transition in Figure 6 are:

1. Define all candidate instances from TOP10NL on their pre conditions:

   a. instances from *land use* class with *land use type* value ‘built-up area’, in OCL:

```
cs1    context cs:built_up_area_10-50
       cs.candidates→select( lu | lu.landusetype =
       ‘built-up area’)
```

   b. instances from the *land use* class with *land use type* value ‘other’ that contain buildings that cover more than 10% of the area of the land use object, in OCL:

```
cs2    context cs:built_up_area_10-50
       cs.candidates→select( lu | lu.landusetype =
       ‘other’
       and b→exists( b | b.geometryPolygon iscon-
       tainedby(lu.geometryPolygon)
       and area(b.geometryPolygon) >
       area(lu.geometryPolygon)*0.1))
```

   c. instances from the *road* class that touch any of the instances in cs1 or cs2 (as a consequence of the collapsing of road polygons to road centrelines), in OCL:

```
cs3    context cs:built_up_area_10-50
       cs.candidates→select( r | r.geometryPolygon
       touches(..cs1_constraint..)
       or rd.geometryPolygon touches(..cs2_constraint..))
```

   d. instances of the *land use* class of any *land use type* located in potential *built-up area* and smaller than *x* m^2, in OCL:

```
cs4    context cs:built_up_area_10-50
       cs.candidates→select( lu | lu.geometryPolygon
       iscontainedby(cs1.geometryPolygon or
       cs2.geometryPolygon)
       and area(b.geometryPolygon) < x)
```

2. Populate the candidate class with the indentified instances.
3. Trigger generalisation processes by post conditions defined on the Candidate Set class for TOP50NL built-up area. Note that post conditions a and d are related to pre conditions b (*cs2*) respectively d (*cs4*):
   a. Instances of land use type ‘other’ can be covered by buildings by at most 10%
   b. Holes are not allowed
   c. Topologically adjacent instances should have no boundary between them
   d. Small instances of *land use* are not allowed
4. Generate a new set of geometries by applying operations on the *candidate object* class based on the post conditions in step 3; assign instances to land use class in TOP50NL and update attribute values
5. Recycle instances that were part of the candidate object class but are not transformed into a TOP50NL land use class of type ‘*built-up area*’ after having finished the process.

UML state diagrams can be added to model the generalisation process in more detail (see Figure 7).

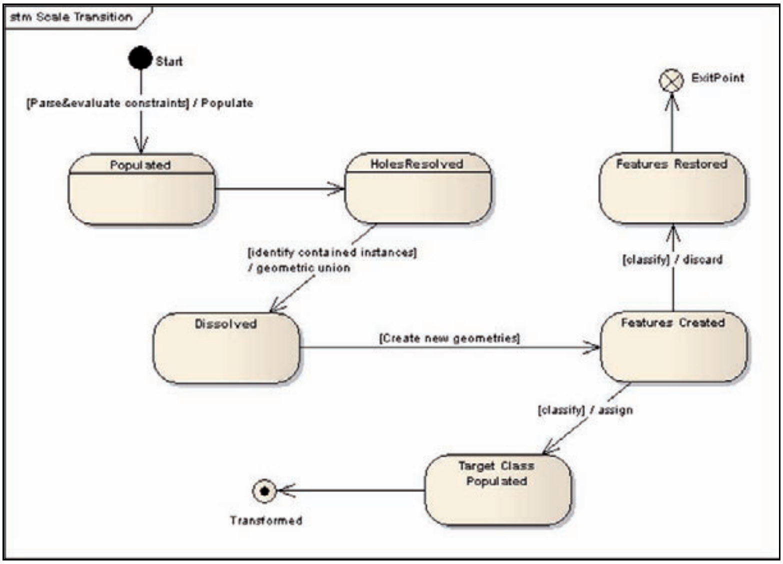
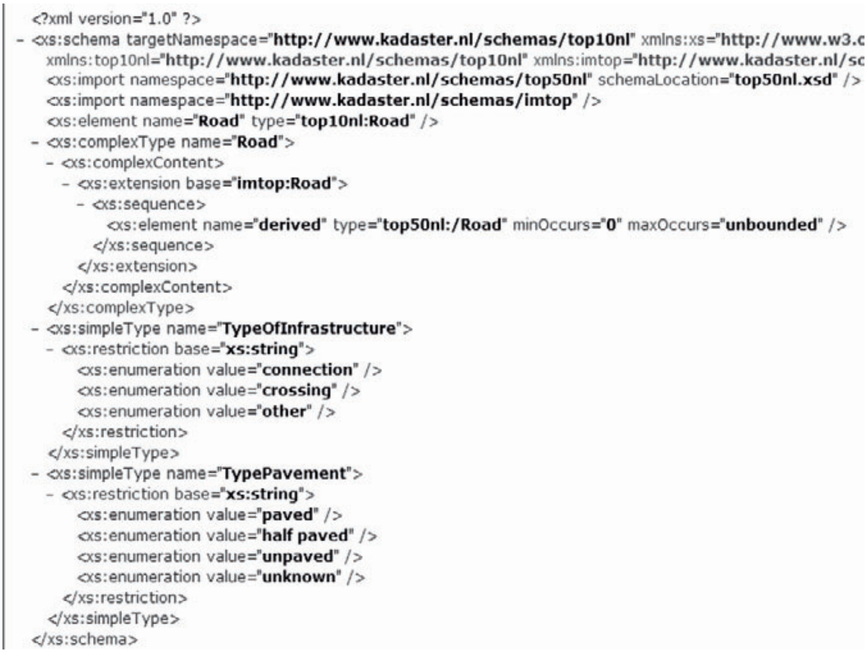


Fig. 7. State diagram for the generation of built-up areas

5. Results of IMTOP with respect to the requirements

Tests were carried out that instantiated the model in order to evaluate IMTOP on the requirements as defined in Section 3.3. From these tests the following conclusions can be drawn. From the model as selected in Section 4 (see Figure 4), it was easy to extract a UML model for a specific scale, since the separated scales could be modelled as separated packages in the model. In addition it was possible to extract an XSD file per map scale from the UML model, as can be seen from the XML fragment in Figure 8.



```
  <?xml version="1.0" ?>
- <xs:schema targetNamespace="http://www.kadaster.nl/schemas/top10nl" xmlns:xs="http://www.w3.c
    xmlns:top10nl="http://www.kadaster.nl/schemas/top10nl" xmlns:imtop="http://www.kadaster.nl/sc
    <xs:import namespace="http://www.kadaster.nl/schemas/top50nl" schemaLocation="top50nl.xsd" />
    <xs:import namespace="http://www.kadaster.nl/schemas/imtop" />
    <xs:element name="Road" type="top10nl:Road" />
  - <xs:complexType name="Road">
    - <xs:complexContent>
      - <xs:extension base="imtop:Road">
        - <xs:sequence>
            <xs:element name="derived" type="top50nl:/Road" minOccurs="0" maxOccurs="unbounded" />
          </xs:sequence>
        </xs:extension>
      </xs:complexContent>
    </xs:complexType>
  - <xs:simpleType name="TypeOfInfrastructure">
    - <xs:restriction base="xs:string">
        <xs:enumeration value="connection" />
        <xs:enumeration value="crossing" />
        <xs:enumeration value="other" />
      </xs:restriction>
    </xs:simpleType>
  - <xs:simpleType name="TypePavement">
    - <xs:restriction base="xs:string">
        <xs:enumeration value="paved" />
        <xs:enumeration value="half paved" />
        <xs:enumeration value="unpaved" />
        <xs:enumeration value="unknown" />
      </xs:restriction>
    </xs:simpleType>
  </xs:schema>
```

Fig. 8. XML fragment showing part of XSD file for TOP10NL, generated from IMTOP

An XSD file for a specific map scale refers to the XSD file for the general model. This reference is generated automatically and is triggered by the specialisation relation between classes in specific scale models and the general model. An XSD file for all map scales could not be easily generated, but is easy to build (once) by hand as its only content are references to all specific map scale XSD files and the general XSD file. The ISO primitives such as TP_edge which are used in the models are not automatically translated to the corresponding GML types. A simple type mapping was implemented for GML types using an Extensible Stylesheet Transformations (XSLT) stylesheet, and applied to the generated XSD files.

The UML model was exported into DDL (Data Definition Language) scripts that can generate the DBMS structure for the storage of the topographic data at the different scales. A relational DBMS with support for spatial types (PostGIS) was used. This experiment showed that the object-oriented hierarchy of a UML model is not easily mapped on flat DBMS tables as also concluded in (Sparks 2001). Therefore some database design choices were made in order to convert the UML model into a DBMS model. The following conversion rules were defined to automatically generate the DBMS model:

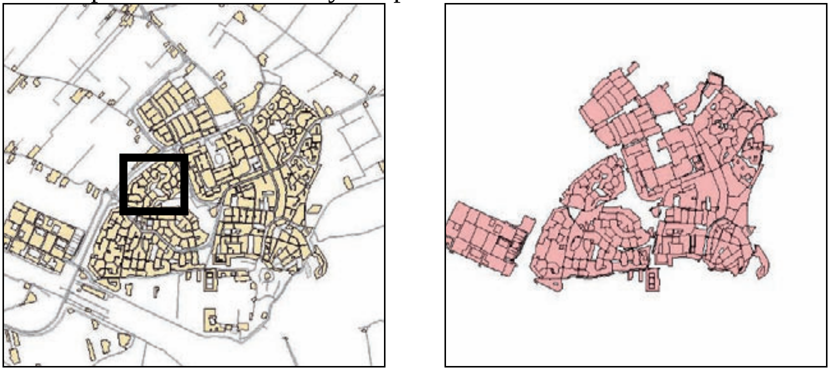
- Each non-abstract UML class (i.e. a class of which instances exist) is mapped to a DBMS table. The table contains columns for all the attributes of that class plus all the attributes of all superclasses.
- Each UML enumeration is mapped to a separate DBMS table and the use of an enumeration as an attribute is implemented as a foreign key to that table.
- References to topology items are implemented as foreign keys to a topological subsystem in the DBMS
- UML Associations are mapped to foreign key relations. 'Many to many' association need an intermediate table for the storage of the association.
- Multi-valued attributes are mapped to association tables.
- UML packages are mapped to DBMS table spaces.
- OCL constraints are mapped to Structured Query Language (SQL) views.

The conversion from a UML model to a DBMS structure (via DDL scripts) can be automated via a transformation language as shown in (Hespanha et al. 2008). The resulting DBMS model fits with the other requirements of IMTOP as defined in Section 3.3.

Also experiments were carried out to test the information on scale transitions specified in IMTOP on real data. For these experiments the TOP10NL part of the database was populated with TOP10NL data. According to IMTOP, instances from the source scale were selected first. After this selection, a procedure was applied to change the selected instances (either discard them, or adapt them to the target scale) based on invariants and post conditions. In this process views were generated on the original TOP10NL data tables using the constraints. For example the constraint specifying that instances of the class 'buildings', with attribute 'type of building' equal to 'glasshouse' should be removed if their area on the target map will be smaller than 0.36 map mm2 can be easily specified (with real world coordinates) in SQL. The view definition for the objects that should be removed based on this constraint could be as follows:

```
SELECT b.ident, b.typeofbuil, b.heightclas, b.id, b.geom
FROM building_area b
WHERE b.typeofbuil::text = 'glasshouse'::text AND
area2d(b.geom) <= 900::double precision;
```

Human interaction was used to express the constraints in SQL. An example to show how the transition model can be applied on a real case, is shown in Figure 9. In Figure 9a candidate instances are identified for built-up area in TOP50NL (see Figure 6) based on the input object classes and the constraints specified in the transition model. In Figure 9b all the polygons of land use type '*built-up area*' are shown that were formed after the transition process as defined by the post conditions of the model.



a. Candidate instances

b. Transition result (Zoom-in)

Fig. 9. Snapshots of the model-based generation of built-up areas

The tests on the scale transitions were also used to validate if strictly applying currently available generalisation specifications specified in constraints, without adding any human interpretation, results in expected output. From the experiments it can be concluded that it is not always straightforward to apply the generalisation-related constraints. On the one hand the experiments showed that cartographer's interpretation is sometimes difficult to formalise. On the other hand they showed that the constraints need to be enriched for machine-based interpretation. An example of a constraint that is not sufficiently formal is the constraint that specifies that roads that are dead ends, classified as 'other', having main traffic use 'mixed traffic' and with a width of '2–4 meter' should be removed, except when the instance is longer than 100 meters and leads to a building. Two things that are not straightforward in this constraint is: how to evaluate 'a dead end' and how to evaluate 'leads to a building'. Connectivity is involved in both ambiguities and could be solved by using topological primitives and expressing the constraints with operations using this topological information.

6. Conclusions

This paper presented a data model (called IMTOP) for an integrated topographical database containing rich semantics on generalisation in order to support automated generalisation. After analysing several modelling approaches in UML, a modelling approach was selected that covers both data content and scale transitions that apply to complete sets of object classes in a transparent way. Transition models are added to IMTOP to model transitions that apply to specific instances. This paper showed that OCL constraints can be used to model valid data content at separate scales as well as pre- and post conditions to trigger generalisation processes.

The results of instantiating the model to evaluate it against the model requirements show that it is possible to formalise information on the integration and generalisation with UML and OCL and to use it to generate a multi-scale database. The formalisation is meant to express generalisation information in an unambiguous way. The experiments show that more work is needed to improve the formalisation of generalisation-related information, specifically the formalisation of cartographer’s interpretation that is currently used in interactive processes.

Besides supporting the workflow of the Kadaster in the generalisation process, IMTOP may also be the basis for users of multi-scale topography. This has important added value compared to independent scales. Future research will further focus on the possibilities of using IMTOP for a Model Driven Architecture approach to support the integration of databases at several scales, as well as automated generalisation. This requires a high level of formalisation. It is also future ambition to extend IMTOP with larger scale base data (1:1k). Finally, vario-scale models and progressive transfer will be investigated, which is outside the scope op the IMTOP project, but within the research of the involved authors (Oosterom 2005).

References

Bédard Y, Larrivée S, Proulx M-J, and Nadeau M (2004) Modelling geospatial databases with plug-ins for visual languages: a pragmatic approach and the impacts of 16 years of research and experimentations on Perceptory. LNCS 3289. Springer, Berlin, pp 17–30.

Buttenfield BP, and Delotto JS (1989) Multiple representations. National Center for Geographic Information and Analysis (NCGIA). Scientific Report for the Specialist Meeting, Technical paper 89–3, 87p.

Devogele T, Trevisan J, and Raynal L (1996) Building a multi-scale database with scale transition relationships. In: International Symposium on Spatial Data Handling, pp 337–351.

Enterprise Architect (2008) http://www.sparxsystems.com.au/, accessed on 22–Jan–2008.

Friis-Christensen A and Jensen CS (2003) Object-relational management of multiply represented geographic entities. In: Proceedings of the Fifteenth International Conference on Scientific and Statistical Database Management. Cambridge, MA, USA, July 9–11, pp 183–192.

Hespanha J, van Bennekom-Minnema J, van Oosterom PJM, and Lemmen C (2008) The MDA approach applied to the Land Administration Domain Model with focus on constrains specified OCL. Proceedings of FIG Working Week, 14–19 June, Stockholm, Sweden

Jones CB, Kidner DB, Luo LQ, Bundy GL, Ware JM (1996) Database design for a multi-scale spatial information system. In: IJGIS, vol 10, 8, pp 901–920.

Kadaster (2002) Specificaties TOPxxvector (Specifications TOPxxvector). Topografische Dienst, Emmen, The Netherlands.

Kadaster (2006) Generalisatievoorschriften TOP50vector (Generalisation regulations TOP50vector) Topografische Dienst, Emmen, The Netherlands.

Kilpelainen T (1997) Multiple representation and generalisation of geo-databases for topographic maps. PhD thesis, Finnish Geodetic Institute.

Lemmen CHJ, and van Oosterom PJM (2006) Version 1.0 of the FIG Core Cadastral Domain Model. XXIII International FIG congress, October, Munich, 18p.

Louwsma JS, Zlatanova S, Lammeren R, and van Oosterom PJM (2006) Specifying and implementing constraints in GIS – with examples from a geo-virtual reality system. In: GeoInformatica, vol 10, 4, pp 531–550.

Molenaar M (1989) Single valued vector maps: a concept in Geographic Information Systems. Geo-Informationssysteme, vol 2, 1, pp 18–26.

NCGIA, 1989, The research plan of the National Center for Geographic Information and Analysis. In: Int. J. Geographical Information Systems, vol 3, 2, pp 117–136.

Oosterom PJM van (2005) Variable-scale topological data structures suitable for progres-sive data transfer: the GAPface tree and GAP-edge forest. In: Cartography and Geographic Information Science, vol 32, 4, pp 331–346.

Oosterom PJM van (2006) Constraints in spatial data models, in a dynamic context. In: Drummond J, Billen R, João E, and Forrest D (eds). Dynamic and Mobile GIS: Investigating Changes in Space and Time, pp 104–137.

Oosterom PJM van, de Vries M, and Meijers M (2006) Vario-scale data server in a web service context. Workshop ICA Commission on Map Generalisation and Multiple Representation, June, Vancouver, 14p.

Parent C, Spaccapietra S, and Zimányi E (2006) Conceptual modelling for traditional and spatio-temporal applications. The MADS approach. ISBN: 3–540–30153–4.

Sparks G (2001) Database modelling in UML. Sparx Systems whitepaper.

Stoter JE, Quak CW, van Oosterom PJM, Meijers BM, Lemmens RLG, and Uitermark HT (2007) Considerations for the design of a semantic data model for a multi-representation topographical database. In: H. Kremers (ed), Lecture notes in information sciences. Berlin: CODATA, pp 53–71.

West-Nielsen P, and Meyer M (2007) Automated generalisation in a map production environment – the KMS experience. In: Mackaness WA, Ruas A, and Sarjakoski LT (eds). Generalisation of geographic information: cartographic modelling and applications. Amsterdam: Elsevier, pp 301–313.

An Interoperable Web Service Architecture to Provide Base Maps Empowered by Automated Generalisation

Theodor Foerster, Jantien Stoter, Rob Lemmens

International Institute for Geoinformation and Earth Observation, P.O.Box 6, 7500AA Enschede, the Netherlands
email: {foerster, stoter, lemmens}@itc.nl

Abstract

Producing customized base maps generated by automated generalisation on the web is an important issue in physical planning. In this web context an interoperable architecture is a key requirement. It integrates the necessary data and the functionality to finally perform the generation of the base map. Additionally, interoperability increases the reuse of the architecture for other domains. In this paper we will describe such an architecture. It has two key features: it supports the user profiles to specify the generalisation constraints and the generalisation-enabled WMS, which generates the base map according to the user profiles. The specialized WMS is especially able to access Web Service-based generalisation functionality. For the implementation of the architecture we used Geoserver, 1Spatial's Clarity and 52° North WPS.

Keywords: Automated generalisation, Web Service architecture, user profiles, physical planning

Introduction

Automated generalisation is a means to extract information by transforming data regarding scale and use. Most of the applications involving generalisation aim at map production. These maps are nowadays frequently provided on the web and are accessible through Spatial Data Infrastructures (SDIs) (McLaughin & Groot 2000). Mostly maps are static and serve only a very general purpose, but the demand for highly customized maps is increasing. Data providers are forced to adapt their products to this changing market setting. In the context of producing and disseminating customized maps based on existing topographical databases, an interoperable Web Service architecture is an important prerequisite.

Research on automated generalisation in the context of Web Services resulted in different approaches (Sarjakoski et al. 2005; Edwardes et al. 2005; Harrower & Bloch 2006). One promising approach is to perform generalisation processes on the web using a Web Service architecture (Burghardt et al. 2005; Foerster & Stoter 2006). Applying this approach to produce and disseminate customized maps on the web is still missing. This paper presents an interoperable Web Service architecture producing customized maps by means of generalisation. Customization in this case is achieved by so called user profiles describing generalisation-related user-information, which drive the generalisation process.

This paper is motivated by the RO-online[1] portal (Ministry of Housing, Spatial Planning and the Environment 2008). This portal will be launched in July 2008 and aims at a web-based dissemination approach for physical planning maps of the Netherlands. It thereby meets the evolving requirements of e-government, public participation and cost-reduction. The physical planning maps that will be delivered through RO-online consist of the plan data and a currently available complete topographic map as base information. This base information is intended to be customized individually for each user group to provide adjusted base maps for better usable physical planning maps. The starting point for this paper is to develop an interoperable architecture, which generates and delivers such customized base maps for RO-online. Interoperability is a key aspect in this context, to be able to incorporate existing architectures such as RO-online, to be able to integrate generalisation functionality provided by other Web Services, and to be flexible for future requirements.

This project is of high interest for topographic data providers, as it touches the main issue of generalisation of topographic data to derive customized maps. It is important to note, that this paper will not solve core generalisation problems, but it will show how generalisation can be applied in an interoperable Web Service architecture and how it might enable web-based dissemination of customized maps. Interoperability is an important aspect for sustainability of investments as it ensures the extensibility

[1] RO stands for Ruimte Ordering (Dutch); in English: physical planning.

and the reuse of architecture components. From a technical perspective, interoperability is achieved through data and service standards (e.g. GML, WMS, WFS and WPS). Those standards are the building blocks of the architecture presented in this paper.

The main component of the proposed architecture is the generalisation-enabled Web Map Service (WMS[2]). It consumes the user profiles and provides the maps according to these user profiles. Therefore the WMS is extended with generalisation capabilities. In our implementation the generalisation-enabled WMS is empowered by an agent-based generalisation system (i.e. 1Spatial's Clarity, Hardy et al. (2003)), which is a widely adopted implementation of the constraint-based generalisation approach. Additionally Clarity is extended to access distributed services providing generalisation functionality (i.e. Web Generalisation Services) and thereby to overcome its lack of functionality.

The paper will first introduce the project in detail and will explain the benefits of generalisation for the project. In Section 3 the paper will examine recent work in the context of generalisation and Web Service architectures. The proposed architecture will be described in Section 4. Section 5 explains some details about the already implemented components. Finally Section 6 will conclude and discuss the major aspects of the proposed architecture and will give some outlook for future work.

Physical Planning Maps on the Web

The currently ongoing shift towards e-government is relevant in the context of physical planning. The RO-online project in the Netherlands focuses on a web-based dissemination of physical plans and is the successor project of the DURP[3] project, which developed an information model for exchanging digital spatial plan information (IMRO[4]) over the web (Ottens 2004).

The RO-online project develops a web portal to provide instant web-based access to physical planning maps, compiled of the evolving digital plan data plus available topographic data. The portal is based on an

[2] More information on the WMS interface specification: www.opengeospatial.org/standards/wms.

[3] DURP stands for Digitale Uitwisseling in Ruimtelijke Processen. In English: Digital exchange for spatial processes.

[4] IMRO stands for Informatiemodel voor de Ruimtelijke Ordening. In English: Information model for physical planning.

interoperable Web Service architecture for the dissemination of the maps (Fig. 1).

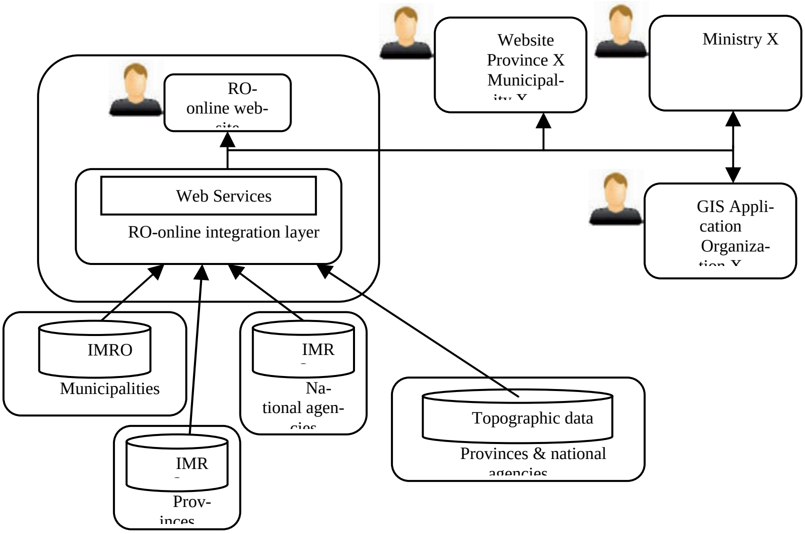


Fig. 1. The RO-online architecture incorporates Web Services (adopted from: www.helpdeskdurp.nl)

The topographic data that is currently used as base map is the complete topographical map at a certain scale (for example 1:10k or 1:50k) depending on the scale of the plan (municipal, provincial or national government level). Therefore the base map might be not optimally applicable to all user groups, as well as to new uses which are made possible by the new environment compared to the paper plans. For example, different physical plans can be combined in the portal and they can also be accessed in an interactive way (i.e. zooming). The map in Fig. 2 demonstrates that zooming might result in such a dense topographic base map, that the complete map might not be usable anymore. At least, the usability of such a map could be increased by a customized base map according to the zoom scale. Therefore the base maps have to be adapted to a specific user group but also to the way the plans are queried (integrated with other plans or not; at a specific scale).

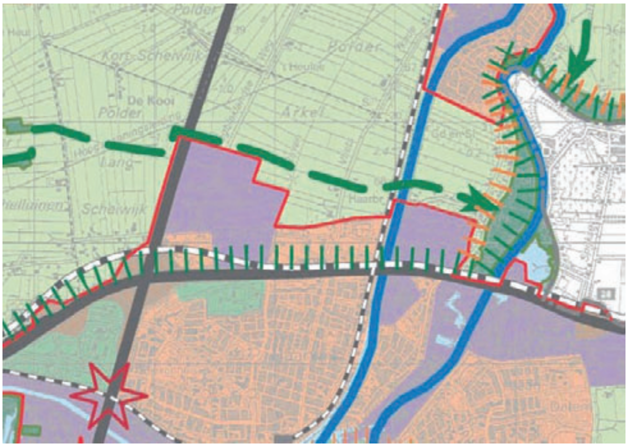


Fig. 2. Example of a physical planning map (provincial plan with 1:50K topographic base map) – in the current static approach, this base map would not have been adapted while zooming

This was the starting point of the *DURP ondergronden*[5] project, which focuses on a methodology to generate the most applicable base map for specific uses and specific user groups based on a topographic database by means of generalisation[6]. The methodology addresses usability-related and technological-related research (Poppe & Foerster 2006). This paper focuses on the latter one. The usability-related research identifies some key use and user requirements for the customized base maps which will be input for the technological-related research. The usability-related research will also yield requirements for the base maps laid down by the level of detail of the physical planning information. For example, a plan with accurate plan information can be portrayed on a detailed base map, whereas a plan that still contains fuzzy description of boundaries should not be portrayed with a highly detailed base map. These requirements are still under research. Therefore, the main requirement for the technological-related research is to develop an extensible architecture, which is able to serve customized base maps by the means of automated generalisation. The technical compliance with RO-online is a requirement for the architecture. An interoperable architecture assures that all information can be exchanged and integrated across systems. It is also important to note, that RO-online can integrate the plan information in the architecture as proposed in this paper.

[5] DURP ondergronden: in English: DURP base maps.

[6] DURP ondergronden project website: www.durpondergronden.nl.

The generalisation will be implemented based on the Large Scale Base Map of the Netherlands (*GBKN* in Dutch) (scale range 1:500 – 1:1000) and TOP10NL (topographical database at scale 1:10 000). The generated base maps will be used in combination with physical planning data at local scale (Dutch: *bestemmingsplannen*) and at provincial scale (Dutch: *structuurvisies*).

Related Literature

Model versus Cartographic Generalisation

Research on automated generalisation has yielded a lot of concepts and applications in the last 20 years. A good overview can be found in McMaster & Shea (1992) or in the most recently published book of Mackaness et al. (2007). Different views on generalisation have been developed, such as the generalisation model by Gruenreich (1992), which separates generalisation into *model generalisation* and *cartographic generalisation*. Model generalisation is concerned with the transformation of data according to a target model and cartographic generalisation is aiming at producing maps out of data by avoiding cartographic conflicts. This approach defines a clear separation between data and representation. The separation seamlessly integrates with current software systems, which also follow this approach (also known as Model-View-Controller pattern).

Cartographic generalisation is about solving conflicts of cartographic features on the map. Cartographic features do already have symbolization attached. The relation of cartographic generalisation and symbolization is depicted in Fig. 3. The presented use case in this paper aims at cartographic generalisation, as it addresses the production and dissemination of physical planning maps.

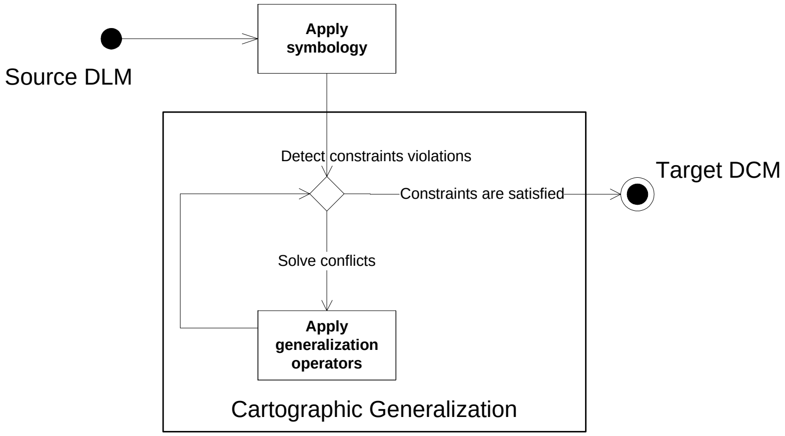

Source DLM
Apply symbology
Detect constraints violations
Constraints are satisfied
Target DCM
Solve conflicts
Apply generalization operators
Cartographic Generalization

Fig. 3. Cartographic generalisation process

Constraint-based Generalisation

Constraints for generalisation have been introduced to overcome the complexity of rule formation of generalisation. They describe conditions that should be met in the output of the generalisation process. Beard (1991) proposed this approach because the established expert systems became too complex in terms of understanding, maintenance and reuse. The idea of constraints is thereby not to formalize of how to reach a specific goal as in the case of rules, but to describe the outcome (i.e. the final map). Lots of previous approaches applied the concept of constraints to implement sophisticated generalisation systems. As constraints might compete with each other depending on the specific cartographic objects and the scale, the generalisation process itself turns into an optimization approach. An overview of the different optimization approaches applying the concept of constraints is given in Sester (2005). In this paper we will utilize the generalisation system Clarity of 1Spatial, which applies an agent-based optimization approach (Lamy et al. 1999). Until now this system has been applied in a group of National Mapping Agencies (Magnet Consortium, Lecordix 2005) and yielded a lot of promising generalisation results as presented by Regnauld (2006) and Lecordix et al. (2007). Applying Clarity in the presented architecture demonstrates also its suitability for Web Service architectures in general.

Web-based Architectures for Generalisation on the Web

Interoperability is the capability of two components to communicate at run-time to meet a common goal (ISO/TC211 2005). Such interoperability is achieved by open standards-based software development. This was the starting point for research on generalisation to propose the Web Generalisation Service (Sarjakoski et al. 2005; Edwardes et al. 2005). The Web Generalisation Service is a means to share knowledge between researchers in terms of generalisation algorithms and to extend existing software systems by remotely available generalisation functionality. In this context Edwardes et al. (2005) introduced a classification of Web Generalisation Services according to their granularity. The classification distinguishes Web Generalisation Services into two classes, the compound generalisation service and the generalisation operator service. The compound generalisation service applies a sequence of generalisation operator services to carry out a complex generalisation process such as building generalisation. The generalisation operator service implements the idea of generalisation operators, as they have been introduced in early research about automated generalisation. The concept of generalisation operators provides an abstract view on algorithms and allows grouping them according to their functionality. Different kinds of functionality have been classified in a distinct set of operators. The latest attempt, which bases its classification on formal data models of ISO and OGC is published by Foerster et al. (2007a).

Currently there are two different proposals for implementing Web Generalisation Services. Foerster & Stoter (2006) suggest to apply the evolving OGC Web Processing Service interface specification (WPS[7]) and implemented it as the 52N WPS framework[8], which is available under open source license at 52°North (Foerster & Schaeffer 2007). Burghardt et al. (2005) suggested the WebGen framework, which applies a concrete XML-RCP model to provide the generalisation functionality on the web. The WPS framework provides more functionality such as different clients. As the 52N WPS framework applies an OGC standardized approach (i.e. WPS interface specification) it provides a higher degree of interoperability in the future than the WebGen framework. Besides the research on Web Generalisation Services, Oosterom et al. (2006) proposed a Web Feature Service

[7] More information on the WPS interface specification: www.opengeospatial.org/standards/wps.

[8] 52N WPS framework website: www.52north.org/wps.

(WFS[9]) supporting progressive transfer to overcome the problem of inefficient data transfer.

In the context of generalisation research there are several projects addressing a web-based architecture such as the GiMoDig project (Sarjakoski et al. 2005) and the WebPark project (Burghardt et al. 2004). The latter one addressed cartographic generalisation in their architecture. They based their complete architecture on already existing standards such as Web Map Service and Styled Layer Descriptor (SLD[10]). This project focused on the dissemination of mobile maps and made no attempt to serve different generalisation needs of different user groups in a flexible way. The WebPark project only took into account different symbolization options, determined by different SLD documents during the generalisation process (static rules) but did not include the concept of user profiles (i.e. user-specific generalisation constraints). Additionally the WebPark project did not apply any Web Service generalisation processing, as it is presented in this paper.

Design of the Architecture

Generating the Physical Planning Map

Before introducing the conceptual overview of the architecture and its components, this subsection addresses the logical issues related to the generation of the physical planning map inside the architecture. This physical planning map is a combination of the requested planning information and the customized base map (finally available through RO-online). It will be a result of generalised topographic information meeting requirements as specified in the user profile, including requirements for symbolization of topographic objects. As already mentioned, the topographic symbolization drives the selection of the topographic feature types and their initialization as cartographic features (Section 3, Fig. 3). The symbolization of the physical plan has been set up in a legal process, in which all the Dutch planning authorities participated and is thereby predefined for the resulting map.

[9] More information on the WFS interface specification: www.opengeospatial.org/standards/wfs.

[10] More information on the SLD specification: www.opengeospatial.org/standards/sld.

To illustrate how the user profile dictates the generalisation of the base map, we define two constraints which could be part of a user profile. It is important to note that those constraints are only exemplary, thus in a user profile for a real application there might be also more constraints defined:

a) topographic objects should never cross the boundary of the overlaying planning object and
b) the building objects should never exceed a certain minimum size on the map.

Those two constraints already have a huge impact on the generalisation of the map. Depending on the map situation (i.e. density of cartographic features and map scale) any generalisation action such as aggregation, enlargement or simplification could harm one of the two constraints. An example of such a map situation is schematized in Fig. 4. Although the example is simple and the solution might not meet cartographic criteria, it demonstrates how constraints encoded as user profiles can drive the generalisation process individually in the proposed interoperable architecture. As mentioned before, the constraints are based on usability-related research, which captures the information defining needs, interests and preferences of the user towards the base map. The constraints are thereby only a concept to transform the findings of the usability-related research into a representation, which can be used as an input for the generalisation system to produce a proper map. The constraints are statically defined and the client application will not be able to change the constraints dynamically. It is important to note, that further issues related to the formulation and verification of such constraints is outside the scope of this paper, but will be addressed in future research. In this context future research will also address the formalization of constraints using for instance the classification of constraints as proposed by Burghardt et al. (2007). The formalization might be expressed using the Object Constraint Language (OCL) (Warmer & Kleppe 2003).

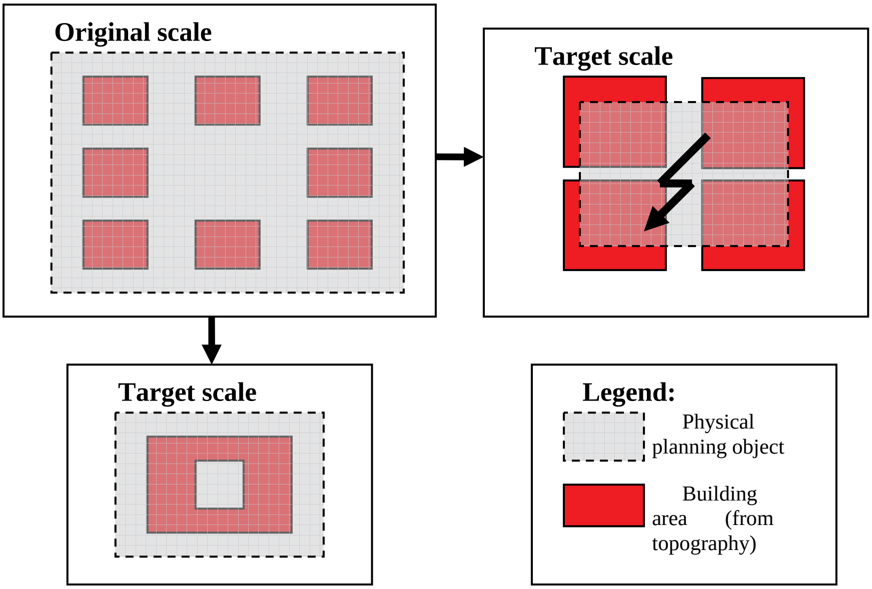


Fig. 4. Generalisation example with the topological constraint – The boundary of the planning object restricts the cartographic objects to move outside by typifying and scaling (e.g. upper left). However, aggregating is allowed (bottom left)

Components of the Architecture

The core component of the architecture is a mapping component, which produces and delivers the physical planning maps including the customized base maps. As the architecture aims at Web Services usage, this mapping component has to be able to access the different data services and has to be able to generalise the data by also accessing Web Generalisation Services. Additionally, as the produced maps have to be accessible from RO-online, the mapping component has to be a Web Service as well. For this reason our proposal is to develop a generalisation-enabled WMS. It is an extension of the WMS interface, as it has to consume the user profiles, to be aware of the generalisation requirements of the specific (remotely located) user. It is important to note, that by consuming remote user profiles, the architecture is in principle capable to serve a physical planning map regarding any set of constraints. Thereby the generalisation-enabled WMS has not to be configured for any newly introduced user group. Additionally the generalisation-enabled WMS has to handle the symbolization for the topographic data, which is specific to each user. This is achieved by SLDs, which are part of the WMS interface specification. To increase the degree of interoperability and to make use of remote services, the generalisation-enabled WMS retrieves the required data for the physical planning map (i.e. physical planning data and topographic data) from WFS.

During the generalisation process, the generalisation-enabled WMS, will access remote generalisation functionality, hosted on Web Generalisation Services (deployed as OGC WPS). The applied Web Generalisation Services in this architecture are intended to be atomic and provide functionality on the generalisation operator level (Foerster et al. 2007). Such services are called Operator Services (Edwardes et al. 2005). The term atomic refers to the functionality of the Generalisation Web Service, which is standalone and only performs generalisation upon a designated part of data and follows thereby the concept of generalisation operators. Applying atomic functionality in a complex sequence of generalisation functionality is promising as it will allow designing a robust generalisation process, in which the single atomic processes do not interfere with each other.

An overview of the introduced components is given in Fig. 5. The WFS serving the data and the client could be used from RO-online. This already shows that the generalisation-enabled WMS is finally easily applicable to the RO-online architecture or any future application.

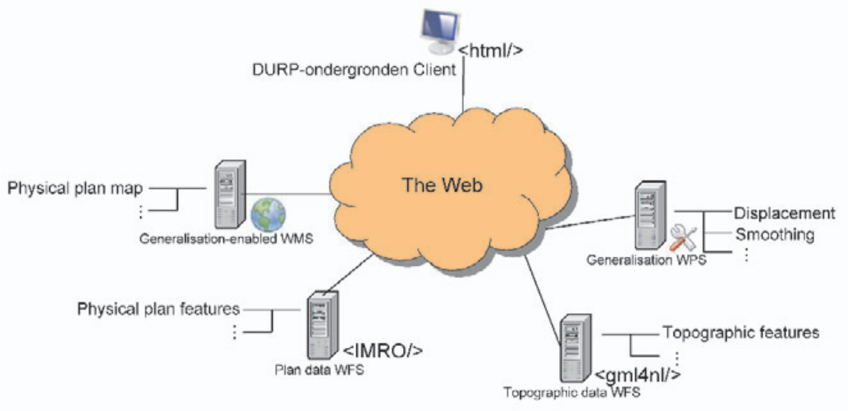


Fig. 5. Components of the interoperable architecture for producing, disseminating and displaying the physical planning map

Exemplary Architecture Walkthrough

This subsection describes the interaction between the introduced components in more detail. Those interactions are depicted in Fig. 6. We start with the client application (i.e. the DURP ondergronden client application), in which the end-user selects his/her applicable user category (e.g.

citizen, architect etc.), the location, optional zoom-level and the specific type(s) of plan information. Each user category is associated internally inside the client application with a user profile and a description of the topographic symbolization. The client application is implemented as a browser-based application, which allows easy integration with other portals such as RO-online. The client application requests a physical planning map (containing both plan information and a base map) from the generalisation-enabled WMS via the *GetMap* operation with some default parameters (map extent, map size), the SLD and the user profile.

The SLD describes the cartographic model for the topographic data encoded as XML. The user profile, which specifies the constraints (Section 4.1), will be incorporated as a vendor specific parameter into the WMS request. It is important to note, that customizing of the WMS service interface via vendor specific parameters is compliant to the WMS specification. The encoding of the user profiles might change with the outcome of the user-related research. However, for the prototype a plain XML-encoding of the constraints (as used in ESRI’s prototype Optimizer (Monnot et al. 2007)) is sufficient. It is important to note, that the generalisation-enabled WMS will also be able to render the planning map without the user profile, in which case the WMS only applies some default generalisation constraints.

Based on the requested physical planning map and the spatial extent, the WMS retrieves the physical planning features from the plan data WFS. Additionally based on the spatial extent plus the listed topographic feature types in the SLD, the generalisation-enabled WMS retrieves them accordingly from the topographic data WFS. The WFS communication is based on the *GetFeature* operation and returns GML-encoded features. The features are encoded regarding the specific schema of the physical planning data (IMRO) and the topographic data (TOP10NL or GBKN). The WMS converts the retrieved data into cartographic features by applying the specific symbolization. The symbolization for the physical plan is fixed, but the symbolization for the topographic features depends on the SLD. All these cartographic features are added to a map. For the further transformation of the map by means of generalisation, the physical planning features are immutable and will not change. During the generalisation process the topographic features will be transformed according to the constraints (as specified in the user profile) to produce the customized base map.

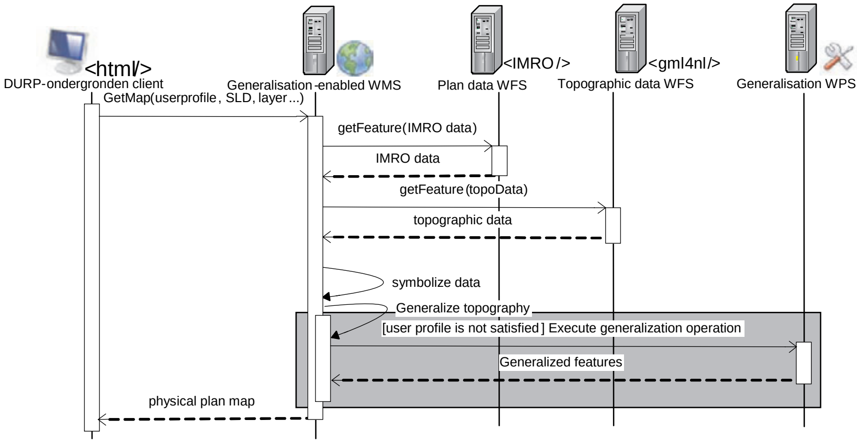


Fig. 6. Interaction within the interoperable architecture - the gray-shaded box marks the complex generalisation processing

During the generalisation process (emphasized in a gray-shaded box, Fig. 6), the generalisation-enabled WMS calls not only local generalisation algorithms, but also remote generalisation algorithms, which are hosted by Web Generalisation Services. The Web Generalisation Service is called by the *Execute* operation according to the WPS interface specification. It is important to note, that the sequence of involved algorithms is not fixed, but depends on the actual map situation (i.e. density of cartographic features and map scale) and the user profile. An example of a map situation and a user profile is given in Section 4.1. The generalisation-enabled WMS carries out the appropriate generalisation process according to the constraints described in the user profile.

Implementation of the Architecture

This section describes the on-going implementation work. As one of the project requirements is to apply open source software whenever it is feasible, Geoserver[11] was chosen as the appropriate software implementation of WMS and WFS. The reason for choosing Geoserver is manifold, at first it is written in the java programming language, at second it provides a comprehensive configuration tool for its services and at third it can be extended easily through *datastores*. The Web Generalisation Service is based on the 52° North WPS implementation (52N WPS) and the browser-based

[11] Geoserver website: www.geoserver.org.

map client is based on OpenLayers[12]. 1Spatial's Clarity is used as generalisation system for the generation of the customized base map as it implements the promising approach of agent-based generalisation (Section 3.1).

The implementation is still on-going. Here we present the parts that were finished as well as the experiences. Geoserver and the 52N WPS are already installed in our test environment as well as the data. In a first attempt, all data are inserted directly in the Clarity database, as we want to focus on the implementation of the generalisation-enabled WMS, which is the core component of the architecture and the main challenge of this project. In a final architecture setup, the data services will be provided by RO-online anyway.

The generalisation-enabled WMS is realized as a combined software component of Clarity and Geoserver. This is possible for three reasons: a) Clarity and Geoserver are written in the java programming language, b) Clarity can be loaded from another application through its underlying Application Programming Interface (i.e. Gothic API) directly and c) Geoserver allows embedding external software applications by the means of datastores[13]. Especially the concept of datastores provides Clarity all specific request parameters of the client application (user profile, SLD) to perform the generalisation individually.

The implementation architecture of the generalisation-enabled WMS and its *ClarityDatastore* is depicted in Fig. 7. The following course of action is performed by the generalisation-enabled WMS, whenever it reveives a *GetMap* request specifying user profile (constraints), SLD (symbolization), location and scale: At first Geoserver hands the request over to the ClarityDatastore. The ClarityDatastore configures the Gothic database through the Gothic API by applying the symbolization (as described in the SLD) and passing the constraints (as described in the user profile) to it. The Gothic API performs the generalisation on top of the Gothic database by calling the local and remote generalisation functionality (e.g. Web Generalisation Service implemented as OGC WPS). In a first attempt the knowledge of remotely available generalisation functionality is incorporated by the concept of actions (e.g. generalisation algorithms) in the Gothic API. By using the concept of actions it is totally opaque to the Gothic API if it calls a local algorithm or a remote functionality hosted on a Web Generalisation Service. It calls the generalisation functionality depending on the map situation and on the user profile. Nevertheless, Clarity carries out the plan finding, the execution of the plans and the evaluation

[12] OpenLayers website: www.openlayers.org.

[13] The concept of datastores is inherited from geotools, as it builds some parts of the backbone of Geoserver.

of the received generalisation results internally. Afterwards all the generalised data will be handed back to the Geoserver application. This forwards the result as a map back to the requesting client application.

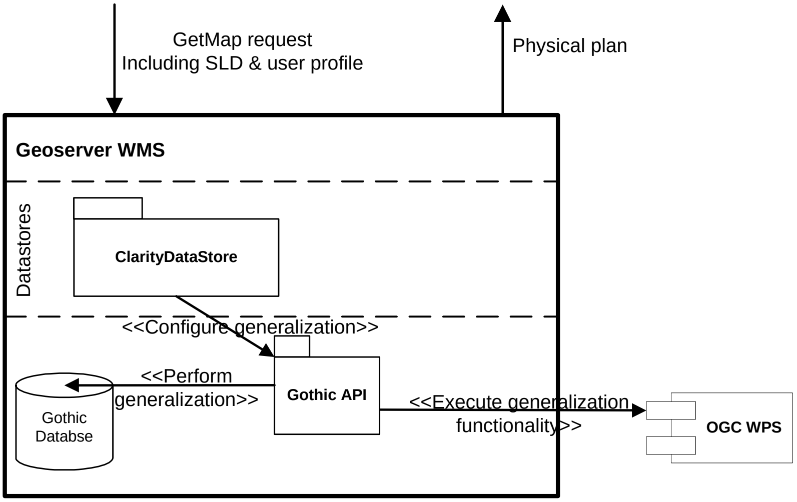

GetMap request
Including SLD & user profile
Physical plan
Geoserver WMS
Datastores
ClarityDataStore
<<Configure generalization>>
<<Perform generalization>>
Gothic Databse
Gothic API
<<Execute generalization functionality>>
OGC WPS

Fig. 7: Implementation architecture of the generalisation-enabled WMS (OGC WPS provides specific generalisation functionality such as displacement, simplification)

Regarding the generalisation-enabled WMS, the most important part of this software architecture is the coupling of Geoserver with Clarity (i.e. the ClarityDatastore). We will briefly introduce the design of the ClarityDatastore (Fig. 8). The coupling of Clarity (i.e. Gothic API) is realized by implementing three interfaces. The *ClarityDatastoreFactory* implements the *DatastoreFactorySpi* and has to maintain the connection to the Gothic database which is represented by the *ClarityDatastore* class. This is representing the *AbstractDatastore* and provides access to the actual data. It has thereby to configure the Gothic API with the specific symbolization and the user profile and to execute the generalisation. The result of the generalisation will be provided by the *ClarityFeatureReader* which allows Geoserver to access the generalised features finally by the *FeatureReader*.

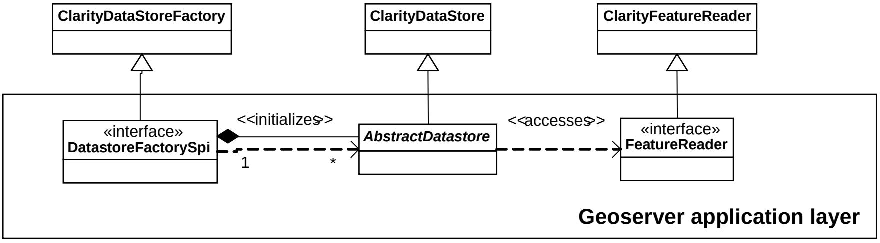


Fig. 8: Class diagram demonstrating the link between the geoserver implementation and Clarity

At the current stage, the design of the generalisation-enabled WMS is finished. It already shows that the proposed architecture is effective. In a first step, we will develop the ClarityDatastore as described in this section (Fig. 8). In a second step we will enable the generalisation-enabled WMS to access Web Generalisation Services. This will be done statically at first by preparing actions within Clarity, which connect specific remote generalisation algorithms through their endpoint URL. However, Clarity already accesses those remote algorithms automatically through the concept of actions. Finally, Clarity should be able to automatically search for different generalisation functionality on the web and incorporate it on-the-fly. This automatic approach also addresses the need for a semantically enriched description. Such a description is required to ensure meaningful interaction with the generalisation functionality and finally to produce usable generalisation results. A first attempt to solve this problem is to combine the proposed classification of operators (Foerster et al. 2007a) with the approach of application profiles for processes. In the future more sophisticated approaches should be applied such as developed in Lemmens (2006) and proposed by Regnauld (2007).

In the final prototype setting, we will incorporate the generalisation-enabled WMS into RO-online by accessing their data services and linking the developed WMS client application to the RO-online portal.

Outlook & Conclusion

This paper is motivated by the RO-online project, which provides physical planning maps via the Web. However, RO-online does not address the aspect of customized base maps, which increase the usability of physical plans, especially now they are no longer only available as single plans on paper, mapped at a specific scale. This paper addresses the issue of generating customized base maps by proposing an interoperable Web

Service architecture, which supports the production and dissemination of such customized base maps. The aspect of interoperability enables a high degree of portability and reuse of the architecture and thereby allows the adoption of customized base maps for other applications. The customization of the base maps is achieved by the means of automated generalisation.

The customization of the base map is driven by user profiles, which specify the constraints for the generalisation process individually according to the specific user. In order to generate maps according to such user profiles, the core of the architecture is the generalisation-enabled WMS. It accesses the remote data services and performs the generalisation process according to the posted user profile (Fig. 5; Fig. 6). The generalisation-enabled WMS requests remote generalisation functionality if required, which is hosted on Web Generalisation Services.

The designed prototype is deployed to a large extent on open source software. The generalisation-enabled WMS is realized by an internal coupling (i.e. datastore) between the Geoserver WMS and the generalisation system Clarity (Fig. 7). Overall Geoserver has demonstrated a clear design by the means of datastores to integrate existing applications for mapping purposes. The Web Generalisation Services will be incorporated by the means of actions into Clarity and allow Clarity to utilize local or even remote generalisation functionality.

Looking at the proposed architecture, there are still some issues left open to be addressed. To increase the interoperability of the architecture, the meaningful description of the Web Generalisation Services is a crucial issue. As already proposed, application profiles for processes or even describing the services by the means of semantic web technologies will enable an automated and meaningful discovery and execution of Web Service-based functionality. By now, the Web Generalisation Service instances are incorporated by manually linking their endpoint URLs into the generalisation-enabled WMS. However, the execution of the remote functionality is achieved automatically by Clarity.

Another crucial aspect of the architecture is its real-time performance. Executing the complete course of action, as it is described in the paper, might lead to bad performance of the architecture. Therefore, certain mechanisms for caching of already generalised maps might be helpful. It would allow processing the map once, but serving it multiple times. Research should analyze the most crucial parts of the architecture in terms of performance and propose feasible solutions. Such solutions could range from the tiling of maps to caching of generalisation results at the Web Generalisation Service. Also approaches of progressive transfer, as it is demonstrated for Web Services already by (Oosterom et al. 2006), might improve the performance of the architecture.

In the near future we will focus on further implementation of the proposed architecture, as it is already described in Section 5.

Acknowledgements

The presented research is part of Theodor Foerster's PhD thesis work which is sponsored by the RGI research program[14]. We kindly acknowledge the valuable comments of the anonymous reviewers and the discussions with Prof. Peter van Oosterom from TU Delft, the Netherlands.

References

Beard, K. M. (1991): Constraints on rule formation. In: Buttenfield, B. & Mcmaster, R. *(eds.),* Map Generalization: Making Rules for Knowledge Representation, *Longman,* 121-135.

Burghardt, D.; Edwardes, A. & Mannes, J. (2004): An architecture for automatic generalisation of mobile maps. Gartner, G. *(ed.), 2nd Symposium on Location based service and telecartography.*

Burghardt, D.; Neun, M. & Weibel, R. (2005): Generalization Services on the Web – A Classification and an Initial Prototype Implementation. *Auto-Carto 2005.*

Burghardt, D.; Schmid, S. & Stoter, J. E. (2007): Investigations on cartographic constraint formalisation. *10th ICA workshop on Generalisation and Multiple Representation.*

Edwardes, A.; Burghardt, D. & Neun, M. (2005): Interoperability in Map Generalization Research. *International Symposium on Generalisation of Information 2005.*

Foerster, T. & Stoter, J. (2006): Establishing an OGC Web Processing Service for generalization processes. *9th ICA workshop on Generalisation and Multiple Representation.*

Foerster, T.; Stoter, J. & Lemmens, R. (2007): Towards automatic web-based generalization processing: a case study. *10th ICA workshop on Generalisation and Multiple Representation.*

Foerster, T.; Stoter, J. & Köbben, B. (2007a): Towards a formal classification of Generalization operators. *ICC 2007.*

[14] RGI stands for Ruimte for Geoinformatie (in English: space for geoinformation); project website: www.rgi.nl.

Gruenreich, D. (1992): ATKIS - a topographic information system as a basis for GIS and digital cartography in Germany. *From digital map series to geo-information systems, Geologisches Jahrbuch Reihe A, 207 – 216.*

Hardy, P.; Hayles, M. & Revell, P (2003): Clarity - a new environment for generalisation using agents, Java, XML and topology. *Fifth Workshop on Progress in Automated Map Generalization.*

Harrower, M. & Bloch, M. (2006): MapShaper.org: A Map Generalization Web Service. *IEEE Computer Graphics and Applications,* 22-27.

ISO/TC 211 (2005): Geographic information – Services. *International Organization for Standardization.*

Lamy, S.; Ruas, A.; Demazeu, Y.; Jackson, M.; Mackaness, W. & Weibel, R. (1999): The Application of Agents in Automated Map Generalization. *19th International Cartographic Conference.*

Lecordix, F.; Regnauld, N.; Meyer, M. & Fechir, A. (2005): Magnet Consortium. *8th ICA Workshop on Generalization and Multiple Representation.*

Lecordix, F.; Gallic, J. L.; Gondol, L. & Braun, A. (2007): Development of a new generalization flowline for topographic maps. *10th ICA workshop on Generalisation and Multiple Representation.*

Lemmens, R. L. G. (2006): Semantic interoperability of distributed geo-services. PhD Thesis, International Institute for Geo-Information Science and Earth Observation (ITC), Enschede, The Netherlands.

Mackaness, W. A.; Ruas, A. & Sarjakoski, L. (ed.) (2007): *Generalisation of geographic information: cartographic modelling and applications.* Elsevier.

McLaughlin, J. & Groot, R. (2000): Geospatial data infrastructure: concepts, cases and good practice. *Oxford University Press.*

McMaster, R. B. & Shea, K. S. (1992): Generalization in Digital Cartography. *American Association of Geographers.*

Ministry of Housing, Spatial Planning and the Environment (2008): Government launches website for spatial plans. Article url: http://international.vrom.nl/pagina.html?id=11061, visited January 2008.

Monnot, J.; Lee, D. & Hardy, P. (2007): Topological constraints, actions, and reflexes for generalization by optimization. *10th ICA Workshop on Generalisation and Multi Representation.*

van Oosterom, P.; de Vries, M. & Meijers, M. (2006): Vario-scale data server in a web service context. *Workshop of the ICA Commission on map Generalization and Multiple representation*.

Ottens, H. (2004): An Information Model for Strategic Spatial Policy Documents. *7th AGILE Conference on Geographic Information Science,* 605-611.

Poppe, E. & Foerster, T. (2006): Automated application-driven generalization of base maps for DURP. *Congresbundel 4e GIN Symposium,* 84-87.

Regnauld, N. (2006): Improving Efficiency for Developing Automatic Generalization Solutions. *ISPRS Workshop: Multiple Representation and Interoperability of Spatial Data,* 1-5.

Regnauld, N. (2007): Evolving from automating existing map production systems to producing maps on demand automatically. *10th ICA Workshop on Generalisation and Multiple Representation.*

Sarjakoski, T.; Sester, M.; Sarjakoski, L.; Harrie, L.; Hampe, M.; Lehto, L. & Koivula (2005): Web generalisation services in GiMoDig - towards a standardised service for real-time generalisation. T. Toppen, F. & Painho, M. *(ed.), AGILE 2005,* 509-18.

Sester, M. (2005): Optimization approaches for generalization and data abstraction. *International Journal Of Geographical Information Science, 19*, 871-897.

Warmer, J. & Kleppe, A. (2003): The Object Constraint Language. *Addison Wesley*, 206 pp.

Combining Three Multi-agent Based Generalisation Models: AGENT, CartACom and GAEL

Cécile Duchêne, Julien Gaffuri

IGN, COGIT Laboratory, 2-4 avenue Pasteur, 94165 Saint-Mandé cedex, France.
email: {cecile.duchene,julien.gaffuri}@ign.fr

Abstract

This paper is concerned with the automated generalisation of vector geographic databases. It studies the possible synergies between three existing, complementary models of generalisation, all based on the multi-agent paradigm. These models are respectively well adapted for the generalisation of urban spaces (AGENT model), rural spaces (CartACom model) and background themes (GAEL model). In these models, the geographic objects are modelled as agents that apply generalisation algorithms to themselves, guided by cartographic constraints to satisfy. The differences between them particularly lie in their constraint modelling and their agent coordination model. Three complementary ways of combining these models are proposed: separate use on separate zones, "interlaced" sequential use on the same zone, and shared use of data internal to the models. The last one is further investigated and a partial re-engineering of the models is proposed.

Keywords: Automated generalisation, Multi-agent-systems, Generalisation models, Models combination.

1. Introduction

In this paper, we deal with automated cartographic generalisation of topographic vector databases. Cartographic generalisation aims at decreasing the level of detail of a vector database in order to make it suitable for a given display scale and a given set of symbols, while preserving the main characteristics of the data. It is often referred to as the derivation of a Digital Cartographic Model (DCM) from a Digital Landscape Model (DLM) (Meyer 1986). In the DCM, the objects have to satisfy a set of constraints that represent the specifications of the expected cartographic product (Beard 1991; Weibel and Dutton 1998). A constraint can be related to one object (building minimum size, global shape preservation), several objects (minimum distance, spatial distribution preservation), or a part of object (road coalescence, local shape preservation). Different approaches to automate generalisation handle the constraints expression in different ways. For instance, in approaches based on optimisation techniques (Sester 2000; Højholt 2000; Bader 2001), the constraints are translated into equations on the point coordinates.

The work presented in this paper relies on an approach of generalisation that is step by step, local (Brassel and Weibel 1988; McMaster and Shea 1988), and explicitly constraint driven (Beard 1991). More precisely, our work is concerned with three complementary models based on this approach, which also rely on the multi-agent paradigm. These three models are respectively dedicated to the generalisation of dense, well-structured data (AGENT model), low density, heterogeneous zones (CartACom model), and to the management of background themes during generalisation (GAEL model). The purpose of this paper is to investigate the possible synergies between the three models.

The next section of the paper presents in a comparative way the major aspects of the AGENT, CartACom and GAEL models. In section 3, three complementary scenarios for a combined use of these models are proposed, and the underlying technical requirements are identified. One of them is further investigated in section 4, where a partial re-engineering of the models is proposed. Finally, section 5 concludes and draws some perspectives for on-going work.

2. Comparative presentation of AGENT, CARTACOM and GAEL

2.1. The AGENT model

The AGENT generalisation model has first been proposed by Ruas (1998, 2000). It has then been used and enriched during the European AGENT project (Barrault et al 2001).

In the AGENT model, objects of the database to generalise are modelled as agent, i.e. autonomous entities that try to reach a goal thanks to capacities of perception, deliberation, action, and communication (Weiss 1999). Two levels of agents are considered. A *micro* agent is a single geographic object (e.g. road segment, building). A *meso* agent is composed of micro or meso agents that need to be considered together for generalisation (e.g. a group of aligned buildings, a urban block). This results in a pyramidal hierarchical structure where agents of one level are disjoints. Cartographic constraints can be defined for each agent (Figure 1). If a cartographic constraint concerns several agents it is translated into a constraint on the meso agent they are part of, thus a constraint is always internal to an agent.

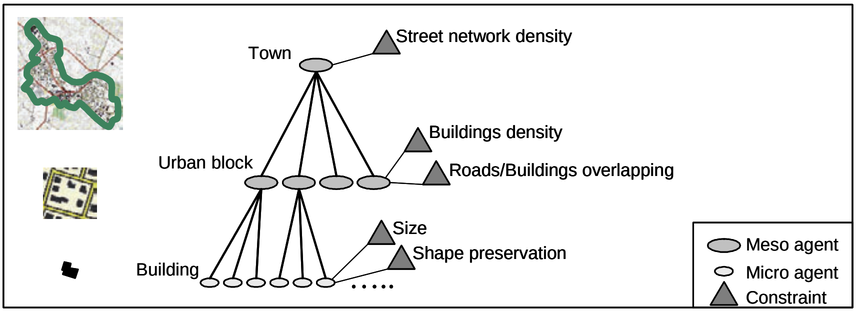


Fig. 1. The AGENT model: agents and constraints

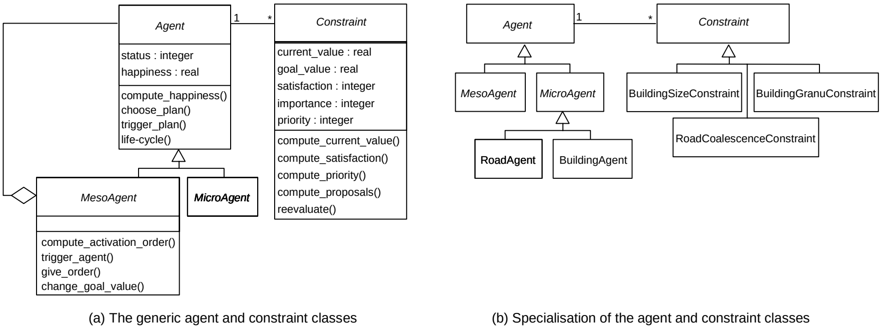
The constraints are modelled as objects. A constraint object can be thought of as an entity, part of the “brain” of an agent, in charge of managing one of its cartographic constraints. In terms of data schema (cf Figure 2a), a generic *Constraint* class is defined, linked to the generic *Agent* class. The attributes defined on the *Constraint* class are as follows:

- *current_value*: result of a measure of the constrained property (e.g. area, for the building size constraint). It is computed by the *compute_current_value* method,
- *goal_value*: what the current value should be,

- *satisfaction*: how satisfied the constraint is, i.e. how close the current value is from the goal value. It is computed by the *compute_satisfaction* method,
- *importance*: how important it is according the specifications that this constraint is satisfied, on an absolute scale shared by all the constraints,
- *priority*: how urgent it is for the agent to try and satisfy this constraint, compared to its other constraints. It is computed by the *compute_priority* method depending on the satisfaction

Two additional methods are defined:

- *compute_proposals*: computes a list a possible plans (generalisation algorithms) that might help to better satisfy the constraint, and
- *re-evaluate*: after a transformation assesses if the constraint has changed in a right way (if it has been enough improved, or at least if it has not been too much damaged)



(a) The generic agent and constraint classes

(b) Specialisation of the agent and constraint classes

Fig. 2. AGENT static model : data schema

The generic *Constraint* class is specialised into several specific constraints classes, one for every kind of cartographic constraint (cf Figure 2b). One agent is linked to one constraint object of every specific constraint class that is relevant to its geographic nature (e.g. for a building, BuildingSizeConstraint, BuildingShapeConstraint, etc.).

When a geographic agent is activated, it performs a life-cycle where it successively chooses one plan among those proposed by its constraints, tries it, validates its new state according to the constraints re-evaluation, and so on. The interactions between agents are hierarchical: a meso agent triggers its components, gives them orders or changes the goal values of their constraints (Ruas 2000).

The AGENT model has been successfully applied to the generalisation of hierarchically structured data like topographical urban data (Lecordix et al 2007) and categorical land use data (Galanda 2003).

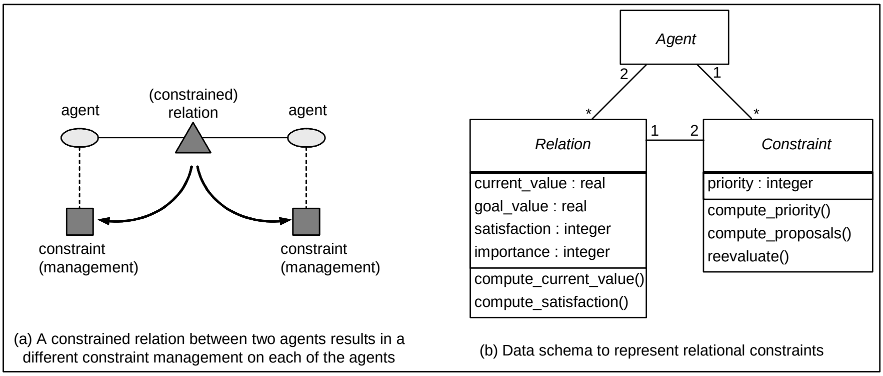
2.2. The CARTACOM model

The CARTACOM model has been proposed by Duchêne (2004) to go beyond the identified limits of the AGENT pyramidal model. It is intended for data where no obvious pyramidal organisation of the space is present, like topographical data of rural areas. In this kind of situation, it is difficult to identify pertinent disjoint groups of objects that should be generalised together, and constraints shared by two objects are difficult to express as an internal constraint of a meso object.

In CARTACOM, only the micro level of agents is considered, and agents have direct transversal interactions between each other. CARTACOM focuses on the management of constraints shared by two micro agents, that we call *relational constraints*. Examples of relational constraints are, for a building and a road, constraints of non overlapping, relative position, relative orientation.

The object representation of the constraints proposed in the AGENT model has been adapted to the relational constraints, which are shared by two agents instead of being internal to a single agent (Fig. 3). Two classes instead of one are used to represent the constraints: *Relation* and *Constraint*. The representation of a relational constraint is split into two parts:

- the first part is relative to the objective description of the state of the constrained relation, which is identical from the point of view of both agents and can thus be shared by them. This description is supported by a *Relation* object linked to both agents,
- the second part is relative to the analysis and management of the constraint, which is different for each agent and should thus be separately described for each of them. This part is described by two *Constraint* objects: one for each agent sharing the relational constraint.



(a) A constrained relation between two agents results in a different constraint management on each of the agents

(b) Data schema to represent relational constraints

Fig. 3. CARTACOM static model: agents and constraints

In order to improve the state of a relational constraint, in CARTACOM an agent can use two kinds of “plans”: either apply to itself a generalisation algorithm, like in AGENT, or ask the other agent sharing the constraint to apply an algorithm to itself.

When activated, an agent performs a life-cycle similar to the AGENT life-cycle. If AGENT internal constraints have been defined on the agent on top of its CARTACOM relational constraints, the agent can perform its internal generalisation through a call to the AGENT life-cycle, which is then seen as a black box. In the case where the agent asks another agent to perform an action, it ends its life-cycle with a “waiting” status, and resumes action at the same point when it is next activated. The agents are activated in turn by a scheduler. Sending a message to another agent places it on the top of the scheduler’s stack, i.e. the agents trigger each others by sending messages.

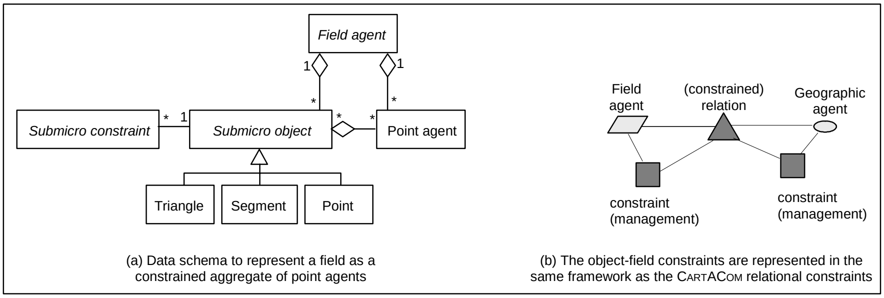
The CARTACOM model has been successfully applied to low density, rural zones of topographical data, where the density is such that few contextual elimination is necessary (Duchêne 2004).

2.3. The GAEL model

The GAEL model has been proposed by Gaffuri (2007). Its is intended for the management of the background themes like relief or land use, during an agent generalisation of “foreground” topographic themes by means of the AGENT or CARTACOM model. The background themes differ from the foreground themes in that they are continuous (defined everywhere in the space) instead of being discrete and, from a generalisation point of view, they are more flexible than the foreground themes (thus they can absorb most of the transformations of the foreground themes). Two types of cartographic constraints are considered in the GAEL model: constraints of shape preservation internal to a field theme, and constraints that aim to preserve a relation between a foreground object and a part of a background field (object-field constraint). An example of an object-field constraint is, for a river and the relief, the fact that the river has to remain in its drainage channel.

In the GAEL model, a field theme is decomposed into subparts by means of a constrained Delaunay triangulation, like in (Højholt 2000). The field’s shape preservation constraints are expressed as constraints on subparts of the triangulations called *sub-micro* objects: segments, triangles, points (Figure 4a). The object-field constraints are expressed as relational constraints between a field agent and a micro agent of the AGENT or CARTACOM model (Figure 4b, not represented in the class diagram of

Figure 4a), and translated into constraints on sub-micro objects. The points that compose the triangulation are modelled as agents. The sub-micro objects are thus groups of point agents. Each internal or object-field constraint that concerns a sub-micro object is translated into forces on the point agents that compose it.



(a) Data schema to represent a field as a constrained aggregate of point agents

(b) The object-field constraints are represented in the same framework as the CARTACOM relational constraints

Fig. 4. GAEL static model : sub-micro level, point agents, sub-micro and object-field constraints

When a point agent is activated, it computes and applies to itself a small displacement in the direction that would enable it to reach a balance between the forces resulting from its constraints. Interactions between agents can be hierarchical or transversal. Field agents can trigger their point agents (hierarchical interaction), and point agents can directly trigger their neighbours (transversal interactions). This results in a progressive deformation of the field in answer to the deformations of the foreground themes.

The GAEL model has been successfully applied (Gaffuri 2007) to the preservation of relations between buildings and relief (elevation) and hydrographic network and relief (overland flow).

2.4. Areas of applications of AGENT, CARTACOM and GAEL: schematic summary

Figure 5 summarizes the main characteristics of the AGENT, CARTACOM and GAEL models. AGENT is based on hierarchical interactions between agents that represent single geographic objects or groups of objects. The considered constraints are described as internal to a single agent and managed by this agent. This model is best suited for generalising dense areas where density and non-overlapping constraints are prevalent and strong contextual elimination is required. CARTACOM is based on transversal interactions between agents that represent single geographic objects. The considered constraints are described as shared by two agents and managed by both concerned agents. This model is best suited for generalising low density areas where more subtile relational constraints like relative orientation are manageable. GAEL is based on transversal interactions between agents that represent points of geographic objects connected by a triangulation, and hierarchical interactions between these agents and agents that represent field geographical objects. The considered constraints are described either as shared by a field agent and a micro agent, or as internal to groups of connected point agents, and handled by these point agents. This model is best suited for the management of side-effects of generalisation on the background themes.

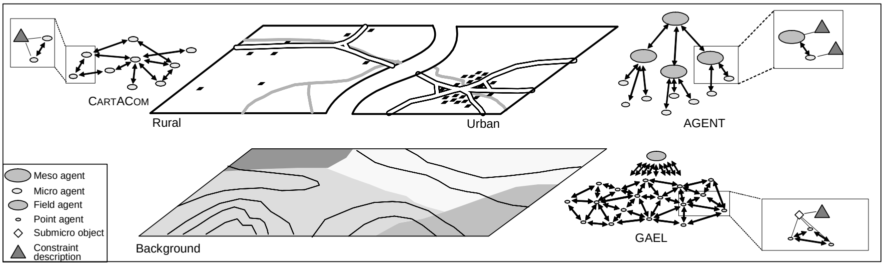


Fig. 5. AGENT, CartACom and GAEL model: target areas of application and levels at which constraints are described (GAEL object_field constraints are not represented)

The three models are best suited for different kinds of situations that are all present on any complete topographic map. Thus they will have to be used together in a complete generalisation process. In the next section, scenarios are proposed for the combined use of the three models.

3 Proposed scenarios to combine AGENT, CartACom and GAEL

In the subsections 3.1, 3.2 and 3.3, three complementary scenarios for the combined use of the models are studied, in which the synergy takes place at different levels. For each scenario, the underlying technical and research issues are identified.

3.1. Scenario 1: separate use of AGENT, GAEL and CartACom on a spatially and/or thematically partitioned dataset

This first scenario concerns the generalisation of a complete topographical dataset. Such a dataset contains both foreground and background themes (everywhere), and both rural and urban zones. In this scenario, we propose to split the space as shown in figure 5, both spatially and thematically, in order to apply each of the three models where it is *a priori* best suited:

- urban foreground partitions are generalised using AGENT,
- rural foreground partitions are generalised using CartACom,
- background partitions follow using GAEL.

Let us notice that this scenario does not cover the complete generalisation process but only a part of it. It is intended to be included in a larger generalisation process or Global Master Plan (Ruas and Plazanet 1996) that also includes steps for overall network pruning, road displacement using e.g. the beams model (Bader 2001), and generalisation of background themes (on top of letting them follow the foreground themes). Actually, these additional steps would also be applied on either thematically or spatially split portions of the space.

This scenario first requires adapted methods to partition the data in a pertinent way (here into foreground and background themes, into urban and rural zones). Then, whatever the partitioning, the resulting space portions are not independent because strong constraints exist between objects situated on each side of the borders: continuity of roads and other networks at spatial borders, inter-theme topological relations, etc. This interdependence requires mechanisms for the management of side-effects on the data, i.e. to ensure that no spatial inconsistency is created with other portions of the space when applying one model on portion of the space. It also requires pertinent heuristics for the orchestration of the process: when to apply which model on which partition.

These issues are not new: they have already been discussed by (McMaster and Shea 1988; Brassel and Weibel 1988; Ruas and Plazanet 1996) regarding the design of generalisation process composed of elementary algorithms. We are just a step forward here, since now we consider the combination of several generalisation processes based on different models.

3.2. Scenario 2: "interlaced" sequential use of AGENT, CARTACOM and GAEL on a set of objects

This second scenario concerns the generalisation of a set of objects included in a single partition of the previous scenario i.e. a portion of either urban foreground space, rural foreground space, or background space. In this scenario, we propose to enable the "interlaced sequential use" of the models, i.e. calling successively two or more of the models on the same objects, possibly several times (e.g. AGENT then CARTACOM then AGENT again).

Indeed, experiments performed with the AGENT and CARTACOM models show that the previous scenario is not sufficient. The limit between a rural space that should a priori be generalised by CARTACOM and a urban space that should a priori be generalised by AGENT is not so clear. In some zones of medium density, CARTACOM enables to solve most of the conflicts while tackling also more subtile constraints like relative orientation, but can locally encounter over-constrained situations. In this second scenario, such locally over-constrained situations can be solved by a dynamic call to an AGENT hierarchical resolution. Conversely, not all the constraints shared by two objects in an urban zone can easily be expressed as an internal constraint of a group (meso agent) and solved at the group level. Thus, in scenario 1, some of them are given up, e.g. the constraint of relative orientation. Scenario 2 enables punctual use of CARTACOM inside a urban zone, which could help in solving such subtile relational constraints for which no group treatment is available. Regarding the thematic split between foreground and background, it seems that this distinction is not so well defined either. This is why in this scenario, some objects can be considered as foreground at one time of the process and background at other times. For instance, buildings are foreground when handling there relational constraints with the roads thanks to a CARTACOM activation; but they are rather background when handling the overlapping constraints between roads, as their behaviour at this time should just be to follow the other feature classes in order to prevent topological inconsistencies.

To summarize, in scenario 2 the geographic objects of a dataset are able to play several roles during a generalisation process: an object can be modelled as an AGENT, CARTACOM and GAEL agent at the same time and be successively triggered with an AGENT, CARTACOM or GAEL behaviour (life-cycle). To be more precise, a same object of the micro level can be modelled and triggered both as an AGENT and CARTACOM agent, and the points that compose it modelled and triggered as GAEL agents (as the GAEL model operates at the points level).

To enable this, some mechanisms are required to detect the need to dynamically switch to another model. This means, a mechanism is needed detect that the currently used model is unable to solve the situation, and identify the pertinent set of objects that should temporarily be activated with another model. Then, some consistency preservation mechanisms are required, not from a spatial point of view (this has already been tackled in scenario 1), but regarding the data in which an agent stores its representation of the world. For instance, if a CARTACOM activation is interrupted and an AGENT activation is performed that eliminates some agents, the neighbours of the eliminated agents should be warned when the CARTACOM activation resumes, so that they can update their "mental state". Otherwise, they could enter in an inconsistent state, with pending conversations and relational constraints with agents that do no longer exist.

3.3. Scenario 3: simultaneous use of AGENT and CARTACOM data on one object

This third scenario concerns the generalisation decisions taken by an agent of the micro level (single geographic object) that is both modelled as an AGENT and as a CARTACOM agent as proposed in scenario 2. Only the AGENT and CARTACOM models are considered here since only these models operate at a common level (micro level).

An agent that is both modelled as an AGENT and as a CARTACOM agent has knowledge both of its internal constraints and of relational constraints shared with other agents. But so far, including in scenario 2 above, only the internal constraints are taken into account when it behaves as an AGENT agent, and only the relational constraints are taken into account when it behaves as a CARTACOM agent (during its CARTACOM life-cycle, it can perform internal generalisation thanks to a call to the AGENT life-cycle as explained in 2.2, but the AGENT life-cycle is then seen as a "black box"). In this third scenario, an agent is able to consider both kinds of constraints at the same time when making a generalisation decision, be it in a CARTACOM or in an AGENT activation. This means that, when choosing the next action to try, the agent takes into account both the proposals made by its internal and relational constraints (with the restriction that an agent activated by AGENT does not try an action consisting in asking another agent to do something). And, to validate the action it has just tried, the agent takes into account the satisfaction improvement of both its internal and relational constraints. This scenario is not intended to introduce more relational constraints in urban zones than in scenario 2. It just proposes that, when such constraints have been defined (like the relative orientation constraint), they can be taken into account at the same time as the internal constraints. Provided relational constraints are parsimoniously added, and the relative importances and the relaxation rules of the internal and relational constraints are well defined, this scenario should not result in over-constrained situations anywhere. And it has multiple advantages:

- The aim of an agent activated by CARTACOM (e.g. a rural building) is still to satisfy both internal and relational constraints, but it can satisfy all of them by trying the actions they suggest, while remaining in its CARTACOM life-cycle. This is less computationally heavy than calling the AGENT life-cycle as a "black box".
- The aim of an agent activated by AGENT (e.g. a urban building) is still first to satisfy as well as possible its internal constraints. But, if it has relational constraints defined, they can prevent it from applying an internal algorithm that would decrease their satisfaction too much. For example, algorithms that square the angles of a building, or that transform it into a rectangle, can easily break relations of local parallelism between the building (or one of its walls) and another building or a road. (Steiniger, 2007, p. II-C-13) proposes to prevent this by forbidding the use of these algorithms in the parts of urban space classified as "inner city", because this problem frequently occurs in this kind of area. This scenario 3 enables to avoid this kind of problem in a more adaptive way (only when it really occurs).
- An agent activated by AGENT can also try internal actions specifically in order to improve the satisfaction of one of its relational constraints (like a small rotation in order to achieve parallelism with a neighbouring road). This is far less heavy than having to stop the AGENT activation and start a CARTACOM activation on the whole urban block containing the building.
- If micro-agents activated with AGENT cannot cope with some relational constraints because of "domino effects", another way of solving these constraints can also be that the meso agent above seeks for a global solution by analysing the relational constraints of its components (e.g., in the above case the urban block identifies the buildings that should rotate).

To enable this scenario 3, it is necessary to re-engineer the parts of the AGENT and CARTACOM static models related to constraint representation so that internal (AGENT) and relational (CARTACOM) constraints can both be handled by an agent within the same methods. Hence, the methods of the "Agent" class that use the constraints have to be modified, both in the

AGENT and in the CARTACOM model, in order to take into account the presence of both internal and relational constraints.

4. How to put the proposed scenarios into practice

4.1. Technical requirements underlying scenarios 1, 2 and 3: summary

In sections 3.1, 3.2 and 3.3 we have presented three scenarios where the AGENT, CARTACOM and GAEL models are used with an increasing degree of combination: separate use on separate zones (scenario 1), "interlaced" sequential use on the same zone (scenario 2), shared use of data internal to the models (scenario 3). The three scenarios are complementary and we intend to put all the three into practice in a medium term. The identified underlying issues are summarized hereafter, starting from the most external elements of the models, to the most internal:

1. Define methods to split the map space into relevant partitions, on spatial and/or thematic criteria [scenario 1]
2. Define a strategy to know which model to apply when on which portion of space [scenario 1]
3. Define mechanisms to manage the side-effects at borders, when generalising one partition with one model [scenario 1]
4. Define mechanisms to dynamically identify a set of geographical objects that require a temporary activation of another model than the currently active one [scenario 2]
5. Define mechanisms to preserve the consistency of data internal to one model, when another model is running [scenario 2]
6. Re-engineer the representation and management of constraints in AGENT and CARTACOM so that internal and relational constraints can be handled together [scenario 3]

The current status of the issues (1) to (5) is briefly described in the next section. The issue (6) is tackled more in deep in section 4.3.

4.2. Status of the technical issues underlying scenarios 1 and 2

The issues underlying the scenarios 1 and 2 (issues 1 to 5) in the list above) are part of a research that is currently beginning. However, for some of them we already have some elements of answer. Regarding the space partitioning (issue 1), previous research like (Boffet 2000; Chaudhry 2007) provide specific methods to identify urban or mountainous areas. Regarding the management of side-effects at thematic borders (issue 3), the *Object-field constraints* have been defined in the GAEL model in order to manage, thanks to a GAEL activation, the side-effects induced on the background themes by the AGENT or CARTACOM activations performed on foreground themes. This has already been implemented and tested for the themes building-relief and hydrography-relief during an AGENT activation (Gaffuri 2007). However, the question of when optimally to trigger GAEL during the AGENT activation (issue 2) is not solved yet. Regarding the interlaced use of two models, (Duchêne 2004) tackles the automatic triggering of group operations during a CARTACOM activation (e.g. with an AGENT meso activation). Consistency preservation mechanisms (issue 5) have been implemented and tested with manually triggered group operations. To detect that a group operation is needed (issue 4), a model has been proposed but not implemented at this time.

4.3. Re-engineering of constraint modelling in AGENT and CARTACOM to support scenario 3

In this section, we focus on the re-engineering of the AGENT and CARTACOM constraint modelling in order to enable that an agent modelled both as AGENT and CARTACOM agent can handle its internal and relational constraints at the same time. This means that in any method of an AGENT or CARTACOM agent that handles constraints, the role of constraint can be played either by an internal or a relational constraint. In other words, the agent has to see its internal and relational constraints within the same framework. In AGENT, an internal constraint is modelled as an entity in charge of both the description and the management of the constraint (Figure 6a). In CARTACOM, because the descriptive part is shared by two agents, two different entities are used for the description and the management of a relational constraint (Figure 6b). To integrate the two representations in a common framework, we first propose to modify the relational constraint modelling in CARTACOM: the descriptive part of the constraint (*Relation* object) is “replicated” on the two linked *Constraint* objects so that a CARTACOM agent “sees” the same thing as an AGENT agent (Figure 6c).

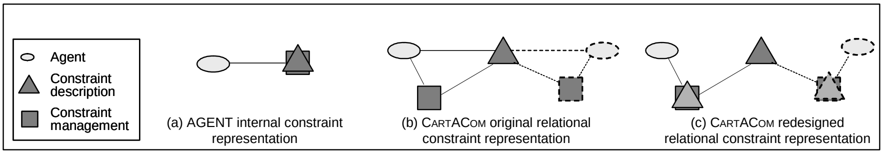
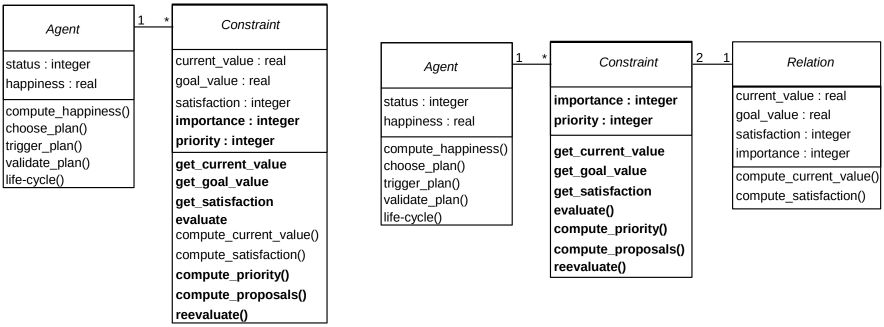

Agent
Constraint description
Constraint management
(a) AGENT internal constraint representation
(b) CartACom original relational constraint representation
(c) CartACom redesigned relational constraint representation

Fig. 6. How to ensure that an agent "sees" its internal and relational constraints in the same framework

More precisely, there is no data replication, but all the information supported by the *Relation* object is made available from the linked *Constraint* objects. The resulting data schema is shown in Figure 7b: getter methods have been added to the CartACom *Constraint* class, which get the values carried by the attributes of the CartACom *Relation* class. The AGENT data schema is modified accordingly (Figure 7a: the same getter methods are added, but they get the values from the attributes of the AGENT *Constraint* class). Attributes and methods in bold are the ones that the agent can use because they are common to internal (AGENT) and relational (CartACom) constraints.



(a) AGENT redesigned constraint modelling

(b) CartACom redesigned constraint modelling

Fig. 7. Formatting the AGENT and CartACom constraint representation in the same framework results in splitting the AGENT Constraint class.

Once this "replication" has been performed, we can merge the AGENT and CartACom data schema (Figure 8). A generic *Agent* class is specialised into *AGENTAgent* and *CartAComAgent*. Similarily, a generic *Constraint* class is specialised into *InternalConstraint* and *RelationalConstraint*. The attributes and methods common to the AGENT and CartACom classes are transferred to the generic classes.

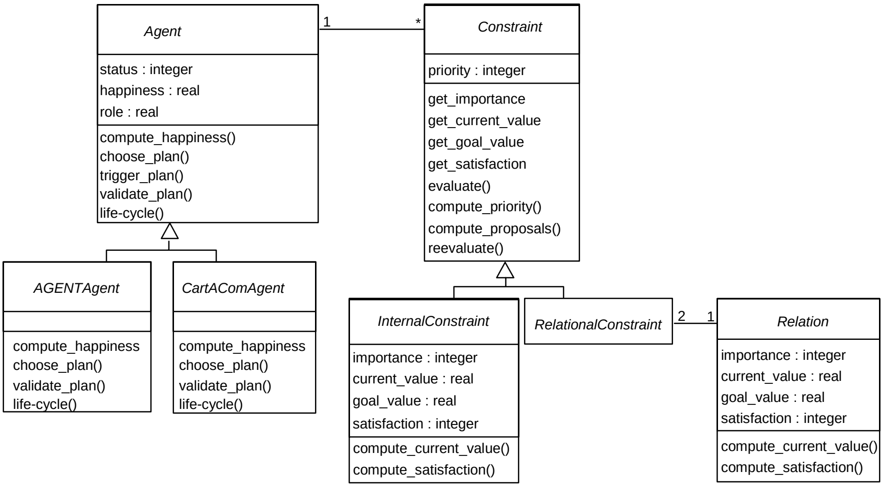


Fig. 8. Factorisation of common aspects of the agents and constraints: re-engineered constraint representation of AGENT and CARTACOM.

As geographic objects can be modelled both as AGENT and CARTACOM agents (i.e. their class can inherit both from *AGENTAgent* and *CARTACOMAgent*), an attribute *role* is added to the generic Agent class. This attribute indicates wether the agent has to be activated as AGENT or CARTACOM agent, i.e. which version of the life-cycle (and the methods it uses) has to be applied to it. Apart from the method that triggers a plan, the other methods used by the life-cycle (and the life-cycle itself) are indeed different for an AGENT and a CARTACOM agent. In other words, these methods, defined at the generic *Agent* level, are specialised in the *AGENTAgent* and *CARTACOMAgent* classes. The *role* of an agent can change during the generalisation process.

The re-engineered data schema presented in figure 8 ensures that an agent modelled both as AGENT and CARTACOM agent can handle its internal and relational constraints at the same time. This was indeed the aim of this re-engineering. But an additional effect of merging the *Agent* and *Constraint* classes, while factorizing the properties and methods common to the models at the most generic level, is to allow an easier maintenance of the system. To go further in this direction, we propose to include the classes of the GAEL data schema (cf. Figure 4, section 2.3) in the merged schema. This is quite straightforward. Regarding constraints modelling, the GAEL model already uses the AGENT modelling for internal constraints associated to point agents and to submicro objects, and the CARTACOM modelling for relational constraints associated with field agents. We just have to add the newly defined *Field-object Relation* class, as a subclass of the CARTACOM *Relation* class. Regarding agents modelling, the GAEL *Field agent* and *Point agent* classes already have the same attributes and methods as the generic *Agent* class of the merged schema. Thus the GAEL agent classes are added as new subclasses of the *Agent* generic class. The final merged schema is shown on Figure 9.

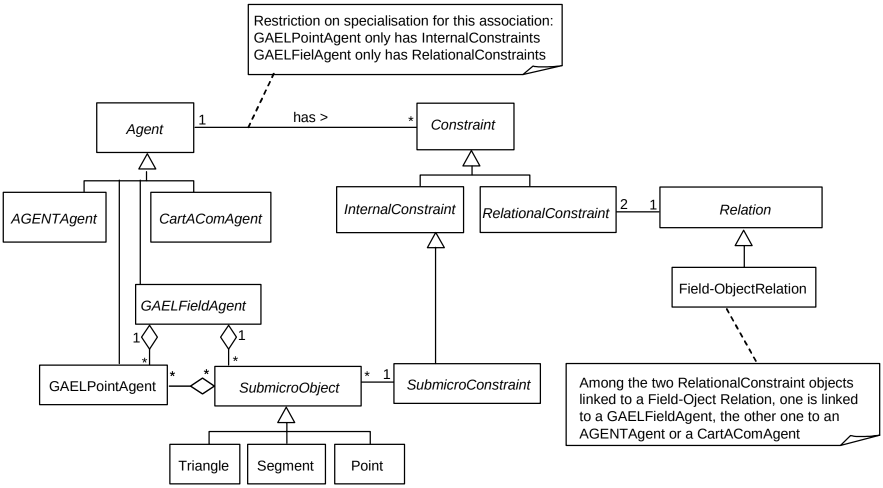

Restriction on specialisation for this association:
GAELPointAgent only has InternalConstraints
GAELFielAgent only has RelationalConstraints
Agent
1
has >
*
Constraint
AGENTAgent
CartAComAgent
InternalConstraint
RelationalConstraint
2
1
Relation
Field-ObjectRelation
GAELFieldAgent
1
1
*
*
GAELPointAgent
*
*
SubmicroObject
*
1
SubmicroConstraint
Among the two RelationalConstraint objects linked to a Field-Oject Relation, one is linked to a GAELFieldAgent, the other one to an AGENTAgent or a CartAComAgent
Triangle
Segment
Point

Fig. 9. Introduction of the GAEL classes to the merged schema

5. Discussion

In the two previous sections, we proposed three scenarios to combine the AGENT, CARTACOM and GAEL generalisation models in order to take advantage of each of them. Among these scenarios, at least the first one can also be extended to other generalisation models. We think that such scenarios are needed to go further in the automation of generalisation, without relaxing the cartographic quality too much.

However, complexity or tractability problems are necessarily attached to a system that would implement these three scenarios. This complexity takes place at two levels: firstly, getting familiar to the system and tuning it for a particular use is fastidious; and secondly, the automated generalisation process itself is of high computational complexity (high numbers of agents, constraints and links between both, time consuming algorithms, etc.).

This double complexity calls some clarifications on the target usages of such a system. If we consider the map series making (for map producers, namely NMAs), this double complexity is not a huge problem: building a new map production line is anyway time and resource consuming, and the computational constraints attached to the actual production of one map are not very strong as long as no memory overflow is encountered. Indeed, if the process is highly automated, it can run on a dedicated machine overnight, and will anyway be far quicker than manual generalisation. Now, if we consider on demand mapping, the computational complexity clearly prevents from using such a system for on the fly generalisation. But it could still be used for off-line customised cartography, provided the tuning (parametrization) can be assisted. Research works that can help in this are (Hubert and Ruas 2003), for the translation of user needs into a generalisation system, and (Taillandier 2007), to help the automated revision of the procedural knowledge within the AGENT model.

6. Conclusion and perspectives

In this paper, we have presented a comparative analysis of three agent-based generalisation models dedicated to three different kinds of geographic data and cartographic constraints: AGENT, CartACom and GAEL. Three complementary scenarios have been proposed to use them in a combined way, with an increasing degree of combination. For each scenario, the underlying issues have been described. The issue that is the most internal to the models has been tackled and as a result, a partial re-engineering of the models has been proposed.

This re-engineered version will now be implemented in Clarity®, the generalisation platform commercialised by 1Spatial, where AGENT and GAEL are already implemented. CartACom, which is for the time being implemented in LAMPS2, will be ported to Clarity on that occasion. The re-engineered model will then be tested on three different topographical data extracts separately:

- an urban zone with classical internal constraints and a few relational constraints defined, that will be generalised with an AGENT activation,
- a rural zone with classical relational constraints and some internal constraints defined, that will be generalised with a CartACom activation,
- a mountainous zone with object-field and internal field constraints defined, where the generalisation of foreground themes will be interactively performed and their side-effects managed by GAEL activations.

The issues related to the first two scenarios, numbered (1) to (5) in section 4.1, are being tackled in a parallel research project. This way, we hope to make significant progress in multi-theme generalisation.

Acknowledgement

The authors would like to warmly thank the anonymous reviewer who sent his marked-up manuscript for his constructive remarks and his English language corrections.

References

Bader M (2001) Energy Minimization Methods for Feature Displacement in Map Generalization. Ph.D. thesis, University of Zürich

Barrault M, Regnauld N, Duchêne C, Haire K, Baeijs C, Demazeau Y, Hardy P, Mackaness W, Ruas A, Weibel R (2001) Integrating Multi-agent, Object-oriented, And Algorithmic Techniques For Improved Automated Map Generalization. In: Proc. of the 20th International Cartographic Conference, Beijing, China, 2001, vol.3, pp 2110-2116

Beard K (1991) Constraints on rule formation. In: Buttenfield B., McMaster R. (eds) Map Generalization: Making Rules for Knowledge Representation, Longman Scientific and Technical, Harlow, Essex, pp 32-58

Brassel K, Weibel R (1988) A review and conceptual framework of automated map generalization. International Journal of Geographic Information Systems, 1988, 2(3):229-244

Boffet A (2000) Creating urban information for cartographic generalisation. In: Proceedings of the 9th International Symposium on Spatial Data Handling (SDH 2000), Beijing, China, pp 3b4-16

Chaudhry O (2007) Automated scale dependent views of hills and ranges via morphometric analysis. In: Proceedings of the 23rd International Cartographic Conference, Moscow, Russia

Duchêne C (2004) The CARTACOM model: a generalisation model for taking relational constraints into account. 6th ICA Workshop on progress in automated map generalisation, Leicester

Gaffuri J (2007) Field deformation in an agent-based generalisation model: the GAEL model. Proceedings of GI-days 2007 - young researches forum, Münster, Germany, 2007, vol. 30, pp 1-24

Galanda M (2003) Automated Polygon Generalization in a Multi Agent System. Ph.D. thesis, University of Zürich

Højholt P (2000) Solving Space Conflicts in Map Generalization: Using a Finite Element Method. Cartography and Geographic Information Science, 27(1): 65-73

Hubert F, Ruas A (2003) A method based on samples to capture user needs for generalisation. 5th ICA Workshop on progress in automated map generalisation, Paris

Lecordix F, Le Gallic J-M, Gondol L, Braun A (2007) Development of a new generalisation flowline for topographic maps. 10th ICA Workshop on Generalisation and Multiple Representation, Moscow, Russie

McMaster R, Shea K (1988) Cartographic Generalization in a Digital Environment: a Framework for implementation in a GIS. Proceedings of GIS/LIS'88, San Antonio, Texas, USA, pp 240-249

Meyer U (1986) Software developments for computer-assisted generalization. In: Proceedings of Auto-Carto, London, 2:247-256

Ruas A, Plazanet C (1996) Strategies for Automated Generalization. Proc. of the 7th International Symposium on Spatial Data Handling, Delft, The Netherlands, pp 6.1-6.17

Ruas A (1998) OO-Constraint modelling to automate urban generalisation process. In: Proceedings of the 8th International Symposium on Spatial Data Handling, pp 225-235

Ruas A (2000) The Roles Of Meso Objects for Generalisation. Proceedings of the 9th International Symposium on Spatial Data Handling, Beijing, pp3b50-3b63

Sester M (2000) Generalization Based on Least Squares Adjustment. International Archives of Photogrammetry and Remote Sensing, vol.33

Steiniger S (2007) Enabling Pattern-Aware Automated Map Generalization. Ph.D. thesis, University of Zürich

Taillandier P (2007) Automatic Knowledge Revision of a Generalisation System. 10th ICA Workshop on Generalisation and Multiple Representation, Moscou

Weibel R, Dutton G (1998) Constraint-Based Automated Map Generalization. In: Proceedings of the 8th International Symposium on Spatial Data Handling, pp 214-224

Weiss G (1999) Multiagent Systems. A Modern Approach to Distributed Artificial Intelligence. The MIT Press

Implementation of Building Reconstruction Algorithm Using Real World LIDAR Data

Rebecca O.C. Tse, Chris Gold, Dave Kidner

University of Glamorgan, Faculty of Advanced Technology, CF37 1DL, Wales.
email: {rtse,cmgold,dbkidner}@glam.ac.uk,

Abstract

An increasing use of three dimensional point clouds for building reconstruction is being driven by the popularity of Airborne Laser Scanning (ALS). Laser scanning data provides rapid and accurate elevation models of buildings, forest and terrain surface. Though the captured data contains X, Y, and Z coordinates, the data volume is huge and does not provide any building information. The challenge is to covert the point clouds into CAD-type models containing vertical walls, roof planes and terrain which can be rapidly displayed from any 3D viewpoint.

An alternative method was developed to locate building blocks and identify the roof structures with the use of the Delaunay Triangulation and its dual Voronoi diagram and simulated data was used to illustrate the algorithm. This paper shows the implementation of the method using real world ALS data.

1 Introduction

Airborne Laser Scanning (ALS) or Light Detection and Ranging (LIDAR) is a laser-based survey for the acquisition of topographical and return signal intensity data. It is a new independent and highly automated technology to produce digital surface models (DSM) and digital terrain models

(DTM) (Ackermann, 99). ALS data provides a rapid 3D data collection over a massive area, but it does not capture any building information. However an increasing need for automated 3D building reconstruction models is due to a variety of applications, from tourism to disaster management. A challenge is to convert the high density 3D point clouds into CAD-type models containing vertical walls, roof planes and terrain which can be rapidly displayed from any 3D viewpoint.

It is common in current research to combine additional data on extracting building information or use pre-defined buildings to fit the roof structures (Brenner and Haala 1998, 1999). Many of them start by removing all buildings, trees and terrain objects from the raw ALS data and generate a bare-earth model (Vosselman 2003). With the help of additional data sources or existing cadastral data they obtain building boundaries (Sohn and Dowman 2003, 2004; Suveg and Vosselman, 2001, 2004; Vosselman and Dijkman, 2001, Rottensteiner 2001, Rottensteiner et al, 2007). Then the roof structure is created and fitted onto some pre-defined building shapes (Brenner and Haala 1998, Brunn and Weidner, 1997, Rottensteiner and Briese 2002, 2003). Finally CAD software is used to create the building and paste it on the bare-earth model. Problems occur when:

- no other data source is available for building outline extraction,
- the target building does not match any pre-defined building shape,
- and the topological connectivity between the building and the terrain surface is needed for further spatial analysis.

Our approach uses an alternative method to reconstruct buildings (Tse et al, 2007) which looks similar to (Vosselman, 1999, Forlani et al, 2003, Wang and Schenk, 2000). However we focus on using LIDAR data with no additional data sources and we use the TIN structure for the whole reconstruction process. We would like to keep the topological connectivity.

The first step is to identify building blocks by searching the vertical walls to separate the high and low land. Then we cluster the data points inside the building outline and find the roof structure. The final step is to extrude the building from the terrain surface with preserved topological connectivity. We used simulated LIDAR data to demonstrate our method, and real world LIDAR data has been used to verify the result.

2 Airborne Laser Scanning (ALS)

An ALS survey is composed of three main technologies to capture high-accuracy data points on the ground. These include: a laser scanning system, a global positioning system (GPS) and an inertial measuring unit (IMU) or inertial navigation system (INS). The laser scanning system, a

LIDAR sensor, is mounted on an aircraft, helicopter or satellites. The sensor measures the time difference between the emission of each pulse to reflect off the ground (or objects on the terrain) and its return to the sensor. An example of an ALS survey is shown in figure 1.

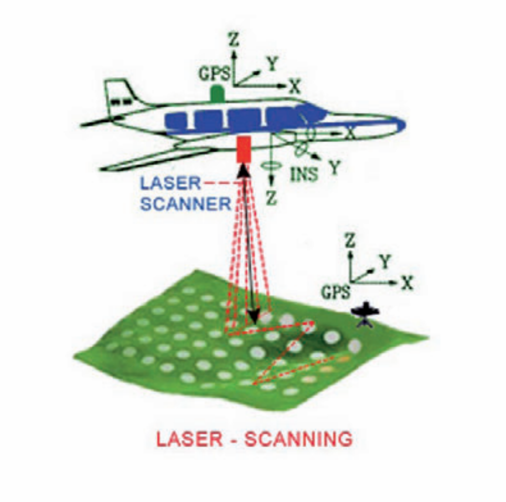


Fig. 1. ALS survey of a piece of land

IMU and GPS are very important for determining the absolute position and orientation of the LIDAR sensors. Airborne GPS is used to determine the x, y and z coordinates of the moving sensor. One ore more GPS based stations on the ground are used to survey the relative position of the airborne GPS. The GPS monitors the altitude and the flight path of the aircraft which observes the 3D data. The IMU is used to correct the errors from the roll, pitch and heading of the plane.

3 Building Blocks Identification

Our method of building block identification starts with looking for vertical wall portions which separate the high from the low land (Tse et al, 2007b). Then the vertical wall segments are connected to form a closed boundary which is the building block. Simulated data of an L-shape and T-shaped building are used to illustrate the method.

There are six main procedures of the algorithm which are:

- Data points resampling
- Vertical segments creation
- Vertical segments tracing

- Building segments clustering
- Building corners determination

3.1 Data Points Resampling

The first step of the algorithm is to convert the raw LIDAR data into a standard triangulated terrain model using the Delaunay triangulation (figures 2). The data points are sampled into a lower resolution triangulation (figure 3) by setting up a disc circle. The sampling method creates a lower resolution triangulation and its dual Voronoi diagram as an index layer. The diameter of the disc circle is between 0.8 to 1 m. Data points are read one by one for testing. A point is accepted if the distance between the first and the second data point is bigger or equal to the diameter of the threshold circle. The sampling method searches point by point and row by row, until it finds another point that is outside the threshold circle (diameter). Each accepted point is separated from the others by at least the diameter of the threshold circle (about 2m).

Fig. 2. High density data points

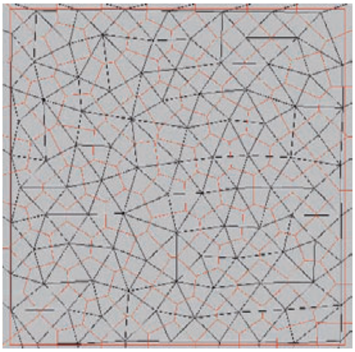


Fig. 3. Lower resolution sample points

The vertical wall portions are captured by partitioning our map into contiguous cells. Some of the big Voronoi cells may contain high and low data points which may contain part of the building boundaries. Our aim is to search and locate the vertical segments in the Voronoi cells.

3.2 Vertical Segments Creation

Our attempt is to detect a vertical surface break in each Voronoi cell by using Principal Components Analysis (PCA) to find the fold axis and "look along" the potential wall segment. The calculation is a 3 x 3 variance-covariance matrix of the coordinates of the points within each cell. The result of three eigenvectors “explains” the overall variance. Figure 4 shows the eigenvector with the smallest eigenvalue to give the line of sight (fold axis) between the high and low points. These segments (thick line in figure 5) are shifted to maximize the difference between the low and high land.

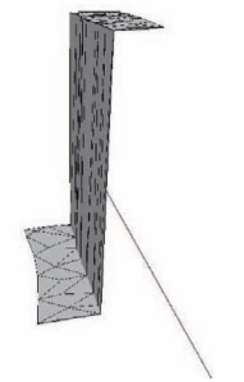


Fig. 4. The eigenvector has the smallest eigenvalue

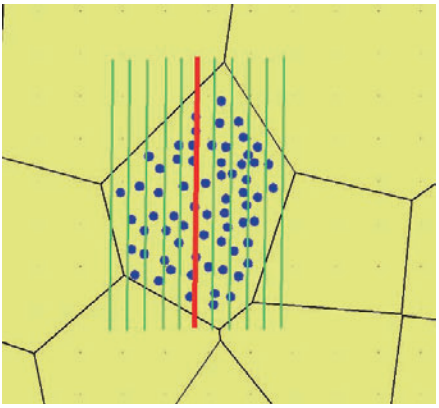


Fig. 5. The splitting (think) line can be shifted to any thin lines positions

3.3 Vertical Segments Tracing

The cells are split along the segments. Then we connect the split edges between adjacent cells to form the closed building boundaries. To identify the closed loop, we start from the first split Voronoi edge and its connections to two other Voronoi edges. Then we search the Voronoi edges in an anti-clockwise order to find all the vertical segments. Figure 6 shows all the vertical segments of an L-shaped building.

When there are extensions to the main building, we need to perform the vertical segments tracing task twice. The searches are the same, but the direction is clockwise. Then two sets of vertical segments produce two sets of building corners. The next step is to compare the two sets of corner points and find the building boundaries with the extensions. Figures 7a and 7b show the two found building boundaries of a T-shaped building with an extension on one side.

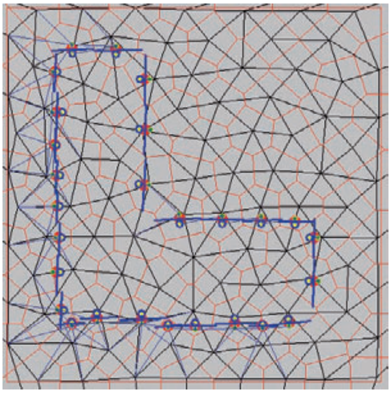


Fig. 6. An L-shaped building boundaries

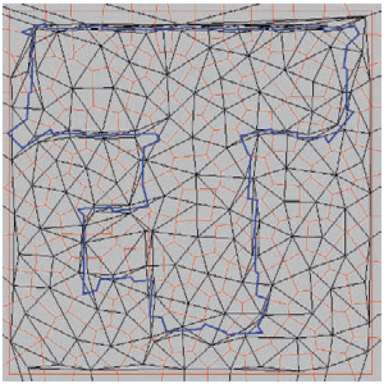


Fig. 7a. Building boundaries include the extension

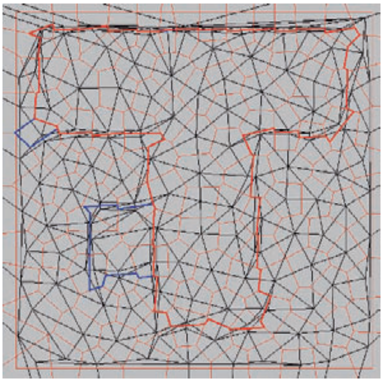
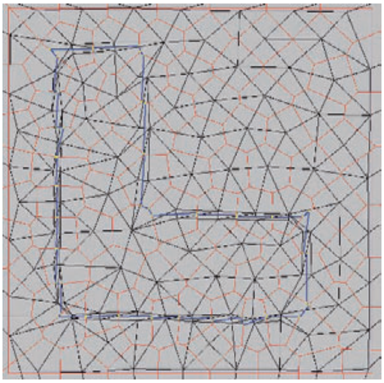


Fig. 7b. Building boundaries without the extension

3.4 Building Segments Clustering

We plot all the split Voronoi edges (vectors) on a circle to cluster the edges according to their orientation in figures 8 and 9. A Delaunay triangulation is formed using the plotted vector points and an MST is formed according to the distances between points. The points are clustered in the same group when they are close together; therefore four groups are formed in figure 9.



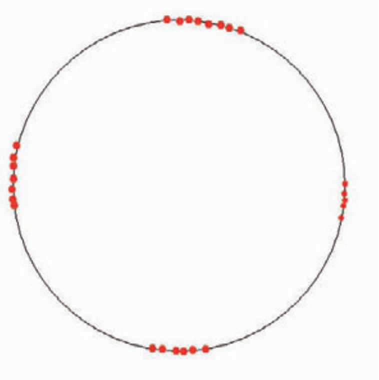


Fig. 8. A closed building boundary **Fig. 9.** Plotted split Voronoi edges on a circle

Each group of edges (same orientation) may represent two different sections of the building outline. Another clustering method is used to separate the edges. If they are geographically not close to each other, they are separated into two groups. Finally each group of the clustered Voronoi edges which share a similar orientation and geographical location, represents each section of the building outline They act as vertical walls of the building to separate the high and low points. Six groups of edges are in six different colours in figure 10.

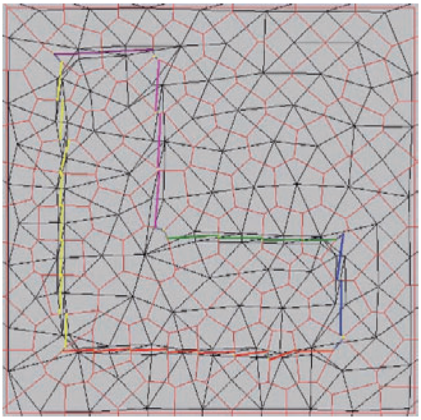


Fig. 10. The clustered Voronoi edges

3.5 Building Corners Determination

An average line is calculated from each group of the clustered segments. Building corners are formed by intersecting the averaged lines. Figure 11 shows the building corners and outline.

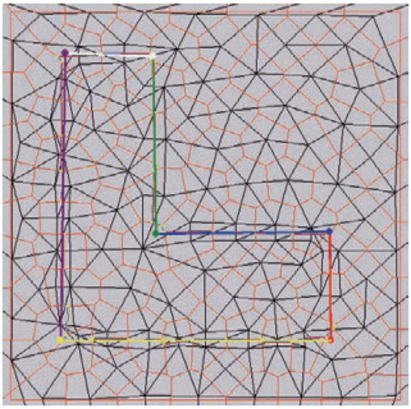


Fig. 11. The L-shape building outline and its corners

In case of a building extension, two building boundaries are created from figures 7a and 7b. Two sets of building corners are formed. The next step is to compare the two sets of corner points, and delete points when they are too close to each other (less than 1m). The final step is to use the remaining points to form the two buildings (figure 12).

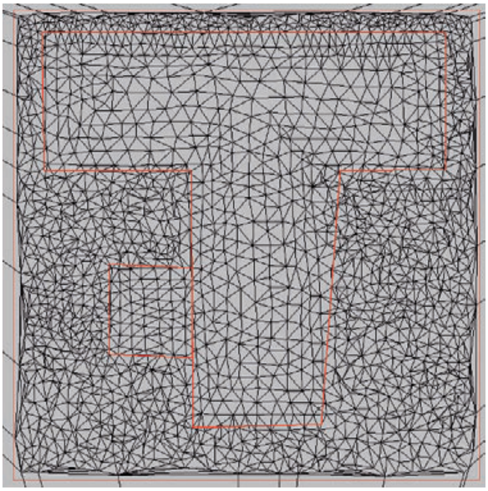


Fig. 12. Two sets of building boundaries

4 Roof Planes Recognition

Once the building boundary is found, and then all interior LIDAR observations and triangles are extracted to identify plane faces. Three clustering methods (Tse et al, 2007a) used to cluster the data points are:

- orientation clustering,
- perpendicular to orientation clustering,
- and geographical location clustering.

Each group of the clustered points represents a single roof plane, with a common description of the plane (its orientation and an averaged “visible” point on it). Then a building is formed by connecting these planar faces and intersections between the roof planes and the vertical walls.

4.1 Orientation Clustering

Orientation clustering separates interior triangles according to their directions. The normal vector of a triangle is a vector perpendicular to it. The first step is to extract the normal vectors from the triangles and plot them on the unit hemisphere (figure 13 and 14).

The normal vectors are connected using the Delaunay Triangulation and an MST is created according to the distance between the normal vectors. Then the vectors are clustered according to the distances between them. The normal vectors are assigned to the same group if they are closer to a threshold. The threshold can be changed or input and is set at 5m in this example.

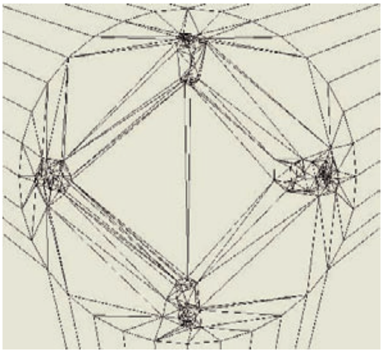


Fig. 13. 2D view of the unit hemisphere

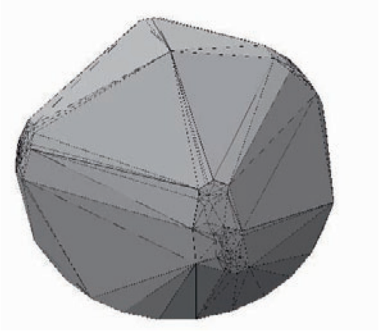


Fig. 14. 3D view of the unit hemisphere

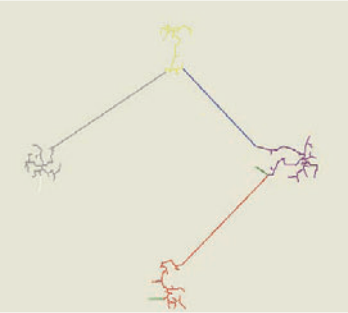


Fig. 15. Four groups of normal vectors

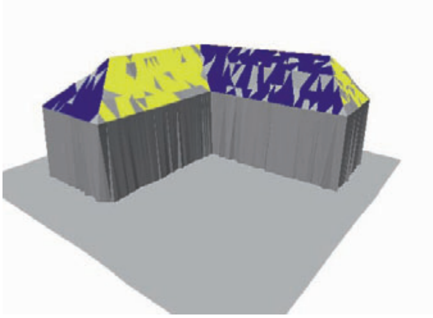


Fig. 16. 3D view of the clustered triangles

Figure 15 shows four groups of normal vectors which represent triangles facing four different directions. Four groups of normal vectors may contain more than four roof planes because coplanar triangles or data points can occur in separate roof portions. A perpendicular to orientation clustering method is used to solve this problem.

4.2 Project onto Average Normal Vector Clustering

To solve the problem, an average normal vector is calculated from each cluster of triangles (darker triangles in figure 16). First we calculate the centre point of each triangle (solid thick lines in figure 17). Then we project the centre points of the triangles onto its averaged normal vector (dashed thick line in figure 17). Equation $\frac{\vec{V} \bullet P}{|\vec{V}|}$ calculates the length between the projected point and the origin of the average normal vector. $\vec{V}$ is the average normal vector and P is the centre points of the triangles.

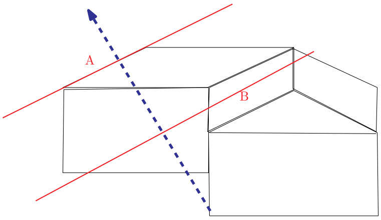


Fig. 17. Roofs A and B are projected onto the averaged normal

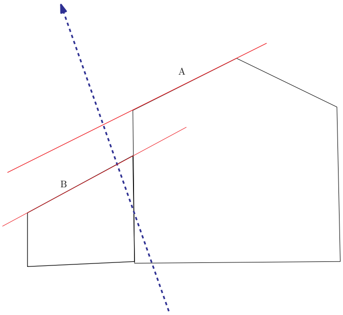


Fig. 18. Main building roof A and extension roof B are projected on its averaged norma l vector

The roof plane of the building extension is separated using this method because a height difference exists between the main and the extended buildings. Triangles on roof A (main building) and B (the extension) are projected onto its averaged normal vector (blue dashed line) in figure 18, and they are separated to two groups.

4.3 Geographical Location Clustering

When the two clustering methods are done, some triangles can still be clustered in the wrong group. For example, triangles on roof A and B in figure 13 are on two roof planes, but they are in the same cluster. Geographical location clustering separates triangles according to their x and y coordinates. A Delaunay Triangulation is created using the centre points under the same group of triangles. The centre points are connected and form an MST, then they are clustered according to their closeness. If the centre points are close to each other (around 2m), they will be clustered into the same group.

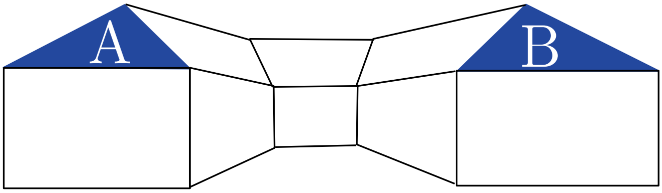


Fig. 19. A building with two hipped roofs facing the same direction

The final step is to find the building corners. Each cluster of triangles represents a roof plane. Region growing clusters the unclassified triangles to find the adjacency relationships between the triangle groups. Unclassified triangles are extracted and grouped according to the nearest classified triangle.

Three planes intersection is used to find the building corners and roof ridges. The found roof planes and vertical walls are intersect each other to form the corner points. Figure 20 shows the intersection points (square points) between the roof planes and the vertical walls of the building.

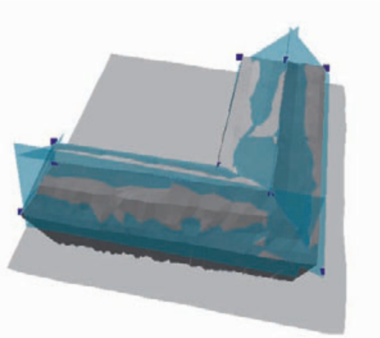


Fig. 20. Building corners

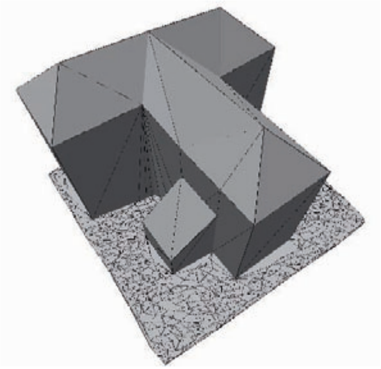


Fig. 21. A reconstructed T-shaped building

5 Building Extrusion

Data points around the building are removed before the reconstruction. The building boundaries are slightly enlarged to avoid any data points which are close to the building walls. A point-in-polygon test is used to extract data points inside the enlarged boundaries for deletion.

We have been successfully using CAD-type Euler Operators to reconstruct our buildings with guaranteed topological connectivity (Tse and Gold, 2001). The building is extruded and the roof is remodelled using the found corner points. Additional Euler operators are used to modify the buildings to allow tunnels or bridges. (Tse and Gold, 2004). Figure 21 shows the reconstructed T-shaped building with a hipped roof.

6 Implementation of Real-world Lidar Data

We have been using real-world LIDAR data to test the roof planes searching algorithm. When an area has multi-scanned lines, the roof clustering method gives an interesting result. Figure 22 is an example of an L-shaped building which has multi-scanned lines. Triangles inside the building are extracted and plotted on the unit hemisphere in figure 23. However the plotted normal vectors do not show any pattern which can be clustered.

Our algorithm does not need perfect data to search for the roof planes, but it does not work if the data is too confusing. We can extract a single scan line to solve the problem. Figure 25 shows a good example of an L-shaped building which is extracted from the multi-scanned lines data.

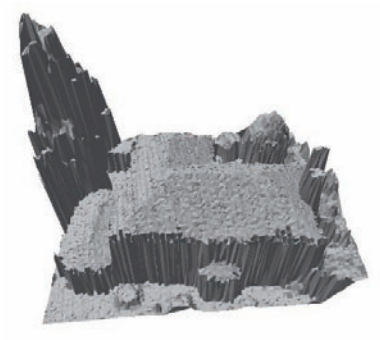


Fig. 22. An L-shaped building

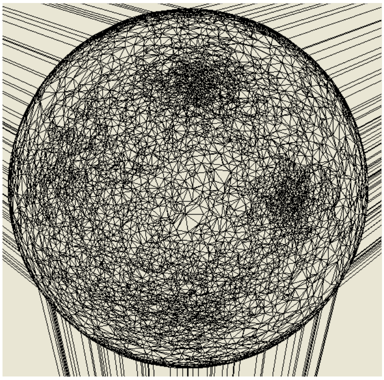
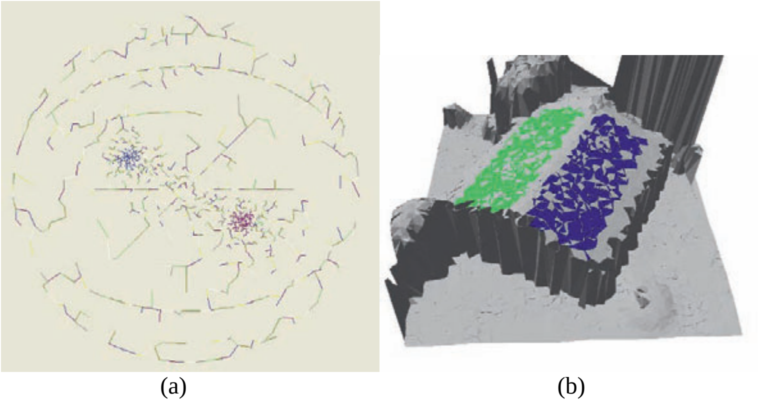


Fig. 23. Normal vectors plotted on the unit hemisphere

However some of the overlapping strips scanned data give a good result for clustering. Starting with a simple two plane gabled roof, it gives a good result in figure 24. The unit hemisphere in figure 24(a) shows the two clustered normal vectors and the 3D view of the two clustered triangles in figure 24(b).



(a) (b)

Fig. 24. An example of a simple two planes gable roof.

Figure 25 is an example of an L-shaped building extracted from multi-scanned lines which can be successfully clustered into four roof planes. Figures 25(a) and (b) show four clusters of normal vectors on the hemisphere and four clusters of triangles. The 3D view of the four roof planes is shown in figure 25(c).

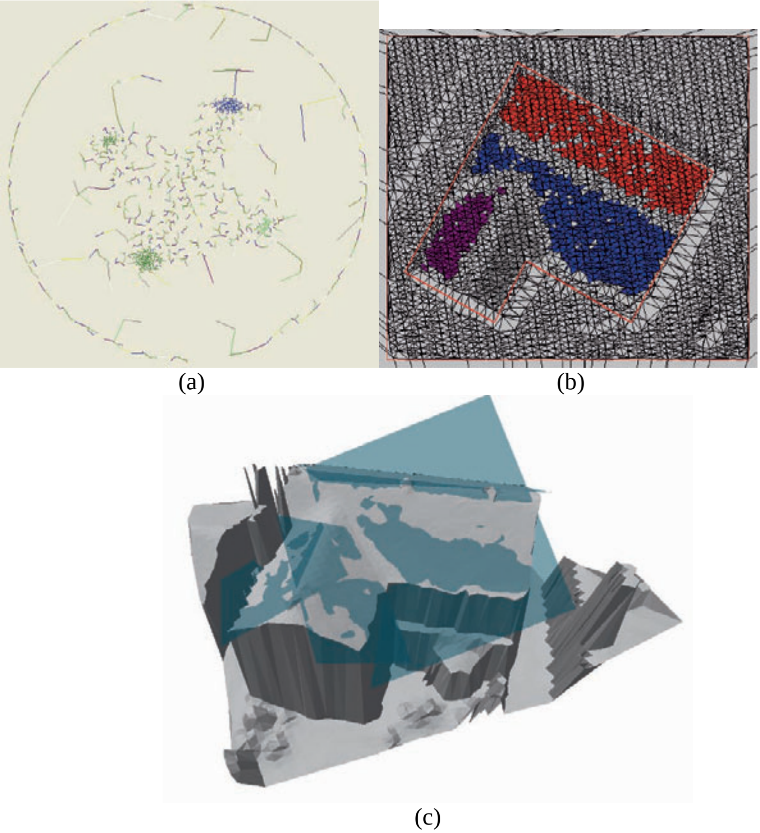
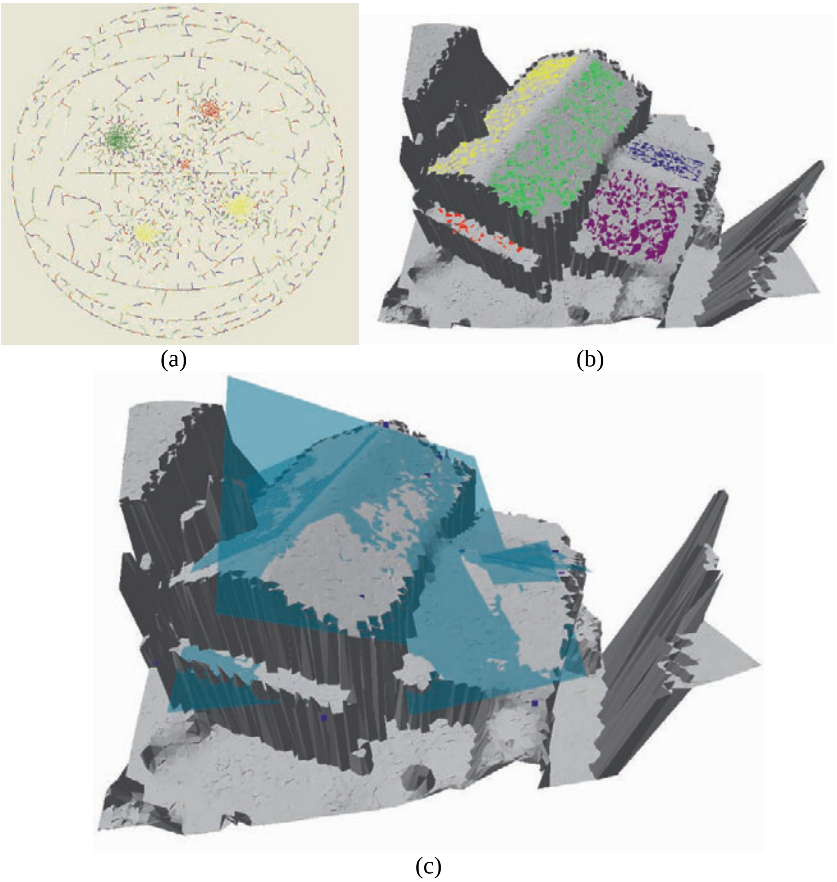

(a) (b)
(c)

Fig. 25. An example of an L-shaped building.

Though multi-scanned line data is confusing to cluster, sometimes it can be clustered successfully. An example of a T-shaped building is extracted from multi-scanned line data (figure 26). Figures 26(a) and 26(b) show the four cluster normal vectors on the hemisphere and the 3D view of the final clustered triangles. Figure 26(c) shows the five roof planes and the calculated intersect points (square points).

7 Conclusion

Delaunay triangulation with the Voronoi diagram gives an alternative approach to reconstruct 3D building models using raw LIDAR data. The advantages of this approach are no initial model of the building shape and no additional data sources. The approach is particularly useful where rapid 3D city models are needed, as little manual intervention is required for many building types.



(a) (b)

(c)

Fig. 26. An example of a T-shaped building.

Modern architecture creates more and more complicated roof structures. Following the research of (Charlesworth et al, 1975), we may be able to model a complicated roof structure like the Wales Millennium Centre with an arch shaped roof structure in the future.

The implementation of building reconstruction algorithms using real-world LIDAR data gives an optimistic result. Problems may occur when the data have multi-scanned lines. However it may be solved by extracting the single-scanned line data.

Real-LIDAR data implementation is time consuming because of the huge amount of data analysed. The algorithm will be capable of reconstructing more than one building at a time in future development.

Acknowledgment

The research discussed and the data used in this paper was supported by the Ordnance Survey, Southampton (Research and Innovation).

References

Ackermann, F. (1999). “Airborne laser scanning - present status and future expectations”, *ISPRS Journal of Photogrammetry & Remote Sensing,* 54(1): p.64–67.

Brenner, C. and Haala, N. (1998) Rapid acquisition of virtual reality city models from multiple data sources. *Int. Arch. Photogrammetry Remote Sensing,* 32 Part 5. Chikatsu, H. and Shimizu, E. p.323-330

Brenner, C.(1999) “Interactive Modelling Tools for 3D Building Reconstruction.” Fritsch, D. & Spiller, R. (ed.) *Photogrammetric Week ‘99’*, p.23-34

Brunn A. and Weidner, U. (1997) “Discriminating building and vegetation areas within digital surface models.” *Technical report, Institute fur Photogrammetrie,* Bonn, Germany.

Charlesworth, H.A.K., Langenberg, C.W. and Ramsden, J. (1975). “Determining axes, axial places and sections of macroscopic folds using computer based methods”. *Canadian Journal Earth Science,* 13, p.54-65.

Forlani, G.; Nardinocchi, C.; Scaioni, M. & Zingaretti, P. (2003) “Building reconstruction and visualization from LIDAR data” *ISPRS International Workshop WG V/4 & INTCOM III/V,* Vision Techniques for digital architectural and archaeological archives, p. 151-156

Rottensteiner, F. and Briese, C. (2002) “Automatic Generation of Building Models from LIDAR Data and the Integration of Aerial Images” In Maas, H.; Vosselman, G. & Streilein, A. (ed.) *Proceedings of the ISPRS working group III/3 workshop ‘3-D reconstruction from airborne laserscanner and InSAR data*, Institute of Photogrammetry and Remote Sensing Dresden University of Technology, 34 Session IV

Rottensteiner, F. and Briese, C. (2003) “Automatic generation of building models from LIDAR data and the integration of aerial images.” In H.-G. Maas, G. Vosselman, and A. Streilein, editors, *Proceedings of the ISPRS working group III/3 workshop ‘3-D reconstruction from airborne laserscanner and InSAR data’*, volume 34 Session IV, Dresden, Germany, Institute of Photogrammetry and Remote Sensing Dresden University of Technology.

Rottensteiner, F., Trinder, J., Clode, S. and Kubik, K. (2007) “Building Detection by Fusion of Airborne Laser Scanner Data and Multi-spectral Images: Performance Evaluation and Sensitivity Analysis” *ISPRS Journal of Photogrammetry & Remote Sensing*, 62, p. 135-149

Sohn, G. and Dowman, I., (2003). “Building extraction using lidar DEMS and IKONOS images.” In H.-G. Maas, G. Vosselman, and A. Streilein, eds., *Proceedings of the ISPRS working group III/3 workshop ‘3-D reconstruction* *from airborne laserscanner and InSAR data'*, volume 34 Session IV. Institute of Photogrammetry and Remote Sensing Dresden University of Technology, Dresden, Germany.

Sohn, G. and Dowman, I. J., (2004). "Extraction of buildings from high resolution satellite data and LIDAR", *Proceedings of ISPRS 20th Congress WGIII/4 Automated Object Extraction.* Istanbul, Turkey. p.345-355.

Suveg, I. and Vosselman, G., (2001). "3D Building Reconstruction by Map Based Generation and Evaluation of Hypotheses." BMVC01.

Suveg, I. and Vosselman, G., (2004). "Reconstruction of 3D building models from aerial images and maps." *ISPRS Journal of Photogrammetry & Remote Sensing,* 58(3): p.202–224.

Tse, R. and Gold, C., (2001). "Terrain, dinosaurs and cadastres -options for three-dimension modelling." In C. Lemmen and P. van Oosterom, eds., *Proceedings: International Workshop on "3D Cadastres*, Delft, The Netherlands. P.243–257.

Tse, R. and Gold, C., (2004). "TIN meets CAD - extending the TIN concept in GIS". *Future Generation Computer Systems (Geocomputation),* 20(7) p.1171–1184.

Tse, R., Gold, C., and Kidner, D., (2007a) "3D City Modelling from LIDAR Data". *Proceedings of Lecture Notes in Geoinformation and Cartography,* Delft, The Netherlands. p.161-175.

Tse, R., Gold, C., and Kidner, D., (2007b) "Building Reconstruction Using LIDAR Data" *Proceedings of Dynamic and Multi-dimensional GIS 2007,* Urmchi, China p.121-126.

Vosselman, G. (1999) "Building Reconstruction Using Planar Faces in Very High Density Height Data" *International Archives of Photogrammetry and Remote Sensing*, 32, part 3/2W5 , p. 87-92

Vosselman, G. and Dijkman, S., (2001). "3D building model reconstruction from point clouds and ground plans." *Proceedings of International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences,* volume 34, part 3/W4,. Annapolis, MA, USA. p. 37–43.

Vosselman, G., (2003). "3D reconstruction of roads and trees for city modelling." In H.-G. Maas, G. Vosselman, and A. Streilein, eds., *Proceedings of the ISPRS working group III/3 workshop '3-D reconstruction from airborne laserscanner and InSAR data*, volume 34, Part 3/W13. Institute of Photogrammetry and Remote Sensing Dresden University of Technology, Dresden, Germany.

Wang, Z. & Schenk, T. (2000) "Building Extraction and Reconstruction from LIDAR Data", *In proceedings of IAPRS,* July, 33, part B3, p. 958-964

A New Approach for Mountain Areas Cartography

Loïc Gondol, Arnaud Le Bris, François Lecordix

Institut Géographique National (IGN France)
2-4 Avenue Pasteur
94165 SAINT-MANDE Cedex – France
email: {loic.gondol, arnaud.le-bris, francois.lecordix}@ign.fr

Abstract

From now on, the French National Mapping Agency (IGN France) is set up with the BD TOPO®. This is a topographic vector database that covers the whole national territory. IGN has decided to product base maps at 1:25k and 1:50k from this database. On topographic mountain maps, rocks areas are among the most difficult map elements to represent, dealing with digital cartography. In the past, they were drawn manually by experienced cartographers, using graphic means and working with aerial photographs. Nowadays, we need to focus on two points with a digital approach. The first one is the detection and an automated classification of concerned areas. The next one is the development of an adapted cartographic representation of rocks and screes areas. This article presents the first results on these problems. As far as possible, we aim at having automated high mountain cartography with lower production costs. Also, we would like it to be as expressive as it was in previous maps. This is to keep the same cartographic quality of the current base map at 1:25k and 1:50k.

Keywords: cartography, representation, mountain, classification, data fusion

1. Context

1.1 The BD TOPO®

The French National Mapping Agency has decided to create a topographic database for several years, named the BD TOPO®. This vector database provides some geographic information about road, rails, electric and hydrographical networks, but also about buildings, administrative boundaries, toponymy, land use and relief. The first version of the BD TOPO® (V1) was mainly acquired by digital restitution, and it included all the needed information for the map at 1:25k. However it required a huge workload and therefore too much time to realise this database on the whole territory. In 2000, IGN France chose a new lighter specification for the BD TOPO® that allowed its completion on the whole territory by the beginning of 2007.

Beyond various applications linked to GIS, the BD TOPO® is also useful to derive the base map, produced and diffused by IGN France. The base map includes a topographic map at 1:25k and another one at 1:50k. Since 1993, 450 out of 1800 maps at 1:25k have entirely been produced from the BD TOPO® V1 on several areas of the national territory.

Nevertheless, up to now, none map covering a high mountain area has been done from the BD TOPO® V1. In fact, the cartographic representation of these areas is knotty dealing with a digital production. On high mountain areas, all current base map versions come from revisions of former versions drawn manually years ago. They are not originally from BD TOPO®. As a result, we currently have 2 distinct processes to collect data, update the database and the topographic map.

1.2. The New Base Map Project

In 2004, IGN France decided to launch the New Base Map Project in order to derive the base map from the BD TOPO® with its new specifications, and to reduce update costs of next base map versions. The process is planned to work on the whole territory, and then on high mountain areas.

Among other issues, this project has to provide solutions to retrieve the needed information for the map that is lacking in the new specifications of the BD TOPO®. In particular, the land use is one of the main incomplete themes. This recovery issue is especially perceptible in high mountain areas where a paramount part of the map information deals with mountain land use: rocks, screes, glaciers… Without these themes, mountain maps would appear uncluttered and could not satisfy users.

It is not enough to recover the needed information in high mountain areas. Indeed, another issue arises that has to be carried out by the New Base Map Project. It deals with the digital representation of these specific themes, in order to get the most expressive possible result on concerned areas.

1.3. From manual to digital

During the history of cartography, a lot of solutions have been tested for the cartographic representation of mountain areas on topographic maps. In (Imhof 1958), a literature review of these solutions is presented. The relief aspect often comes from a combination of different techniques: hill shading, graphic means, contour lines with a colorimetry depending on areas (glaciers, rocks, etc.). But up to now, the solutions that have been carried out and the best provided results are only the outcome of traditional graphic techniques. At present, to edit topographic maps (at 1:25 000 or 1:50 000) in high mountain areas, the National Mapping Agencies are not able to produce numerically new maps without some elements which were drawn manually in the past by traditional cartographers. Figure 1.a. shows an extract of an IGN topographic map at 1:25k, where an example of this cartographic result on the Alps is shown.

In the beginning of the 2000's, IGN France launched the first studies in the purpose of considering the feasibility of substituting digital solutions for traditional methods (Le Men et al. 2002). These works focused on the automatic extraction of map needed information (rocks areas, screes, glaciers) and provided the first steps for data representation. These study results provided in a digitally manner are presented on figure 1.b..

The New Base Map Project has been continuing these first works. In particular, it looks at improving final cartographic results. Dealing with data retrieval, the MATIS, which is a research laboratory of IGN in image processing, has been carrying out the study.

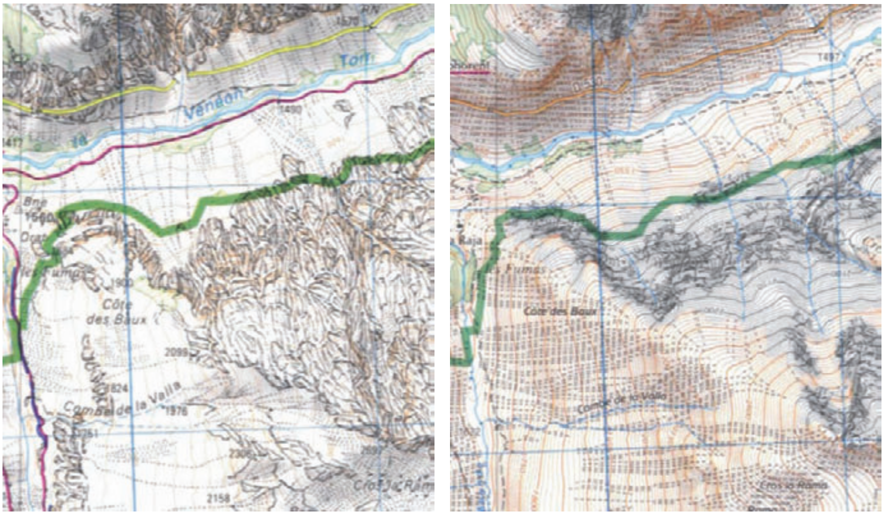


Fig. 1. On the left (fig.1.a.) IGN present handmade topographic map at 1:25k and on the right (fig.1.b.) first map digitally obtained by (Le Men et al. 2002) on the same area

2. Information extraction

2.1 Problems and proposed solutions

Two ways of getting back landcover information that lacks from BD TOPO® but is necessary to obtain the topographic map are possible: it can be extracted either from aerial (ortho-)images (through (semi-)automatic classification) or from present maps. As the cartography of these missing landcover themes is not up to date in present maps, the first solution has been chosen (Le Men et al. 2002). Furthermore, it could also be used afterwards for map updating. So, the chosen solution consists in extracting landcover information out of aerial orthophotos from IGN France's orthoimage database BD ORTHO® through a supervised classification method. Nevertheless, it must be said that extracting such information from satellite images with bands dedicated to remote sensing (such as in (Paul 2003)) could have been possible too, but in the present case, using aerial data is interesting since it is available and captured regularly by IGN.

The only lacking landcover themes necessary to make the map are rocks, screes and glaciers. However the classification legend must contain more items than these three seeked lacking themes to obtain a landcover classification of whole mountainous areas. That's why it consists of the six following classes: rocks, screes, glaciers, forests, high mountain pastures and water areas.

In high mountain areas, landcover extraction using aerial photographs is bothered by several phenomena:

- Shady areas are often very large.
- The radiometry of a same theme can greatly vary within the image. This phenomenon can be due to illumination variations related to the rough relief, but it can also be "artificial" (since the image is in fact a mosaic of orthorectified aerial images which have not been captured at the same time and have undergone several different radiometric treatments) or natural (as in case of changes in geology inside the area).
- Some of the landcover themes have a very close radiometry: they look like each other on the image as for instance some screes (especially riverbed screes) which are almost as light as glaciers or lakes which are often difficult to distinguish from rocks in shadow, or even rocks and screes. This phenomenon is increased by the variations of radiometry explained above.

As a consequence, image information is not sufficient to obtain a correct classification. Nevertheless, the introduction of complementary information in the classification process can improve the result. Two kinds of external knowledge are useful:

- Mountainous landcover is strongly related to the relief, it means to altitude, slope and orientation. Those variables can be easily computed from a Digital Terrain Model and used to obtain a probability to find the different themes.
- Knowledge from another database can also be used. In the present case, European database CORINE Land Cover 2000 (CLC2000) dealing with land use has been used. It is more generalized than the base map since the smallest mapped area is 25 hectares and the better scale to use the data base is 1:100k (Bossard et al. 2000) (fig.2.). Its legend is different too, with a varying semantic precision : for example, different kinds of forests are separated, whereas rocks and screes are contained within a unique class. Besides, in other cases, some CLC2000 classes describe intermediate situations between several of our classes: "forest and evolving shrubby vegetation" is linked both to forest and pastures.

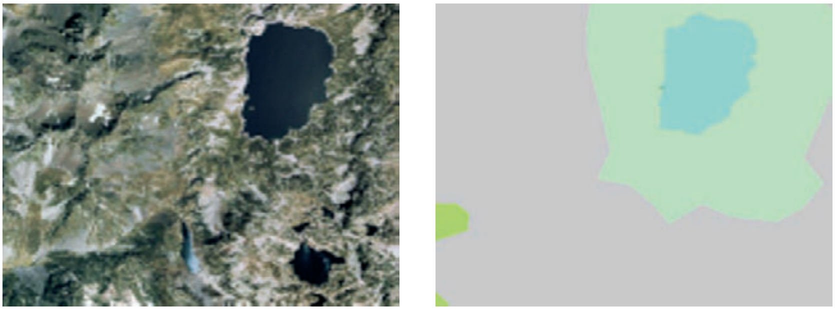


Fig. 2. CLC2000 is a more generalized database

As shadow areas are important on the image, they must be taken into account to obtain a classification of the whole area. A first way to achieve this could consist in correcting the radiometry in shadow areas (after having detected them). However, it would be limited by uncertainties concerning the DTM and the fact an orthoimage (it means a mosaic of merged orthorectified aerial photographs captured at different times and having undergone different radiometric treatments) is used. As the main problem with shadows is the fact that the radiometric model of a class will be completely different in shadow and in light, a second solution consists in dividing each class “C” in two classes “C in shadow” and “C in light” so that two distinct models are computed for each theme. Of course, these two classes are aggregated at the end of the classification. The second solution has been retained for the present method. Nevertheless, a method using radiometric correction in shady areas has been successfully developed by (Le Men et al. 2002) during the first studies about this issue.

Even though the available information is not sufficient to precisely detect and correct shadows, it can be used to compute an approximate prior probability for each pixel of the image to lie in shadow knowing the DTM (since the beginning and final time of data capture of all images are known even if the exact capture time of each pixel of the orthoimage is not precisely known). It could help to discriminate dark themes (such as water) in light from themes in shadow.

New channels can be computed from the original bands (red-green-blue-infrared) of the orthoimage. Associations of these derived channels can lead to better results since some channels are more efficient than others to separate some classes.

2.2. Method

The landcover extraction method consists of two steps.

- The orthoimage is firstly segmented into homogeneous regions (figure 3). A description of the used segmentation tool can be found in (Guigues 2004) and (Guigues et al 2006).

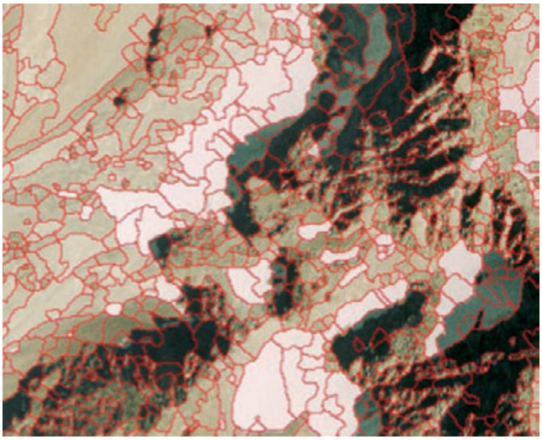


Fig. 3. The image is segmented in homogeneous regions

- Segmentation’s regions are then classified by the tool presented in (Trias-Sanz 2006) and (Trias-Sanz and Boldo 2005). The way image information is used in the classification process consists of two steps :
  - A model is computed from training data captured by an operator. First, for each class, the best parameters of several statistical distributions (such as gaussian, laplacian laws but also histograms (raw or obtained by kernel density estimation)...) are computed to fit to the radiometric n-dimensional histogram of the class (with n standing for the number of channels used in the classification). Then the best model is selected thanks to a Bayes-Information-Criterion (Schwarz 1978) allowing to choose an alternative between fit to data and model complexity (The more complex the model is, the more degrees of freedom it has, the better it is able to describe training data, but the most it is also at risk to “stick” too much to training data without describing the whole groundtruth as well as a simpler one.) Once this model has been computed, it becomes possible to compute the probability that a pixel s has a certain value $I(s)$ if it belongs to class c, it means to obtain $P_{radiometricmodel}(I(s)|c(s)=c)$ with $c(z)$ standing for region or pixel “z's class”.
  - In the present case, a MAP per region classification algorithm is used since it allows to take easily into account external information (from relief, from CLC2000 and concerning shadow probability in the

present case) as prior probabilities. Classifying regions prevents from obtaining too noisy results. With this classification method, the label $c_0(R)$ given to a region R is its most probable class according to the radiometric model previously estimated and prior probabilities. Hence, $c_0(R)$ is the class c that maximizes the following function: $\prod_{i \in \text{extern information sources}} (P_i(c(R)=c))^{a_i} . \left(\prod_{\text{pixel}\, s \in R} P_{\text{radiometricmodel}}(I(s)|c(s)=c)\right)^{\frac{1}{card(R)}}$ with $I(s)$ standing for the radiometry vector of pixel s, $c(z)$ meaning region or pixel "z's class" and $P(c(z)=c)$ standing for the probability for pixel or region z to belong to class c. The a_i terms stands for weight parameters balancing the different prior probability sources.

External information is then introduced as prior probabilities in the classification process (Le Bris and Boldo 2007). This requires an interpretation of this knowledge in terms of probability.

As previously said, in mountainous areas, landcover is strongly related to relief. Thus it depends on altitude, slope and orientation making it possible to compute the probability to find the different themes knowing these three parameters. Such a model is proposed by (Le Men et al. 2002) from physical geography knowledge (such as such as the lowest and highest limits of landcover themes, etc.) presented in (Elhai 1968) and (Lacambre 2001). It consists of two distinct models (made of piecewise linear mathematical functions) depending on altitude and slope (figure 4). Orientation has a significant influence only on forests and glaciers. Thus it is taken into account only for these themes as a modification of the altitude.

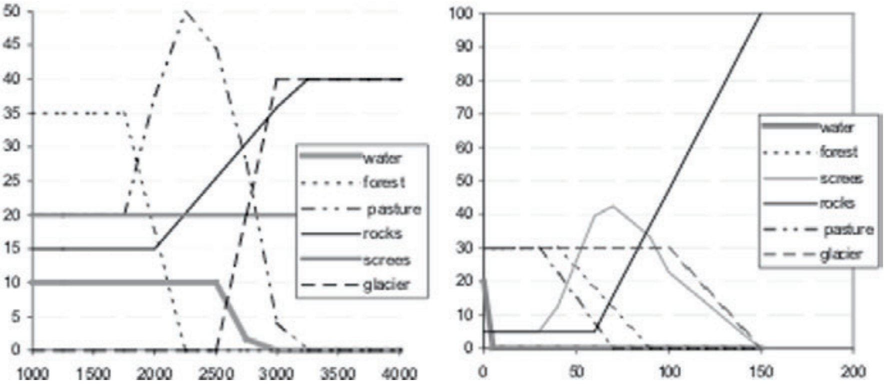


Fig. 4. On the left [right], probability of finding themes knowing altitude [slope].

CLC2000 information must also be introduced into the classification process. As this database is more generalized than expected results, a CLC2000 area can contain several classification themes. Besides, as its legend is different, several themes of our classification can be related to one CLC2000 item, and vice versa. Therefore the introduction of information from CLC2000 in the classification process must deal with those two kinds of uncertainties. Therefore CLC2000 is interpreted in terms of probability with an empirical probability model: for each CLC2000 item $T_{CLC\,2000}$ and for each classification class $T_{classif}$, a probability value $P(T_{classif} \mid T_{CLC\,2000})$ is empirically defined. For instance, for CLC2000 class "forest and evolving shrubby vegetation", the probability to find water and glaciers areas is null but equals 77% for forests, 20% for pastures, 1% for rocks and 2% for screes.

Shadow knowledge is also taken into account as probabilities of being in shadow given the relief, i.e. the DTM, and the interval of image capture time.

A balance between these different sources is applied, allowing to give more or less strength to some of them but also to modify the generalisation level of the result (the higher CLC2000 weight is, the more generalised results are).

2.3. Results

This method has been tested in three study cases. The first test zone, located in the Alps, near St-Christophe-en-Oisans, has already been the test zone of the preliminary study (Le Men et al. 2002). All the classification themes are present, but only an old 3-bands orthoimage made from argentic scanned photographs is available there. Due to very important radiometric variations within the image, it was sometimes difficult, even for an operator, to identify themes. The second test area is located in the Pyrenees, around the Ossau-peak. Orthophotos have been produced from digital pictures in red-blue-green-infrared. All classification themes, except significant glaciers, were present. The third test zone has been chosen in the Alps near Modane, where 4-bands orthoimages (captured by a digital camera) are available and where all the classification themes are present. This area is interesting since available data (orthoimage and DTM) corresponds to what would be used in a production context. Furthermore, digital topographic data captured by operators in the 1990's are available there and can be used to evaluate the classification results.

Results have been visually evaluated (on whole images) revealing no major errors. Nevertheless, over detection can occur concerning glaciers on Modane area : snow and glaciers can be mistook. Furthermore, most of the classification regions have a meaningful size to be relevant on the map.

They have also been numerically (on smaller test zones in the image) evaluated by computing confusion matrices comparing test data captured by an operator to classification results. Nevertheless, it is sometimes difficult to evaluate the obtained results since even a human operator can find it hard to discriminate landcover themes in some parts of the test zones. These results can be seen in table 1.

Table 1. Evaluation of the classification results on the three test zones. us-ac corresponds to the probability for a classified pixel to be really part of its class whereas pr-ac corresponds to the probability for a ground pixel belonging to a given class to be well classified

	St-Christophe-en-Oisans (Alps)		Ossau (Pyrenees)		Modane (Alps)	
	With complementary knowledge					
	us-ac	pr-ac	us-ac	pr-ac	us-ac	pr-ac
Water (Lakes)	/	/	100.0	76.4	98.3	40.6
Forest	81.9	65.2	89.0	95,8	98.7	88.8
Pasture	71.1	52.2	96.3	85.1	/	/
Rocks	76.4	69.9	71.0	87.2	54.2	76.5
Screes	54.7	73.3	88.0	83.1	68.9	52.3
Glaciers	58.6	69.5	98.3	72.2	91.5	67.3
Well classified pixels	67.0 %		87.4 %		75.6 %	
	Only image information					
Well classified pixels	55 %		75 %		67.6 %	

Several parameters have been tested. In particular, these tests have shown that almost equivalent satisfying results are provided by several channels associations such as the intensity-hue-NDVI (Normalized-Difference-Vegetation-Index) one, or the three channels of the Karhunen-Loève colorimetric space (Wang et al. 2003). The significance of the external knowledge and the importance of the balance between them have also been proven.

At the end of the classification, classes “c in shadow” and “c in light” are aggregated in a single class “c”. Then, only themes lacking from BD TOPO® (rocks, screes and glaciers) are kept since information about other themes is already available. So all needed landcover information is then available to make the 1:25k-scale topographic map and now needs to be mapped.

3. Cartographic representation

3.1. Further processes for cartographic representation

Classification results can not be directly used to draw the map: they have to be both simplified and enriched for cartographic requirements. Thus additionnal processes have been designed to post-process them.

On one hand, too small areas (without cartographic meaning) are filtered out because they would be irrelevant and unreadable at a given scale (1:25k).

On the other hand, screes, rocks and glaciers areas form the heart of the cartographic representation. They delineate areas we are interested in the present study. More particularly, rocks and screes correspond to rough relief areas: their cartography should therefore also be a cartography of relief, an interpretation of terrain characteristics. Thus, classification results concerning these themes have to be enriched with additionnal information extracted from DTM to improve the final cartographic representation.

Firstly, rocks and screes areas are split up depending on the slope value in order to associate a proper symbolisation to these different kinds of areas. A threshold allows to distinguish steep slopes from gentle ones: a 100% slope is the limit for rocks areas whereas it is 50% slope for screes ones.

A rocky areas classification following slope orientation criterion is then processed. As previously, it aims at differentiating the cartographic representations depending on terrain characteristics. On one hand, this classification has to contain enough details to illustrate as best as possible the terrain diversity and complexity. On the other hand, areas have to be big enough to be readable on a map at 1:25k. After several tries, the creation of 18 classes at 20° regular intervals has appeared to be the best choice in the present case. However, be careful not to generalise without any cares this result. The first point is to respect qualitative criteria explained above. Finally, rocks areas are divided in 36 different classes, depending both on slope orientation and slope value.

Screes areas characterised by a steep slope will be drawn with growing points along the slope. Therefore, main slope lines (i.e. the symbolisation skeleton in these areas) must be extracted, taking into account density constraints. A way of doing this consists in visiting each node of a meshing covering the area and then considering the current node as the starting point of a slope line along which we go down until exiting the steep slope screes area or being nearby another existing line.

Upper borders in steep slope rocky areas are also required to improve the map readability by illustrating breaks in slopes, considered to be especially dangerous for hikers. These lines are extracted keeping only outline sections of rocky areas with their slope oriented within the area. Besides, they have to be longer than a threshold and their pixels' altitude have to be greater than at least half of their neighbours.

Ridges extraction is also currently tested to improve the map readablity.

3.2. Method

In the previous paragraph, we looked over useful data for cartography in mountain areas. Now we need to affect a symbolisation to these data. This is what the user really sees and reads on the map. The whole issue is to represent in the best possible way the terrain complexity. The main point is to understand accurately the map, so that users can not be mistaken and thus avoid dangerous zones for example. When most of users look at the topographic map at a big scale, they trust in its accuracy. This kind of map has to keep this asset.

Besides reflecting the reality, the representation has to be as automatic as possible, as one of the main purposes of this study is to reduce the map production cost. Up to now, rocks have been drawn manually by experienced cartographers, using graphic means and working with aerial photographs. This technique produces very good results but the point is that it is very expensive. (Hurni et al. 2001) has studied some methods to automate the cartographic representation. However, he advocates the combination of manual tasks with automatic ones. According to him, digital methods can be applied only in a limited spectrum of tasks. Control of the full process by an experienced operator is still necessary and desirable, in order to keep the graphic quality that is characteristic of Swiss maps. At IGN France, we wonder as well how we can improve current results in digital cartographic representation. So we have kept carrying out some researches about this issue.

According to several case studies done in mountain areas, hachures seem to be a relevant way for rocks representation. They provide both a good perception relief and an appreciated graphic result. This technique outcomes from a former one more general, used historically to depict the relief with hachures. They were generally drawn along the biggest slope so that the user could mentally think mounts and valleys. In our study case, we judge necessary to have ridge-lines depicting major contours and ridge crests so they complement hachures representation. Maps from the Swiss National Map Series produced by the Federal Office of Topography illustrate this way of representing rocky areas. Fill hachures are plotted manually either along the biggest slope where this one is steep or along contour lines. An operator estimates the slope value looking at aerial photographs. As the human eye is a subjective tool, there is no mathematic threshold to determine if a slope is steep or not.

This way of representing rocky areas has been adapted in this study, trying to incorporate it in a digital process. A hatched pattern has been designed to fill in rocky areas where the slope exceeds the given threshold of 100%. In order to have a fine visual perception, (Imhof 1982) gives in particular some advices dealing with hachures design and related parameters. They have been taken into account to draw up this pattern. Following the different orientation slope classes, this one is always oriented according to the steep slope. For instance, rocky areas with a slope oriented from 0 to 20° will have a 10° oriented hatched pattern. Hence the interest to have a classification depending on the slopes orientation. This is necessary to let the pattern follow the neighbour steep slope, whatever the slope orientation value is.

Another hatched pattern has been defined for rocky areas with gentle slopes. In this case, hachures are more spaced. Added points symbols and areas randomly disposed are part of the pattern to mean the presence of isolated rocky blocks. The pattern is oriented following the average tangent to the contour line for each class. The hachures logic spacing used in this case is the same for contour lines. The denser they are, the steeper the slope is, and vice versa.

Hachures are voluntarily irregular to cleave better to the real world. In fact, structure lines and the general rocks texture are hardly ever regular and geometric. Thus, when we had to design a pattern, we paid special attention to the possibility of reproducing it without any spatial discontinuity. An off-setting on the edge of the pattern does not produce a satisfying visual effect. This is especially true if there are no structure lines to hide it.

The screes representation is characterised by points symbols. There are small round points and irregular round shapes that mentally suggest to the map reader stones or rocks. Where there are steep slopes, screes tend to form streams because of unsteady rocks rolling down. These rocks crumbling leave a trace on the ground that can be observed on aerial photographs. We tried to duplicate as precisely as possible those characteristics. Depending on the slope value, we differentiated the screes representation. (Jenny 2008) presents a digital method for scree representation developed at Zürich Institute of Cartography. It gives almost similar results, though there is a difference in his process with shadow areas

Lines along the steepest slope are only extracted in screes areas where the slope rises above 50%. They are necessary for the representation support in these zones. In fact, round points symbols are computed and placed at regular intervals along these lines. The symbols diameter increases in a linear way when we go down along the slope. Visually, the crumbling effect appears thanks to the lines support. They are not visible on the map but we can easily imagine them.

Dealing with screes representation located in slopes under 50%, a pattern has been drawn. It is made up of punctual symbols and irregular round shapes, and suggests rocks randomly disposed. On the contrary of pattern used in rocky areas, this last is not oriented. This is to keep the randomness of these zones structures.

Upper borders in rocky areas complement the cartography and give a global structure of these mountain areas. They are useful to mean cliff tops and broken grounds. They are drawn with two lines. The first one is black to underline the break; the other one is dark grey to mean the shadow, which is a frequent characteristic within these zones. The black line has a varying width. This is to avoid a symbolisation that would appear too much geometric. Besides structuring the map in mountainous areas, these upper borders warn against potential dangers that may have not been clearly seen with hachures.

Previous listed elements are put together with more classic and often used ones talking about mountain cartography and relief: I mean contour lines and hill shading.

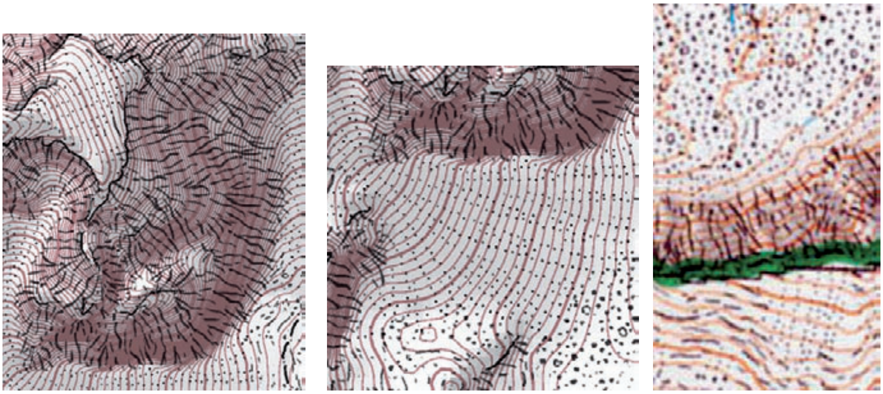


Fig. 5. Representation of rocks and screes depending on slope

3.3. Results

This representation has been tested in Pyrenees nearby the Ossau peak. More recently, other tests have been done in French Alps, within the Vanoise National Park region nearby Modane.(Figures 7 and 8)

Two softwares have been used to match digital data and obtain a resulting paper map since data were prepared (i.e. imported and georeferenced) thanks to the GIS Geoconcept (produced by Geonconcept SA ltd) before being exported and integrated in Mercator (produced by Star-Apic ltd), a cartographic symbolisation software allowing to manage all data representation with layers in order to do map-printing. This software has been used at IGN France for a long time to manage the end of topographic maps production flowlines, until they are printed.

We have got some results that seem to be hopeful. The map has a good global readability. The whole coherency between themes is satisfying as no layers overrides others. Furthermore, we can easily distinguish the different kind of symbolised zones. The representation rationale and the patterns used here allow the map reader to associate the ground nature. Finally, this is promising all the more that this cartography has been entirely done in a digital way, which was one of the goals.

A visual comparison has been made with the former map manually realised. Besides, some expert users have been asked to compare the two maps. They let us know their feedback. This evaluation enables us in particular to identify a weakness dealing with the digital hill shading. In fact, it is far less expressive than the old one drawn manually. So it still needs to be improved, given that there are a lot of current researches in this field. The cartographic contrasts on the digital map need to be underlined, in order to have a better information structure. For instance, dangerous areas should be perceived directly. Another problem deals with glaciers and the detection of this theme. Depending on when aerial photographs have been taken, snow-covered areas can be still present even in summertime. For the time being, these areas are automatically classified as being glaciers whereas they really correspond to screes or rocky areas. We currently try to find out a solution to this problem.

Of course, this method is to be reproduced on other mountain areas. Remind us that the New Base Map project aims at producing a homogeneous representation on the whole territory, and in particular concerning rocky areas. Currently, this is not the case in France. The result is different and depends on both cartographers and when it has been produced. In order to test the reliability and the scalability of this method, we understand better why this kind of test has to be generalised to other areas.

4. Conclusion

The New Base Map project carried out several studies dealing with image processing and mountain cartography. One of the main purposes is to get an automated digital high mountain cartography as good as the one we had before with manual processes. This study has been done combining several data sources, i.e. orthophotos, DTM and external knowledge, an automated extraction of rocky, screes and glaciers areas, and additional data relief-related. These data allow to obtain an automated cartographic representation that aims at improving the relief and the "expression" of high mountain zones. The whole process is described on figure 6.

The first results obtained in Pyrenees and Alps are hopeful. Further tests could include a customer survey, especially concerning mountain map users. Remarks from customers could help us to improve the final process and particular stages, in particular the hill shading. Then, we might be able to introduce this global solution within the IGN base map production flowline.

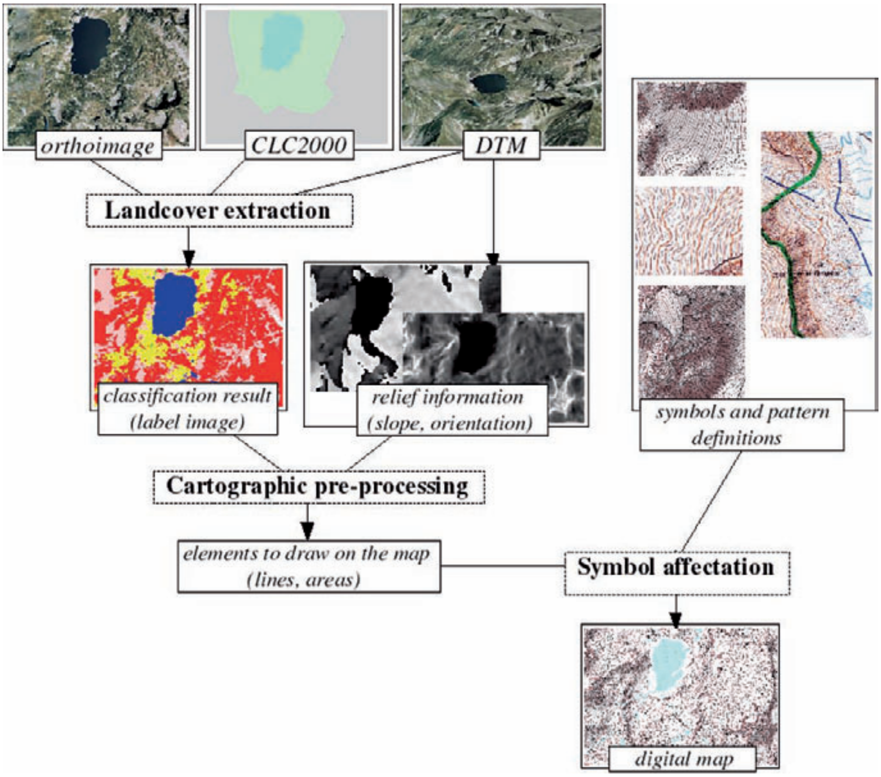

orthoimage
CLC2000
DTM
Landcover extraction
classification result (label image)
relief information (slope, orientation)
symbols and pattern definitions
Cartographic pre-processing
elements to draw on the map (lines, areas)
Symbol affectation
digital map

Fig. 6. Description of the whole mountain cartography process

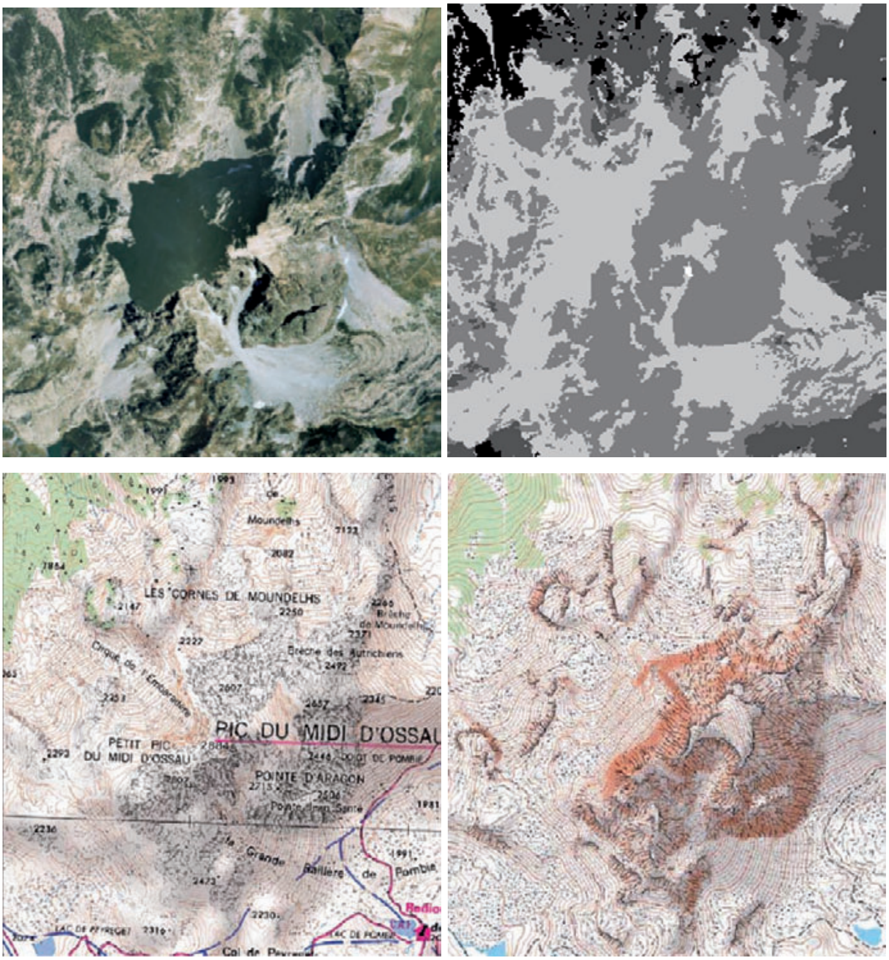


Fig. 7. On the first line, crop of the orthoimage near Ossau-Peak and obtained classification on the same area (from the darkest to the lightest: forest, pasture, rocks, screes and glaciers). On the second line, present handmade topographic map and new fully digital made map on the same area

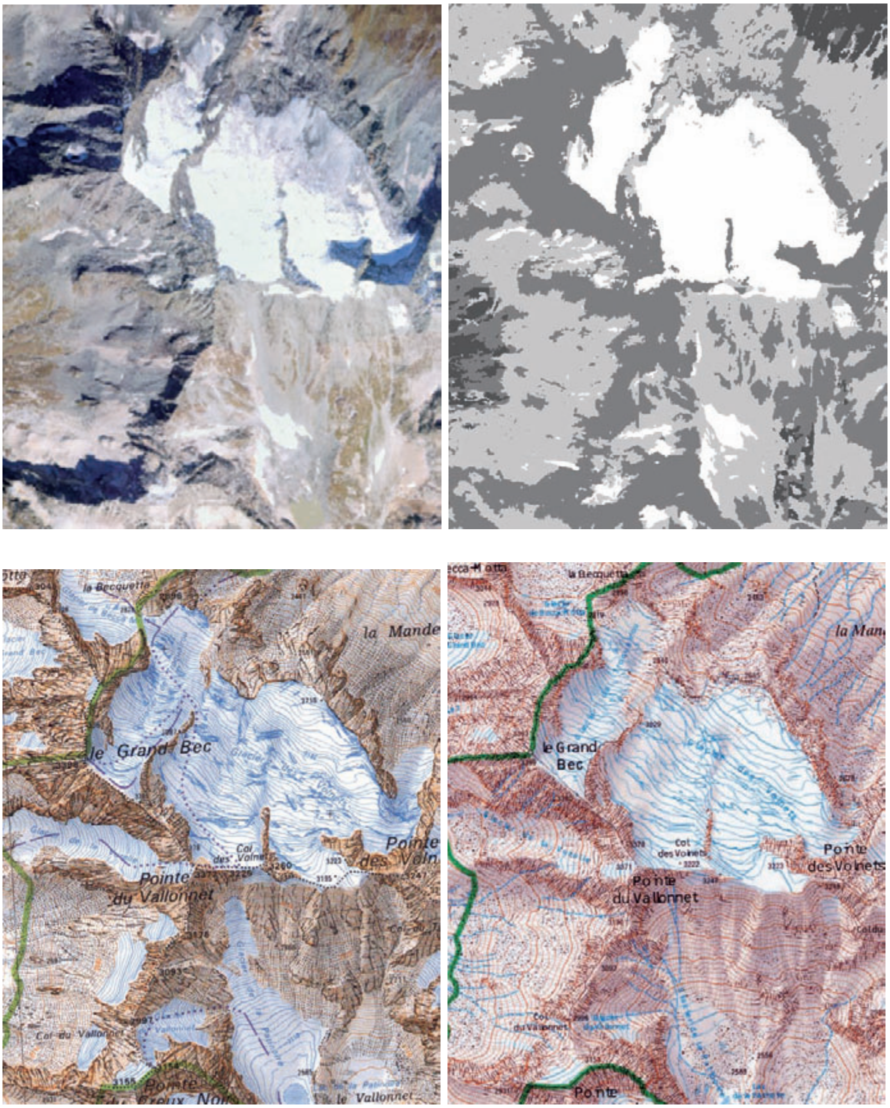


Fig. 8. From left to right: on the first line, crop of the orthoimage near Modane and obtained classification on the same area (from the darkest to the lightest: pasture, rocks, screes and glaciers). On the second line, present handmade topographic map and new fully digital made map on the same area

References

Bossard M, Feranec J, Otahel J (2000) CORINE Land Cover technical guide - addendum 2000. technical report n° 40. Technical report, European Environment Agency

Elhai H (1968) Biogéographie, Paris: Armand Colin

Guigues L (2004) Modèles multi-échelle pour la segmentation d'images, Cergy-Pontoise: Ecole doctorale Sciences et Ingénierie de l'Université de Cergy-Pontoise

Guigues L, Coquerez J-P, Le Men H (2006) Scale sets image analysis. International Journal of Computer Vision, n°68(3)

Hurni L, Dahinden T, Huztler E (2001) Digital Topographic Drawing for Topographic Maps: Tradional Representations by Means of new Technologies. International Publications on Cartography, vol.38, Zurich: Institute of Cartography, Swiss Federal Institute of Technology

Imhof E (1982) Cartographic Relief Presentation, W. De Gruyter

Jenny B (new release, 2008) Automatic scree representation for topographic maps, Zürich: Institute of Cartography, Swiss Federal Institute of Technology

Lacambre A (2001) Aléas et risques naturels en milieu montagnard; apport et limite d'un système d'information géographique. PhD. thesis, Paris: Université Paris 4

Le Bris A, Boldo D (2007) Extraction of landcover themes out of aerial orthoimages in mountainous areas using external information. Proc. of the ISPRS Conference Photogrammetric Image Analysis (PIA), 19-21 September, 2007, Munich, pp 123-128

Le Men H, Trevisan J, Boldo D (2002) Automatic extraction of landcover themes on digital orthophotos in mountainous area for mapping at 1/25k. Proc. of the ISPRS Commission II, 20-23 August, 2002, Xi'an, China, pp 331-337

Paul F (2003) The new Swiss Glaciers Inventory: application of remote sensing and GIS. Zürich: Zürich University

Schwarz G (1978) Estimating the dimension of a model. The Annals of Statistics n°6, pp 461-464

Trias-Sanz R (2006) Semi-automatic high-resolution rural landcover classification. PhD. thesis, Paris: Université Paris 5

Trias-Sanz R, Boldo D (2005) A high-reliability, high resolution rural land cover classification into forest and non forest. Proc. of the Scandinavian Conference on Image Analysis (SCIA), Lecture notes on computer science, vol. 3540, Finland: Springer, Joensuu, pp 831-840

Wang Z, Ziou D, Armenakis C (2003) Combination of imagery - a study on various methods. Proc. of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Toulouse, France

Slope Accuracy and Path Planning on Compressed Terrain

W. Randolph Franklin[1], Daniel M Tracy[1], Marcus A Andrade[12], Jonathan Muckell[1], Metin Inanc[1], Zhongyi Xie[1], Barbara M Cutler[1]

[1] Rensselaer Polytechnic Institute
Troy, New York, 12180–3590, USA
frankwr@rpi.edu – http://wrfranklin.org/, tracyd@rpi.edu, marcus.ufv@gmail.com, muckej@rpi.edu, inancm@rpi.edu, xiez@rpi.edu, cutler@cs.rpi.edu

[2] DPI - UF Viçosa - Brazil

Abstract

We report on variants of the ODETLAP lossy terrain compression method where the reconstructed terrain has accurate slope as well as elevation. Slope is important for applications such as mobility, visibility and hydrology. One variant involves selecting a regular grid of points instead of selecting the most important points, requiring more points but which take less space. Another variant adds a new type of equation to the overdetermined system to force the slope of the reconstructed surface to be close to the original surface's slope. Tests on six datasets with elevation ranges from 505m to 1040m, compressed at ratios from 146:1 to 1046:1 relative to the original binary file size, showed RMS elevation errors of 10m and slope errors of 3 to 10 degrees. The reconstructed terrain also supports planning optimal paths that avoid observers' viewsheds. Paths planned on the reconstructed terrain were only 5% to 20% more expensive than paths planned on the original terrain. Tradeoffs between compressed data size and output accuracy are possible. Therefore storing terrain data on portable devices or transmitting over slow links and then using it in applications is more feasible.

Keywords: terrain compression, slope accuracy, path planning, ODETLAP.

1 Introduction

As ever larger quantities of higher resolution terrain data become available, such as using IFSAR and LIDAR, more efficient compression techniques become more important. This is especially true when it is desired to store the data on portable devices or to transmit the data over slow links. High-resolution data may also compress differently when it is qualitatively different from the older data produced by interpolating contour maps derived from aerial photographs, since the latter are often artificially smooth.

Compression may be either *lossless*, where the restored data is identical to the original data, or *lossy*, where an error is introduced. This choice is not unique to terrain; audio data is also usually compressed lossily. Lossy compression is appropriate when the increased efficiency (i.e., the decreased size of the resulting file) is worth it, or when the original data is imperfect. That is, if the original data has an RMS error of 5 meters (m), then a compression algorithm introducing an average error of $0.5m$ is overkill.

The desired application for the terrain data influences the appropriate metric for evaluating the compression. The easy metric is RMS elevation error, Franklin and Said (1996). However, some parts of the terrain may be more important than others. For example, sharp points in the profile along the skyline are what viewers recognize. This author has had the experience of looking at a mountain range on the horizon while simultaneously looking at a commercial rendition of that same scene, and being unable to correlate the real world with the computer model. The problem resides in the computer model's lack of high spatial frequencies. This may be caused by using calculus tools such as Fourier or Taylor series that assume that the terrain is differentiable many times, and that high frequencies are less important than low frequencies. Both assumptions are false. Not only does nothing in the physics of terrain formation select for smoothness, but rather the reverse. Erosion causes undercutting and slumping leading to cliffs, that is, elevation discontinuities.

Slope is one terrain property that is important to represent accurately. The slope of terrain influences *mobility* (it is difficult to drive up a cliff), *accessibility by air* (aircraft cannot land on a slope), *hydrology* (steeper slopes erode more quickly) and *visibility* (changes in slope are recognizable, and observers sited on a break in the slope may be able to see more).

Slope is often ignored because the assumption is that it comes for free once the elevations are represented sufficiently accurately. However, differencing any imprecise function amplifies the errors. Also, from math analysis we know that approximating a function $f(x)$ more accurately, i.e., $\limsup_{i\to\infty}|(f_i(x)-f(x))| \to 0$, gives no guarantees about $\limsup_{i\to\infty}|(f_i'(x)-f'(x))|$, which may increase without bound. Indeed, it was such paradoxes that motivated the formalization of calculus in the 19th century.

The compression methods introduced here are extensions of ODETLAP, Franklin et al. (2007), and summarized in Figure 1. Briefly, ODETLAP solves a sparse overdetermined system of linear equations for the elevations z_{ij} in an array where a few points' elevations h_{ij} are known. Each known point has an equation

$$z_{ij} = h_{ij} \tag{1}$$

Every non-border point, known or not has an equation

$$4z_{ij} = z_{i-1,j} + z_{i+1,j} + z_{i,j-1} + z_{i,j+1} \tag{2}$$

Border points form a messy special case of no deep theoretical interest, but with the following practical difficulties. Not including equations for border points may lead to the system being underdetermined. A careless choice of equation may bias the surface to being horizontal at the borders, without physical justification.

Since there are more equations than unknowns, the system is overdetermined; we solve for a best fit. The two classes of equations are weighted differently, depending on the relative importance of accuracy versus smoothness. Weighting Equation 1 more highly relatively to Equation 2 makes the resulting surface more accurate but less smooth. A small degree of inaccuracy enables a large degree of smoothness. Indeed, a design requirement for ODETLAP was that, when interpolating between contour lines, that the contour lines not be visible in the resulting surface. Also, broken contours and even isolated points may be processed. These desirable properties are not always shared by competing surface fitting techniques.

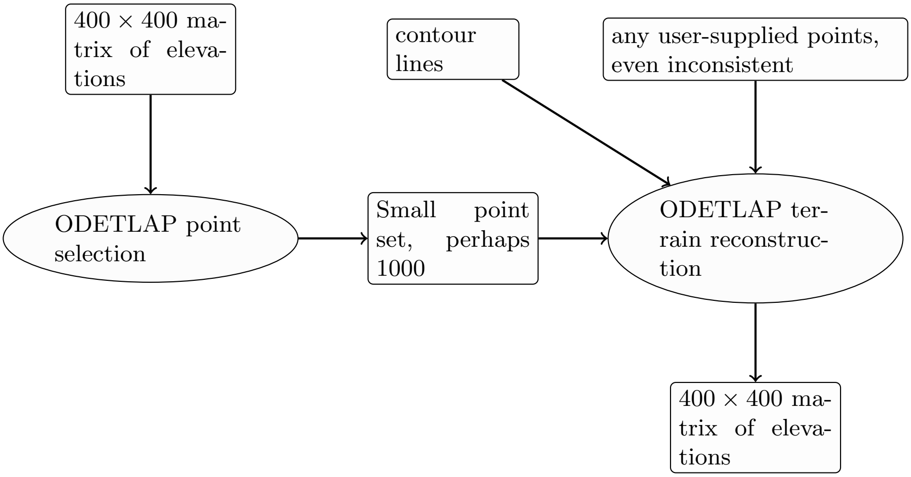
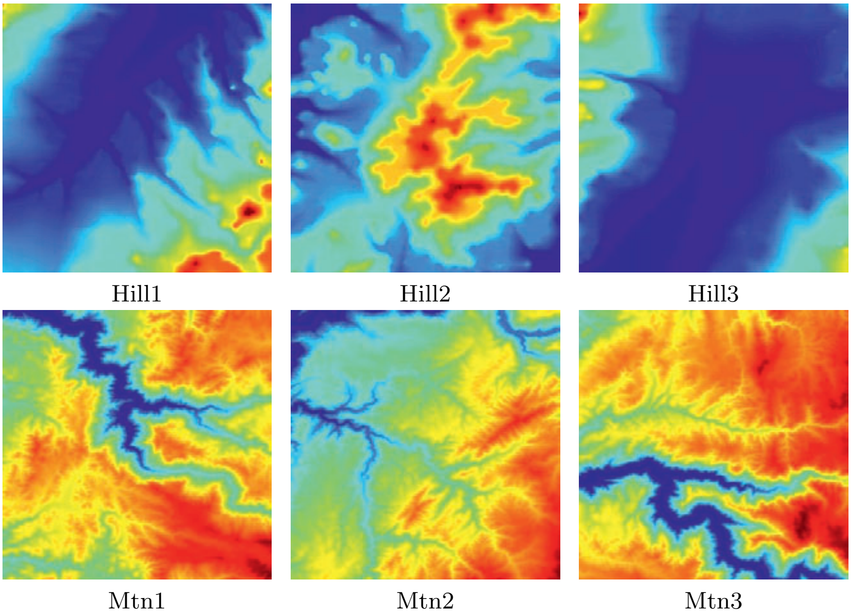

400 × 400 matrix of elevations
contour lines
any user-supplied points, even inconsistent
ODETLAP point selection
Small point set, perhaps 1000
ODETLAP terrain reconstruction
400 × 400 matrix of elevations

Fig. 1. ODETLAP Process

There is little prior art on compressing slopes, apart from some descriptions of fundamental limits. A resolution of 25m or lower cannot identify steep slopes correctly Kienzle (2004). A resolution of 30m with elevations in meters results in a precision of slope calculations no better than 1.9°, Hunter and Goodchild (1997).



Hill1 Hill2 Hill3

Mtn1 Mtn2 Mtn3

Fig. 2. Sample level-II Datasets

Table 1. ODETLAP TIN+Greedy Results

	Hill1	**Hill2**	**Hill3**	**Mtn1**	**Mtn2**	**Mtn3**
Elevation range	505m	745m	500m	1040m	953m	788m
Original size	320KB	320KB	320KB	320KB	320KB	320KB
Compressed size	2984B	5358B	1739B	9744B	9670B	9895B
Compression ratio	107:1	60:1	184:1	33:1	33:1	32:1
(Compressed/Orig size), %.	1.68%	1.33%	1.66%	0.91%	1%	1.23%
# pts selected	1040	2080	520	4160	4160	4160
RMS elevation error	8.49m	9.93m	8.31m	9.48m	9.55m	9.68m
RMS slope error	2.81°	5°	1.65°	8.34°	8.36°	7.87°

2 Terrain Data Structures

The underlying terrain data structure for the research presented in the paper is a matrix or array of elevations. There are other possibilities. One alternative would be high-order spherical harmonics as used in geopotential modeling. However, they are not as applicable to terrain, if only because their complexity grows quadratically with their accuracy. Wavelets of various types are used somewhat, and may become more popular in the future. The major alternative to an array of elevations is a Triangulated Irregular Network (TIN). Franklin (1973) did the first implementation (under the direction of Douglas and Peucker) of the TIN in Geographic Information Science. In the next section we use an updated version of that program, Franklin (2001). In contrast to Isenburg et al. (2006), Franklin (2001) operates incrementally, in the spirit of the Douglas-Peucker line generalization algorithm, Douglas and Peucker (1973) (independently discovered by Freeman and Ramer, Ramer (1972)). In each iteration, it greedily inserts the point that is farthest from the current surface. It can process arrays of up to $10^4 \times 10^4$ points in core. The time to completely TIN a level-I DEM with 1201^2 points (until the max error is under 0.5m) is under 30 CPU seconds on a laptop. Also in contrast to Isenburg, it imposes no restrictions on the size of the generated triangles. However, because it operates out of core, Isenburg can process much larger datasets.

One disadvantage of a TIN compared to an array is the increased complexity of storing the data compactly, since in a naive implementation, most of the storage will be devoted to the topology. Also, rendering the terrain without producing a triangular appearance can require either very many triangles or a smoothing operator. Finally, representing slope accurately, one topic of this paper, appears problematic with a TIN. On the other hand,, unlike an array a TIN is not tied to a particular coordinate system and can better represent large regions of the earth.

3 ODETLAP TIN+Greedy

The first question is, how well does ODETLAP represent slopes? Slope is qualitatively somewhat different from elevation: its autocorrelation distance is smaller, but it requires fewer significant bits.

We used six 400 × 400 test datasets, three hilly and three mountainous, extracted from level-2 DEMs. 400 × 400 is a resolution that we can easily process using the default sparse linear equation solver in Matlab; larger resolutions are possible with other techniques, such as the Paige-Saunders method used by Childs (2003, 2007). ODETLAP TIN+Greedy, the basic version of ODETLAP, selects points with the following two stage process.

1. Use our incremental triangulated irregular network (TIN) program to select $\mathcal{P}$, an initial set of important points.
2. Fit a surface $\mathcal{S}$ to $\mathcal{P}$.
3. If $\mathcal{S}$ is sufficiently accurate then stop.
4. Otherwise, find the 10 to 30 points of the original 400×400 points that are farthest from $\mathcal{S}$. When forming this batch of points to insert, we assume that very close points are redundant, and require points to be at least a couple of pixels apart. Increasing this *forbidden zone* beyond that confers no additional advantage. Points are inserted in batches because of the time to recompute the surface in step 2.
5. Insert the new points into $\mathcal{P}$.
6. Go back to step 2.

The (x, y) are compressed by forming a 400 × 400 bitmap showing the points' locations, then compressing it with a runlength code. The resulting size is not much worse that the information-theoretic limit. The z are compressed with various methods such as *bzip2*.

ODETLAP's running time depends greatly on the input elevation matrix's sparsity. Basic ODETLAP on a 400 × 400 terrain took about 4.6 minutes when 5.4% of the elevations were known, and about 7.5 minutes on a 2.2GHz processor when 7.5% of the elevations were known. Denser input matrices required over 15 minutes.

Table 1 summarizes the results. ODETLAP TIN+Greedy compressed these terrains by factors ranging from 30:1 to 100:1 compared to the original binary file, with RMS elevation errors less than 10m and a slope error ranging from 1.7°to 8.4°, depending on the terrains' ruggedness. The next question is, what is the tradeoff of size versus accuracy? Figures 3 and 4 answer this.

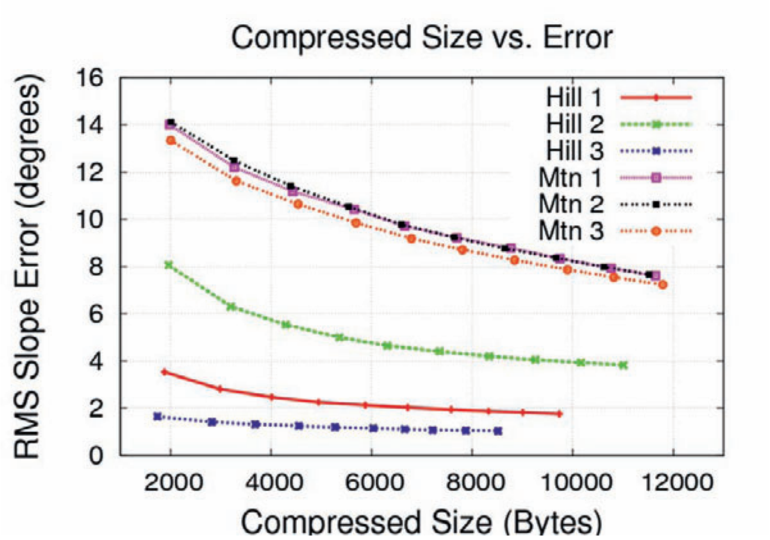


Fig. 3. ODETLAP Tin+Greedy Size – Elevation Accuracy Tradeoff

A major advantage of ODETLAP TIN+Greedy is that it selects the points in order of importance, and so permits progressive transmission of the points. However there will be a size penalty since compressing points incrementally is less efficient than compressing them in one set. Indeed, the former method stores the order of the points, which the latter does not. Therefore, for N points, the penalty will be at least $N \lg N$ bits (the information content of selecting one permutation from $N!$ permutations), but will probably be more. A larger storage cost of this method compared to the following one is caused by these points' positions being irregular.

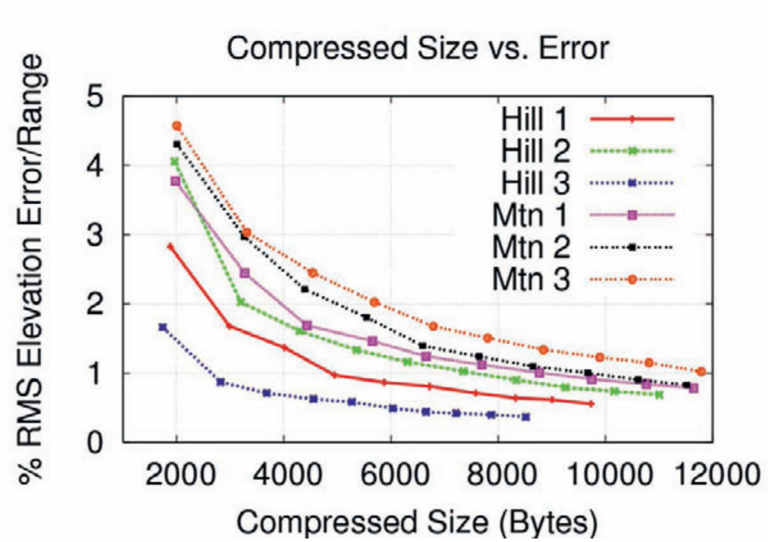


Fig. 4. ODETLAP Tin+Greedy Size – Slope Accuracy Tradeoff

Table 2. ODETLAP Regular grid w/o DCT Results

	Hill1	**Hill2**	**Hill3**	**Mtn1**	**Mtn2**	**Mtn3**
Elevation range	505m	745m	500m	1040m	953m	788m
Original binary size	320KB	320KB	320KB	320KB	320KB	320KB
Compressed size	619B	1591B	315B	4710B	4659B	4777B
Compression ratio	517:1	201:1	1016:1	68:1	68:1	67:1
# pts selected	529	1369	256	4489	4489	4489
RMS elevation error	8.4m	9.2m	9.1m	9.1m	8.9m	8.8m
RMS slope error	4.2°	6.5°	3.0°	9.9°	9.9°	9.4°

4 ODETLAP-Regular Grid

With this alternative, instead of greedily selecting the N most important points, we select points on a regular grid uniformly spaced, say 40×40, or every $10th$ point in x and y. The first advantage is that the points' locations (x, y) do not need to be stored. Second, since the z form a regular array, using any image processing compression technique becomes easy. However, since ODETLAP-Regular grid does not adapt to changes in the spatial complexity of the terrain, it will require more points and it may miss small features. Is this tradeoff worth it?

Table 2 shows the results. For each dataset, the number of points was increased, keeping a square grid of points but selecting more points equally spaced in columns and rows, until the RMS elevation error was under $10m$. For the same number of points, the compressed size varied slightly because

Table 3. ODETLAP Regular grid with DCT Results

	Hill1	**Hill2**	**Hill3**	**Mtn1**	**Mtn2**	**Mtn3**
Elevation range	505m	745m	500m	1040m	953m	788m
Original binary size	320KB	320KB	320KB	320KB	320KB	320KB
Compressed size	306B	807B	172B	2194B	2027B	2013B
Compression ratio	1046:1	397:1	1860:1	146:1	158:1	159:1
# pts selected	529	1600	225	4489	4489	4489
RMS elevation error	9.6m	10.0m	9.7m	9.7m	10.0m	9.9m
RMS slope error	4.3°	6.5°	3.0°	10.°	10.°	9.9°

different sets of z compress differently. After achieving an RMS elevation error smaller than 10, the z coordinate of the selected points are compressed using *bzip2*. Comparing with the ODETLAP TIN+Greedy results, the compression ratio is about 2 times better. On the other hand, the RMS slope error is a little worse.

As an extension, we lossy compressed z as follows. The selected z values were rounded off while preserving an RMS error less than 10 and then transformed with a Discrete Cosine Transform (DCT). A DCT, widely used in image compression, is similar to a Fourier series, but uses a set of higher and higher frequency square waves instead of sines and cosines to approximate a function. The more square waves are used, the more accurate the approximation is, but the more space it takes, Wikipedia (2008).

Then the resulting sequence was compressed using *bzip2*. For a given elevation or slope error, this method compresses better. See Table 3.

5 Path Planning

The next test of our compression algorithm was for *path planning* on terrain, where the traveler is hiding from a set of observers who have been optimally positioned, Franklin and Vogt (2006); Franklin (2002). That is, if we use the compressed terrain to plan a path, how good is that path? We chose the following metric, designed to incorporate several factors affecting real paths.

$$\mathcal{C} = \sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2} \cdot \left(1 + \max\left(0, \frac{\Delta z}{\sqrt{\Delta x^2 + \Delta y^2}}\right)\right) \cdot (1 + 100v) \quad (3)$$

The first term says that shorter paths are better. The second says that moving uphill is expensive. The third term says that being seen by an observer is very expensive ($v = 1$ if the traveler is in sight, 0 otherwise). Note that the uphill term means that this metric is not symmetric; the optimal path from a to b has a different cost, and is not simply the reverse of, the optimal path from b to a. Therefore some other path planning algorithms will fail. Further, since a 400×400 dataset has 400^2 points, graph traversal algorithms

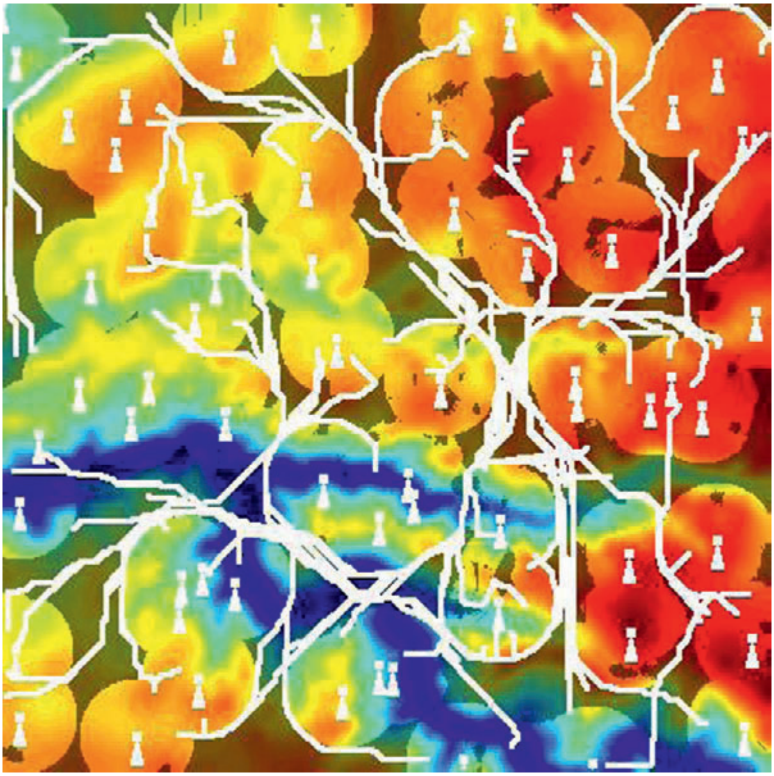


Fig. 5. Many optimal paths avoiding viewsheds on Mtn3

employing an explicit cost matrix are infeasible. Finally, some search strategies that climb hills in parameter space (unrelated to climbing hills on the terrain) stop at local optima, which is undesirable. To address all these concerns, we created a modified A* search procedure, Tracy et al. (2007), and used it to plan paths between many pairs of sources and destinations on each dataset. Figure 5 shows many paths plotted on the *mtn3* dataset. Each little white lighthouse represents an observer. The surrounding colored region is the observer's viewshed. Gaps in the viewsheds are caused by ridges hiding the terrain behind them. The dark regions of the figure are invisible to all the observers. Figure 5 also shows choke points in the terrain, which are traversed by many paths. Those would be candidates for siting future observers.

How to evaluate the path computed on the compressed terrain is also important, and the obvious choices may be wrong. For example, the cost of the path computed on the compressed terrain is meaningless. Indeed, if the terrain were compressed to be flat, then paths computed on it would have no cost for moving uphill and so would be artificially cheap, which is wrong. Even comparing the distance between two paths is meaningless for evaluating them.

Table 4. Increased cost of paths computed on compressed terrain

Data	Compressed size	Compression ratio	Cost increase
Hill1	1763	182	5.5%
Hill2	1819	176	6.1%
Hill3	1607	199	4.4%
Mtn1	1925	166	19.2%
Mtn2	1884	170	18.2%
Mtn3	1946	164	17.0%

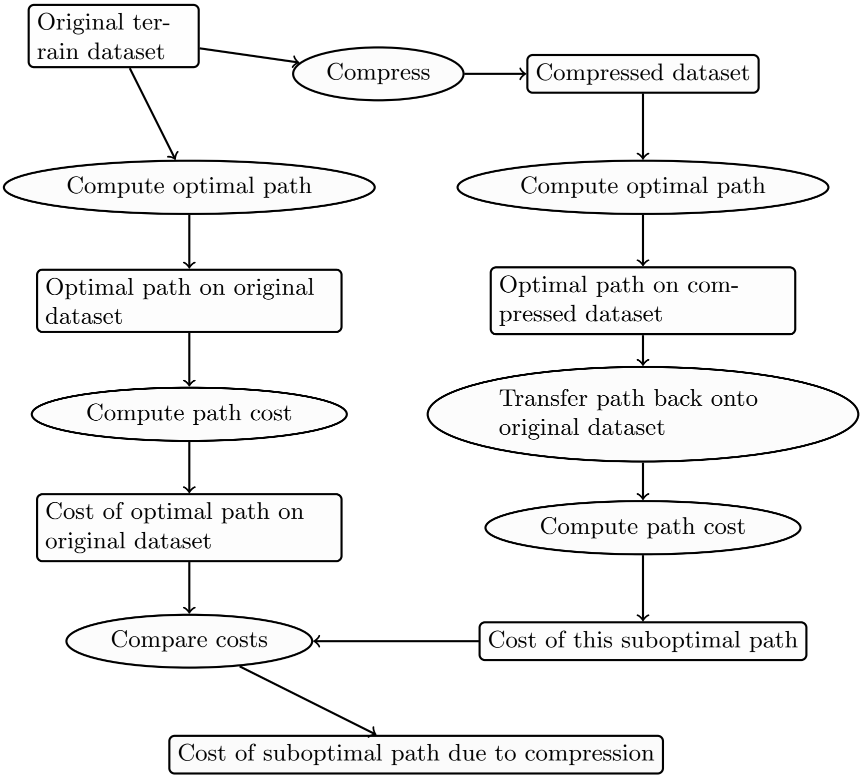

Original terrain dataset
Compress
Compressed dataset
Compute optimal path
Compute optimal path
Optimal path on original dataset
Optimal path on compressed dataset
Compute path cost
Transfer path back onto original dataset
Cost of optimal path on original dataset
Compute path cost
Compare costs
Cost of this suboptimal path
Cost of suboptimal path due to compression

Fig. 6. Compressed path evaluation algorithm

Indeed, two paths may be legitimately quite different but have the same cost; we don't care. Our metric recognizes that the purpose of computing a path on any terrain, compressed or original, is to use it in the real world. Therefore, we transfer the path back to the original terrain dataset, and evaluate it there, as shown in Figure 6.

Table 4 shows the path inefficiency when our six terrains are compressed by factors of at least 164:1. Paths computed on these very compressed terrains were suboptimal by only 6% to 19%.

6 ODETLAP+Slope

With ODETLAP TIN+Greedy, we insert points with the greatest absolute elevation error. Since the goal is to represent slopes accurately, one obvious improvement would be to insert points with large slope errors. Another possibility would be to insert groups of close points since fixing the elevations of a set of close points should also fix the slope in that neighborhood. Both these ideas, and many other experiments not detailed here, had disappointing results. It was time to extend the ODETLAP equations themselves.

Three different representations of the terrain need to be distinguished in order to understand this section.

Original representation This is the original 400 × 400 matrix of elevation posts that we wish to compress.

Compressed representation This compact version is what would be transmitted or stored on portable devices.

Reconstructed terrain The compressed representation would be reconstituted into this new 400 × 400 matrix in order to be used.

For ODETLAP+Slope, we supplement the two existing types of equations, 1 and 2 with a new type of equation designed to force the slope in x and y to be more accurate.

$$z_{i+1,j} - z_{i-1,j} = h_{i+1,j} - h_{i-1,j} \tag{4}$$

$$z_{i,j+1} - z_{i,j-1} = h_{i,j+1} - h_{i,j+1} \tag{5}$$

This sets the Δz between the northern and southern neighbors equal to its known value, and sets the Δz between the western and eastern neighbors equal to its known value. The elevation of the center point is not used. It was done this way because these two Δzs are the values used by the Zevenbergen-Thorne method, Zhou and Liu (2004), a common method for computing slopes, Zevenbergen and Thorne (1987). (The cross product of the two vectors becomes the normal to the surface.) Our system permits the indices to be chosen arbitrarily, to allow for pairs of nonadjacent points to be used; this is a topic of potential future research.

Since the system is overconstrained, the relative weights of the different types of equations can be set depending on the relative importance of slope accuracy, elevation accuracy, or smoothness. The idea for this addition is that the extra freedom of allowing elevations to drift somewhat, provided that the slopes remain accurate, may allow greater slope accuracy.

Either ODETLAP TIN+Greedy or ODETLAP Regular grid may serve as the basis for adding equations 4 and 5. In the former case, we iterate the process of greedily inserting the points whose reconstructed slopes are the worst. ODETLAP TIN+Greedy requires fewer points but ODETLAP Regular grid requires less space to store each of the points on the grid (though any extra irregular points off the grid will take the same space as in ODETLAP TIN+Greedy. As before, we add points in batches for efficiency, and use forbidden zones around the points to prevent close pairs of points to be added in the same iteration, although a point $\mathcal{P}$ added in one iteration may be adjacent to a point added in an earlier iteration, if $\mathcal{P}$'s error is sufficiently large.

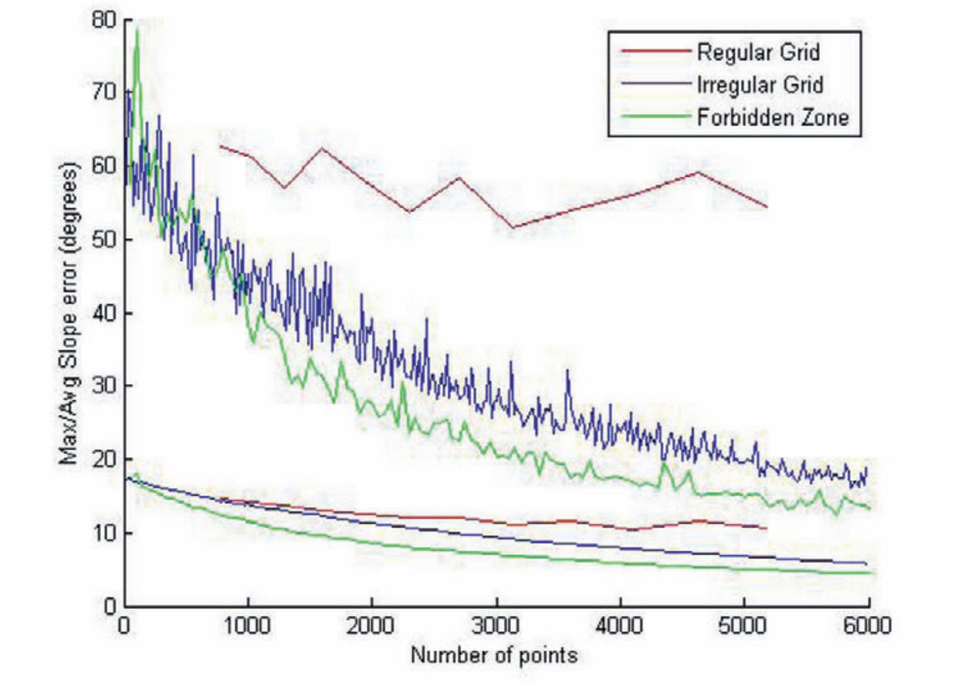


Fig. 7. Slope accuracy vs number of points for Mtn2

Figure 7 shows how three variants of this idea perform on the Mtn2 dataset. They are: selecting points in a regular grid, greedily selecting irregular points, and greedily selecting irregular points using an 11 × 11 forbidden zone. The x-axis is the number of points in the compressed representation (out of a total of 160000 points). The y-axis shows the average and maximum slope errors (the three *max* curves are the higher ones). The best method is greedily selecting irregular points using a forbidden zone.

7 Conclusions and Future Work

Representing terrain, including slope, with ODETLAP has great potential. We are now exploring some of its variations, and applying it to high resolution urban data. The major problem to be addressed is the computation space and time required. We are also extending our path planning algorithm for road construction. Here, we are allowed to modify the terrain with cuts and fills when planning the path. Another application of ODETLAP is terrain smoothing, which might be applied to any other terrain representation. Indeed that ODETLAP was created to smooth or interpolate between contours so that those contours would not be visible in the resulting surface.

One problem with all compression techniques is that they do not preserve *Hydrology*. Regions of the world where the terrain was formed by erosion caused by surface water flow have distinctive properties. There are almost no actual local minima (basins, depressions), because they become lakes. In the few depressions in the coterminous USA, such as the Great Salt Lake, Salton Sea, and Crater Lake, the water either evaporates or percolates away. However, there are many fictitious depressions caused by errors in measuring the terrain or by insufficiently fine sampling, Maidment et al. (1997). That is, the water may exit a depression via a canyon that is so narrow that it fits between two adjacent elevation posts, and so is missed. We are now studying how the hydrological properties of the terrain under compression. This is an instance of the general problem of compressing multiple layers of cartographic data where preserving the relationships between the layers after reconstruction is at least as important as preserving the individual layers' accuracy.

The most general problem is to construct the terrain from a set of mathematical operators that force the resulting terrain to have the desired properties. For instance, suppose that we carved the terrain out of a block of earth with a shovel, with repeated applications of the following operation. Place the shovel touching the earth at a some point. Move the shovel along any trajectory ending at the edge of the earth, provided that the shovel always gets lower and lower. Then, repeat with another shovel path, etc. The terrain that is created will never have an interior local minimum. That is, it will be hydrologically "correct". Can we reduce this idea to practice?

8 Acknowledgement

This research was supported by NSF grants CCR-0306502 and DMS-0327634, by DARPA/DSO under the Geo* program, and by CNPq – the Brazilian Council of Technological and Scientific Development.

References

Childs J (2003) *Development of A Two-Level Iterative Computational Method for Solution of the Franklin Approximation Algorithm for the Interpolation of Large Contour Line Data Sets*. Master's thesis, Rensselaer Polytechnic Institute.

Childs J (2007) Digital elevation modeling journal. `http://terrainmap.com/`, [accessed 28-May-2007].

Douglas DH and Peucker TK (1973) Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. Canadian Cartographer, 10(2):112–122.

Franklin WR (1973) Triangulated Irregular Network Program. `ftp://ftp.cs.rpi.edu/pub/franklin/tin73.tar.gz` [accessed 23 May 2006].

Franklin WR (2001) Triangulated Irregular Network computation. `http://wrfranklin.org/pmwiki/Research/TriangulatedIrregularNetwork`, [accessed 8-June–2007].

Franklin WR (2002) Siting Observers on Terrain. In D Richardson and P van Oosterom, editors, *Advances in Spatial Data Handling: 10th International Symposium on Spatial Data Handling*, pages 109–120. Springer-Verlag.

Franklin WR, Inanc M, Xie Z, Tracy DM, Cutler B, Andrade MVA, and Luk F (2007) Smugglers and Border Guards - The GeoStar Project at RPI. In *15th ACM International Symposium on Advances in Geographic Information Systems (ACM GIS 2007)*. Seattle, WA, USA.

Franklin WR and Said A (1996) Lossy Compression of Elevation Data. In *Seventh International Symposium on Spatial Data Handling*. Delft.

Franklin WR and Vogt C (2006) Tradeoffs when multiple observer siting on large terrain cells. In A Riedl, W Kainz, and G Elmes, editors, *Progress in spatial data handling: 12th international symposium on spatial data handling*. Springer, Vienna.

Hunter G and Goodchild M (1997) Modeling the uncertainty in slope and aspect estimates derived from spatial databases. Geographical Analysis, 29(1):35–49.

Isenburg M, Liu Y, Shewchuk JR, Snoeyink J, and Thirion T (2006) Generating Raster DEM from Mass Points Via TIN Streaming. In *GIScience*, pages 186–198.

Kienzle S (2004) The Effect of DEM Raster Resolution on First Order, Second Order and Compound Terrain Derivatives. Transactions in GIS, 8(1):83–111.

Maidment DR, Olivera F, Reed S, Ye Z, Akmansoy S, and McKinney DC (1997) Water balance of the Niger river basin in West Africa. In *17th Annual ESRI User Conference*. San Diego, CA.

Ramer U (1972) An iterative procedure for the polygonal approximation of plane curves. Computer Graphics and Image Processing, 1:244–256.

Tracy DM, Franklin WR, Cutler B, Andrade MA, Luk FT, Inanc M, and Xie Z (2007) Multiple observer siting and path planning on lossily compressed terrain. In *Proceedings of SPIE Vol. 6697 Advanced Signal Processing Algorithms, Architectures, and Implementations XVII*. International Society for Optical Engineering, paper 6697-16, San Diego CA.

Wikipedia (2008) Discrete cosine transform — Wikipedia, The Free Encyclopedia. `http://en.wikipedia.org/w/index.php?title=Discrete_cosine_transform&oldid=196623065`, [accessed 8-March-2008].

Zevenbergen LW and Thorne CR (1987) Quantitative analysis of land surface topography. Earth Surface Processes and Landforms, 12(1):47–56. doi:http://dx.doi.org/10.1002/esp.3290120107.

Zhou Q and Liu X (2004) Analysis of errors of derived slope and aspect related to DEM data properties. Computers and Geosciences, 30(4):369–378.

Processing 3D Geo-Information for Augmenting Georeferenced and Oriented Photographs with Text Labels

Arnoud De Boer[1], Eduardo Dias[2], Edward Verbree[3]

1 Utrecht University, 3508 TC Utrecht, NL
email: arnouddeboer@cs.uu.nl
2 Geodan S&R, President Kennedylaan 1, 1079 MB Amsterdam (NL)
email: eduardo@geodan.nl
3 Delft University of Technology - Reseach Institute OTB,
Section GIS-technology, 2600 AA Delft, (NL)
email: e.verbree@tudelft.nl

Abstract

Online photo libraries face the problem of organizing their rapidly growing image collections. Fast and reliable image retrieval requires good qualitative captions added to a photo; however, this is considered by photographers as a time-consuming and annoying task. In order to do it in a fully automated way, the process of augmenting a photo with captions or labels starts by identifying the objects that the photo depicts. Previous attempts for a fully automatic process using computer vision technology only proved not to be optimal due to calibration issues. Existing photo annotation tools from GPS or geo-tagging services can only apply generic location information to add textual descriptions about the context and surroundings of the photo, not actually what the photo shows. To be able to exactly describe what is captured on a digital photo, the view orientation is required to exactly identify the captured scene extent and identify the features from existing spatial datasets that are within the extent. Assumption that camera devices with integrated GPS and digital compass will become available in the near future, our research introduces an approach to identify and localize captured objects on a digital photo using this full spatial metadata. It proposes the use of GIS technology and conventional spatial data sets to place a label next to a pictured object at its best possible location.

Keywords: photo annotation, object identification, label placement.

1 Introduction

The increasing availability of consumer digital cameras and integrated all-in-one devices (e.g. camera phones) enables people to capture and upload digital photos at any place and any time. The organization of these rapidly growing image collections is a major challenge for online photo libraries. Good qualitative descriptions of the content added to a photo enable easier retrieval of an image, but unfortunately, captioning photos is experienced by photographers and users as a time-consuming and annoying task (Dias et al. 2007). Those who do not caption their photos encounter problems at a later stage when users are searching for a particular photo. This issue encouraged researchers to develop tools that enable the automatic captioning of digital photos using positioning information – either automatically by a GPS device or manually by georeferencing on a map – to add descriptions about the context and surroundings (Naaman et al. 2004). However, using positioning information only, these state-of-the art photo annotation tools are limited to the adding of descriptions to a photo about its surroundings, and not about the objects that are actually pictured (Chang 2005).

Assuming that in the near future digital cameras will include a GPS chip and digital compass (to capture position and orientation), the work presented here is an approach that extends the captioning of photos and benefits from this full spatial metadata (geographic positioning, altitude, and pitch) in order to produce an abstraction of the captured scene and to identify objects on a photo.

Our process of object identification in digital photos proposes an alternative for computer vision-based image recognition and photogrammetric coordinate conversions (from pixel to terrain coordinate system). Available GIS technology and established spatial data sets are applied to identify what is visible and where it is located on a digital photo by using a perspective viewer service. This tool renders a three-dimensional model based on input view parameters (the full spatial metadata) and outputs a 2D image that is a virtual abstraction corresponding to the pictured scene. Linking the virtual scene to the three-dimensional model, attributes (i.e. street names) from the spatial data sets can be picked and associated with the objects. At this stage, the image can be augmented with captions of the objects. We go one step further: to actually label the objects in the photo with the just determined captions. To do this, constraints and rules are added to a label engine in order to place a label next to a pictured object at its best possible location. This last step of labeling photos can be especially relevant in the accessibility field. It can be the basis for developing new tools used to improve accessibility for visually impaired users to "sense" digital photos using large-sized label fonts or sounds on mouse over, as objects on the photo are well-identified and well-localized.

This paper is organized as follows: Section 2 describes previous research on automatic photo annotation tools and label placement in 3D environments. Section 3 defines the collection of digital photos having full spatial metadata and the spatial data requirements for the preparation of the extrusion models. Section 4 describes our approach for object identification in digital photos. Section 5 proposes some rules that could be applied in order to find the best location to place a label on a digital photo. Finally, Section 6 provides some discussion and conclusions and recommends future research.

2 Related Research

Cartography is described as the graphic principles supporting the art, science, and techniques used in map making maps, which are intended to communicate a certain message to the user. The process of text insertion in maps is referred to as label placement. Label placement is one of the most difficult tasks in automated cartography (Yamamoto and Lorena, 2005). Positioning text requires that:

- overlap among texts is avoided;
- cartographic conventions and preferences is obeyed;
- unambiguous association is achieved between each text and its corresponding feature;
- a high level of harmony and quality is achieved.

Good placement of labels avoids as much as possible overlap of labels with objects and mutual labels; and is applied to provide additional information about a particular feature. Automatic label placement is therefore one of the most challenging problems in GIS (Li et al., 1998):

- optimal labeling algorithms are very computational expensive for interactive systems;
- labels compete with data objects for the same limited space.

Augmented reality (AR), considered to be part of the Multimedia Cartography research field, is an environment that includes both virtual reality and real-world elements and forms part of the research. AR is a field of computer research which deals with the combination of real world and computer generated data. Its principle is to overlay the real world (captured by a camera device) with supplementary information (e.g. labels). It enables users to interact with their environment e.g. by hyperlinking labels inside an AR view (Cartwright et al., 2007). Interactivity is one of the key components of multimedia.

Photo labeling refers to the act of placing labels that describe the features visible on the photograph itself. Saving the labels obtained from the virtual scene to a transparent layer enables to put labels associated with an object onto an image. As such, the photo annotation issue is considered to be part of Multimedia Cartography and AR as well; visible tags in images and AR applications enable user interaction with the environment; numerous ubiquitous and/or augmented reality applications are discussed by Kolbe (2004), Toye et al. (2006) and Schmalstieg and Reimayr (2007).

As Li et al. (1998) observe, object and label placement in limited screen spaces is a challenging problem in information visualization systems. Images also have a limited screen space and therefore (particularly automatic) label placement is of concern for this research in order to avoid overlap of labels mutual and labels with objects.

Numerous researchers already examined the problem of automatic label placement in 2D maps. Recent work of Maass and Döllner (2006a), Azuma (2004) and Götzelman et al. (2006) also focused on the placement of labels in 3D landscapes and Augmented Reality views referred to as view management (Bell et al. 2001). Götzelman et al. (2006) offer complex label layouts which integrates internal and external labels of arbitrary size and shape, and real-time algorithms. Maass and Döllner (2006b) describe two point-feature dynamic annotation placement strategies for virtual landscapes including priority aspects.

Labeling is further divided into internal and external (object) annotation (Maass and Döllner 2006a). An internal annotation is drawn on the visual representation of the referenced object and partially obscures that object. An external annotation is drawn outside the visual representation of the reference object and uses a connecting element such as a line or arc to associate the annotation with the reference object.

Hagedorn et al. (2007) describe the use of a Web Perspective Viewer Service (WPVS) for the annotation of three-dimensional geographical environments (a.k.a. geo-environments). Furthermore, a three-dimensional Web View Annotation Service (3D WVAS) is proposed as an extension to a WPVS. The perspective view together with a depth image is forwarded to the 3D WVAS together with annotation definitions. This annotation technique calculates the positions of the labels, renders them into a separate image buffer, and combines the resulting image in a depth-sensitive way with the input color image (see Fig.1.).

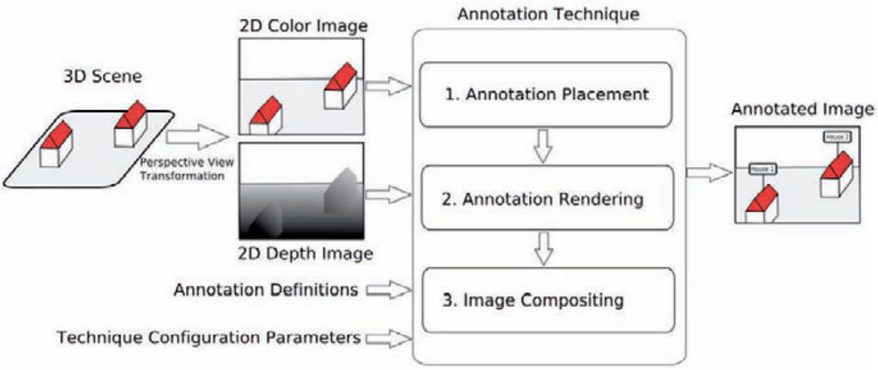


Fig. 1. Process for the annotation of 3D scenes as proposed by Hagedorn et al. (2007)

Our work links up with the previous research in chaining a Perspective Viewer Service with an Annotation Service, although the approach is more simplistic in the sense that we have chosen to use components from a commercial GIS package (i.e. ArcScene and Maplex by ESRI ArcGIS) to demonstrate the concept of internal annotation within digital photos. Our label placing strategy, concentrates in:

- linking the labels to the object they refer to;
- determining the ‘free’ labeling space, i.e. open sky;
- placing the labels at the best possible location.

3 Data collection and preparation

3.1 Image collection

Our concept and ideas were tested by collecting three-dimensional geo-referenced and oriented digital photo at the Market square in the historic city centre of Delft, the Netherlands. We created two collections of test photos:

1. Low-resolution and high-spatial accuracy photos captured using a Topcon GPT-7003i © imaging total station, and

2. High-resolution and low-spatial accuracy photos were captured using a Nikon D-100© digital camera mounted with a 3-axis electromagnetic digital compass and a GPS data logger (see Fig.2.)

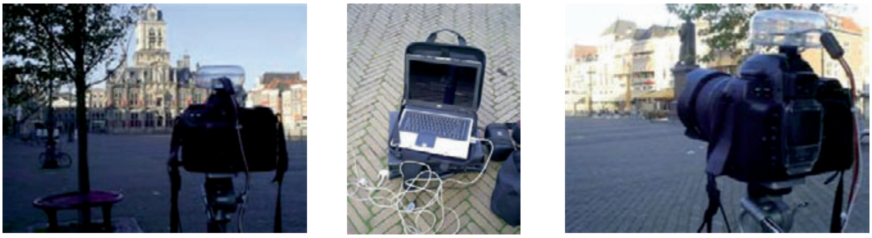


Fig. 2. Collage of the image collection on the Market square in Delft using the Nikon D100 camera mounted with digital compass on the hot shoe cover

The Topcon GPT-7003i distance and direction measurements are connected to a Large Scale Base Map of the Netherlands (GBKN) resulting in a position accuracy of approximately 0.5m and a directional accuracy of approximately 0.5 degrees. The camera included with the Topcon station has the disadvantage of low resolution (0.3 megapixels). On the other hand, the photos captured with the Nikon camera have high resolution (10 megapixels) but low spatial resolution. The position accuracy is around 10 meters and the compass orientation is very inaccurate due to distortion in the compass heading caused by the electromagnetic field of the camera. However, these images are particularly used to identify the misidentification of objects due to lens distortions, GPS and compass inaccuracies.

3.2 Spatial data requirements

A three-dimensional building model is required to serve as input for creating the virtual scene using a perspective viewer service (for our research we used the 3D visualization tool: ESRI ArcScene©). The three-dimensional building model was created by extruding 2D building footprints, extracted from a 1:10,000 topographic base map (the TOP10 vector dataset of the Dutch National Mapping Agency), based on the height values from a raster detailed elevation model (from an Airborne Laser Altimetry dataset, the AHN). This approach results in a gridded footprint in which each feature is associated with an object identifier (OID) from the building footprint dataset and a height value from the elevation raster. The advantage of this approach is that height values inside the buildings are known and the building footprint geometry is preserved. After extrusion of the features and randomly coloring based on the building OID, the model of Fig.3. is obtained.

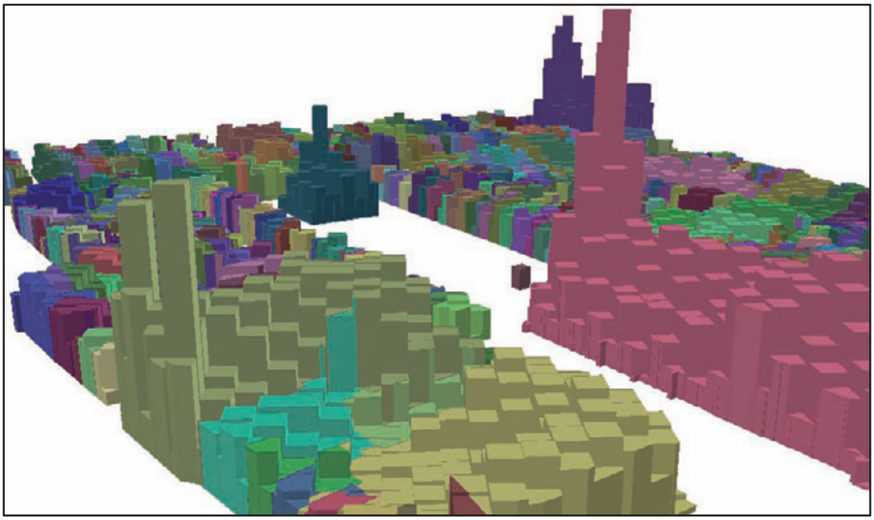


Fig. 3. 3D model created from intersecting the building footprints with the vectorized elevation raster.

Even though we believe that such a 3D model is not interesting in aesthetics terms, it was considered to be the optimal solution for our research, because it is simple to produce and reproduces the shapes of the buildings with sufficient accuracy. This was the model applied as input to create the perspective views representing the digital photo. At this step, the names of buildings and shops located around the Market square are added to the dataset based on a commercial “Points of Interest” dataset. These will be the names to pick after the objects are identified.

4 Object identification

The core of our research is dedicated to the object identification problem: “What is captured by a digital photo?”, and “Where is it located in the photo?” The latter question is very important in order to place a label next to an object. Nevertheless, knowing where the objects are located in a photo enables other innovative applications besides labeling, for example: hyperlinking the objects in the photos to dynamic descriptive pages or tools that help visually impaired users to understand image content by using sounds (for legally blind users) or large sized text labels (for users with low vision) on mouse-over.

As a perspective viewer service, we applied ESRI ArcScene to render our three-dimensional model based upon the view parameters (the full spatial metadata) to create a virtual abstraction that matches the pictured scene. The main issue to be solved is how to link the virtual scene to the three-dimensional model in order to pick the names. Since the virtual abstraction is returned in raster format, its coordinates are in a local pixel coordinate system so a spatial join with the three-dimensional model is not possible. The solution to this problem as proposed in our work is to color each object of the three-dimensional model based on its unique OID. Therefore, the decimal OIDs are converted to RGB color values using the relationship: OID ≡ RGBdecimal = 65536 * Red + 256 * Green + Blue

Fig. 4b shows the output of the building features from the three-dimensional model (Fig.4a) when colored with RGB color values corresponding to their OID. Subsequently, a virtual scene (Fig.4c) corresponding to the digital photo is created by using the view parameters derived from the full spatial metadata of the same digital photo. This virtual scene has to be exported to a lossless compression image format (e.g. PNG) in order to maintain a seamless color throughout the object and, in this way, avoiding additional features to intrude in the form of averaged colored pixels on the object borders.

Now we have a raster image with representation of the visible features. To further analyze this result, it was necessary to convert the virtual scene to vector features again. Therefore, first the RGB color bands of the virtual scene are summed up using the RGB-OID relationship to obtain OIDs again. Next, the vectorized virtual scene (Fig.4d) is joined with the building features of the three-dimensional model (or 2D spatial datasets) based on the OID and names are picked from these datasets (Fig.4e) to label the objects visible on the digital photo.

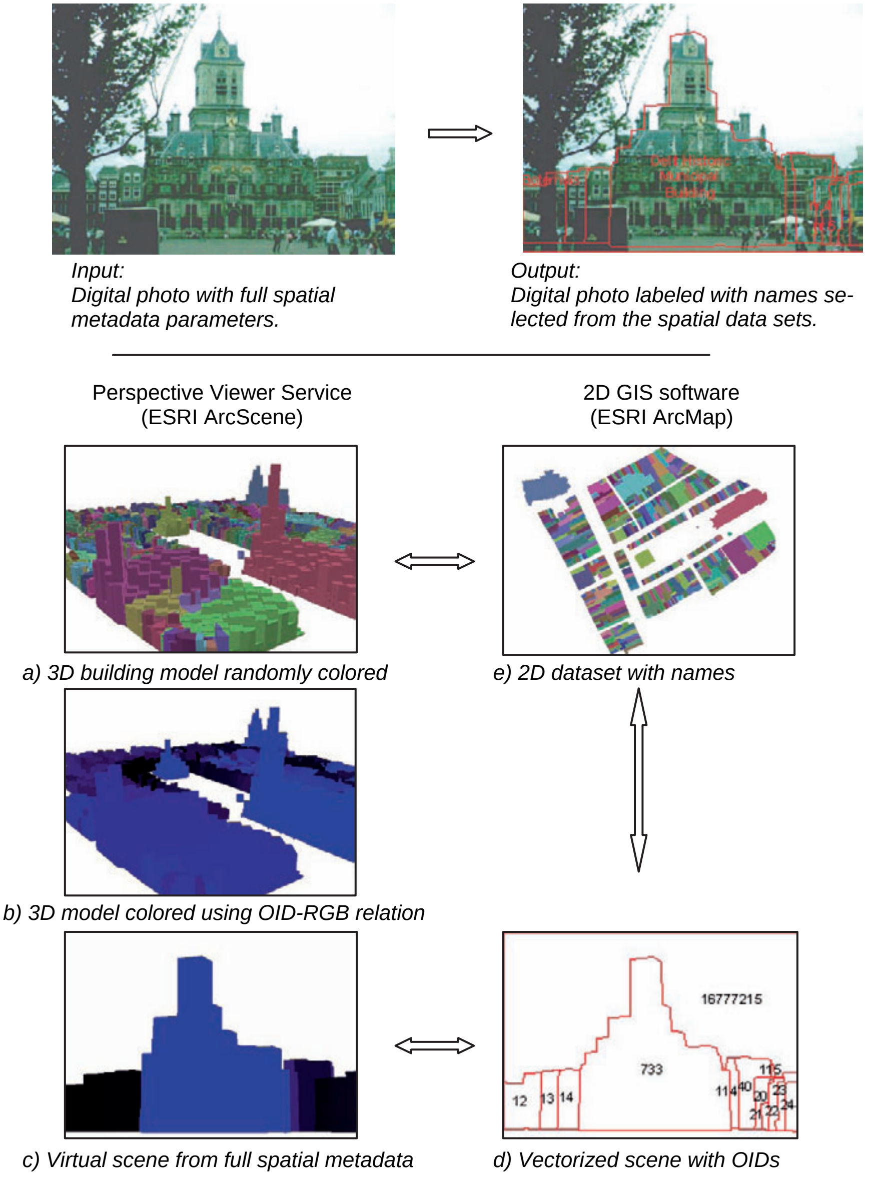

Input:
Digital photo with full spatial metadata parameters.
Output:
Digital photo labeled with names selected from the spatial data sets.
Perspective Viewer Service
(ESRI ArcScene)
2D GIS software
(ESRI ArcMap)
a) 3D building model randomly colored
e) 2D dataset with names
b) 3D model colored using OID-RGB relation
c) Virtual scene from full spatial metadata
d) Vectorized scene with OIDs
16777215
733
12
13
14
114
40
115
20
23
21
22
24

Fig. 4. Concept and process of object identification using the OID-RGB relation.

5 Label Placement

After the object identification, in which pictured objects are identified and localized on a digital photo, the next part of our research focused on label placement: proposing constraints and rules on where to place a label on a digital photo to identify a certain object. We assumed that the best location for label is at an empty area, which is defined as the area where there are no objects inside the virtual scene (or digital photo). Using ESRI ArcMap©, the virtual scene is overlaid with the digital photo and labeled. The labeling was optimized by using constraints and rules on the label engine extension Maplex©, including, among others, to avoid overlap between labels and between labels with objects, i.e. labels are placed outside visible objects using connectors (see Fig.5)

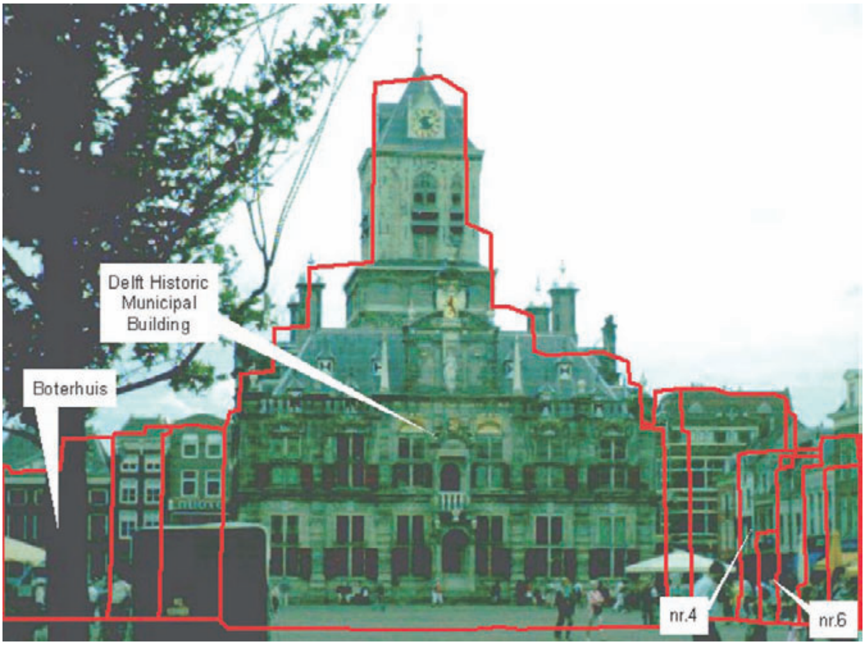


Fig. 5. Visible objects are externally annotated using connectors avoiding as much as possible overlap of labels with objects and among labels.

However, because the tree in front of our pictured scene is not included in our 3D building model, the label engine assumes that this location is a good location to place a label. But, since we do not wish to overlay objects in the picture, we needed to identify in the picture, the areas without any features. Therefore, the digital photo is reclassified to a binary image. In our example (see Fig.6 upper left), we used the median as a cut-off value. After this, we combined the binary image derived from the virtual scene with the building model. This limits the label engine to avoid placing a label overlapping objects in the united layer. The result is shown in Fig.6.

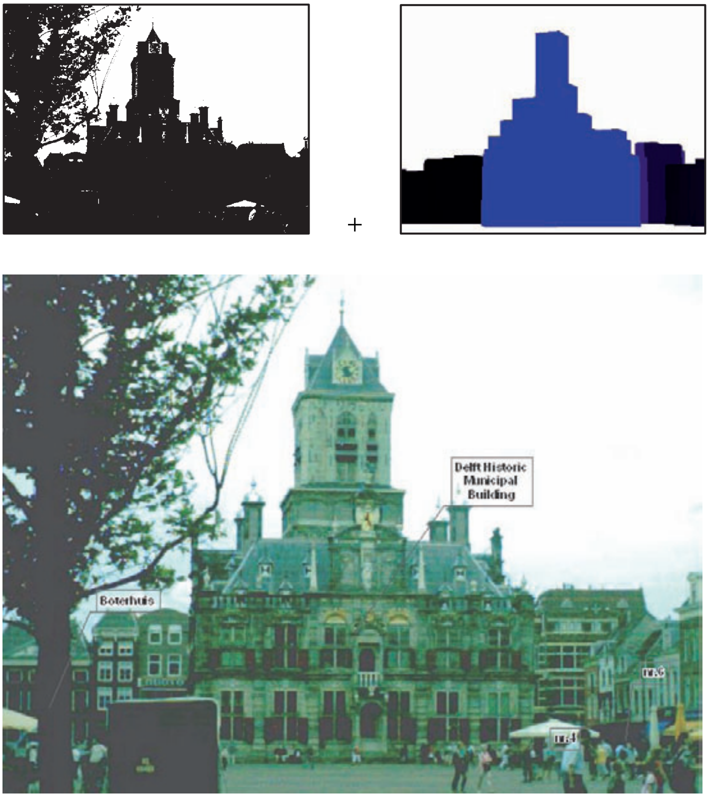


Fig. 6. A binary image and the virtual scene are used to place a label at its best possible location (assumed to be at the empty areas).

Our second proposal is to apply a depth image (a.k.a. depth map) to vary label font sizes with the distance from the observer to the objects, maintaining the perspective view. A depth map is created (and returned) by the renderer to identify what is visible or not to build the two-dimensional abstraction of the 3D model. We created our depth (see Fig. 7 on the left). image by:

1. calculating the distance from observer to an object;
2. updating these as an attribute to the building model;
3. coloring the three-dimensional model based on the distance-attribute;
4. creating and exporting a virtual scene using the view parameters.

Finally, the digital photo is labeled using the depth image showing the varying label font sizes of Fig.7.(right hand picture).

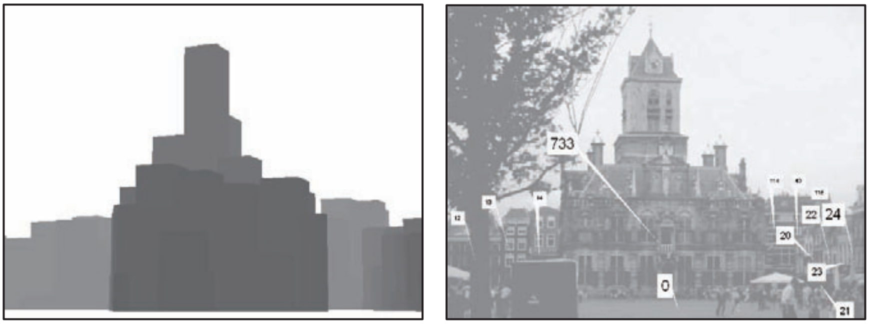


Fig. 7. Depth images as output of the perspective viewer services (left) used to vary in label font sizes depending on the distance from observer to pictured object (right).

The amount of labels that should be placed on a digital photo depends on the number of visible objects and user preferences. By adding a constraint to the label engine that only objects of a minimum specified size should be labeled (e.g. based on polygon perimeter), the amount of labels that are placed on a photo could be reduced. Fig. 8 shows an example of a digital photo labeled with names of shops located around the Market square of Delft; on the left hand side all visible objects are labeled; on the right hand side only visible object with a polygon perimeter larger than 640 m^2 are labeled.

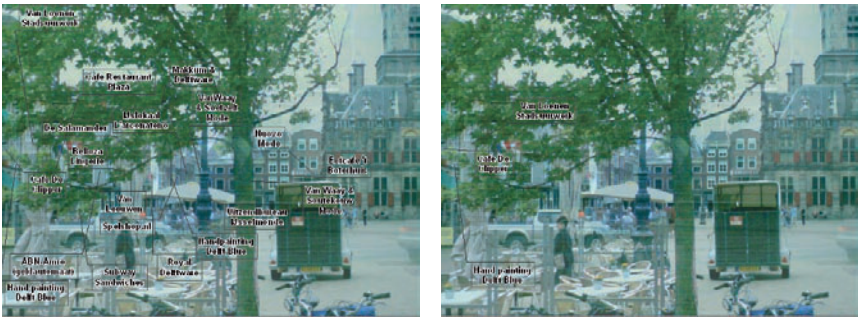


Fig. 8. Varying the amount of labels to place: on the left hand side no constraints are added with respect to object size; on the right hand side the larger visible objects are labeled.

6. Results and conclusions

This research showed how objects on a digital photo can be identified using the photo's full spatial metadata (with position and orientation). In addition, we investigated the best location for a label to annotate an object within the photo. The results of the object identification for the photo collection with high-quality spatial metadata (acquired with the Topcon imaging total station) were very positive. There was a good match between the 3D building abstractions and the photos. The results for the photo collection with lower spatial accuracy (acquired with the Nikon camera connected to a GPS receiver and a digital compass) revealed less successful results than the Topcon, directly related with the inaccuracies of the GPS and compass devices. As expected, it was also observed that the amount of misidentification increases with increasing inaccuracy of GPS and compass and decreasing field-of-view angles.

This work proposes a methodology for object identification in digital imagery alternative from the existing methodologies: computer vision technology and photogrammetric equations. It is concluded that the use of GIS technology and spatial data to create a virtual scene as output of perspective viewer services is appropriate to apply in object identification and localization. In doing so, the problem of label placement in three-dimensional geographic environments is reduced to a two-dimensional map labeling problem. The best location of label placement was determined using constraints and rules to be applied to the virtual scene and the reclassified-to-binary image of the input photo. In addition, depth maps enable the variation of label font size depending on the object distance to the photographer.

Two main limitations were identified in this approach. The first is that using current consumer devices (GPS receiver and digital compass) to acquire the geographical spatial metadata resulted in increased misidentification of features owing to inaccuracies of the sensors, when compared to the high-accuracy professional device. The other limitation found relates to the performance when handling a large amount of features in the extrusion model (from the vectorized elevation model). The tools we used for this study create a single 3D model that owing to its large size limits the spatial extent of the area to analyze. To solve this issue, it is proposed to implement this process in the form of a webservice that uses the data stored on dedicated spatial databases. In this way the service can create on-the-fly the 3D model based on only the relevant area for the photo, making the area extent not a limitation since we eliminate the need for a unique 3D model. Only the availability of the data for any region is the limitation.

Further research is recommended to evaluate with real users the constraints and rules for the label algorithm, since the strategies to place the labels should be user driven or based on user preferences.

Acknowledgements

The authors gratefully acknowledge all who contributed to this research, in particular Peter van Oosterom. Parts of this work were supported by the IST FP6 project TRIPOD (045335), see: www.projecttripod.org.

References

Azuma R (2004) Overview of augmented reality. International Conference on Computer Graphics and Interactive Techniques ACM SIGGRAPH 2004 Course Notes, Los Angelos.

Bell B, Feiner S, Höllerer T (2001) View Management for Virtual and Augmented Reality. Proceedings of the 14th annual ACM symposium on User interface software and technology 2001 pp 101-110.

Cartwright W, Gartner G, Peterson MP (2007) Multimedia Cartography Second Edition. Springerlink, Berlin Heidelberg New York.

Chang EY (2005) EXTENT: Fusing Context, Content, and Semantic Ontology for Photo Annotation. ACM SIGMOD CVDB Workshop, Baltimore.

Dias E, de Boer A, Fruijtier S, Oddoye JP, Harding J, Matyas C, Minelli S (2007) Requirements and business case study. Project deliverable D1.2. TRIPOD: TRI-Partite multimedia Object Description. EC-IST Project 045335 (www.projecttripod.org).

Li J, Plaisant C, Schneiderman B (1998) Data object and label placement for information abundant visualizations. Proceedings of the 1998 workshop on New paradigms in information visualization and manipulation, Washington D.C., pp 41-48

Götzelman T, Hartman K, Strothotte T (2006) Agent-based annotation of Interactive 3D Visualizations. 6th International Symposium on Smart Graphics, Vancouver, pp 24-35.

Hagedorn B, Maass S, Döllner J (2007) Chaining Geoinformation Services for the Visualisation and Annotation of 3D Geovirtual Environments. 4th International Symposium on LBS and Telecartography, Hong Kong.

Kolbe TH (2007) Augmented Videos and Panoramas for Pedestrian Navigation. 2th International Symposium on LBS and Telecartography, Vienna.

Naaman M, Harada S, Wang QY, Garcia-Molina H, Paepcke A (2004) Automatic Organization for Digital Photographs with Geographic Coordinates. Proceedings of the Fourth ACM/IEEE-CS Joint Conference on Digital Libraries, pp 53-62.

Maass S, Döllner J (2006a) Dynamic Annotation of Interactive Environments using Object-Integrated Billboards. Proceedings 14-th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision, WSCG’2006, Plzen, pp 327-334.

Maass S, Döllner J (2006b) Efficient View Management for Dynamic Annotation Placement in Virtual Landscapes. 6th Int. Symposium on Smart Graphics, Vancouver, pp 1-12.

Schmalstieg D, Reitmayr G (2007) The World as a User Interface: Augmented Reality for Ubiquitous Computing. Location Based Service and TeleCartography, Springer, pp 369-391.

Toye E, Sharp R, Madhavapeddy A, Scott D, Upton E, Blackwell A (2006) Interacting with mobile service: an evaluation of camera-phones and visual tags. Personal and Ubiquitous Computing, pp 1 – 10.

Yamamoto M, Lorena LAN (2005) A Constructive genetic approach to point-feature cartographic label placement. Metaheuristics: Progress as Real Problem Solvers, Springerlink, New York.

Interactive Geovisualization and Geometric Modelling of 3D Data - A Case Study from the Åknes Rockslide Site, Norway

Trond Nordvik[1], Chris Harding[2]

[1] Division of Geomatics, Norwegian University of Science and Technology, N-7491 Trondheim, Norway
[2] Department of Geological and Atmospheric Sciences, Iowa State University of Science and Technology, Ames, Iowa 50011-2274, USA

Abstract

This paper reports on the 3D visualization and modelling of multiple 3D data sets from the Åknes rockslide site, one of the world's largest and most complex rock slide assessments. Using the open-source architecture OpenSceneGraph (OSG), we created an interactive 3D application which enables the project's geoscientists to visualize combinations of several different 3D data sets from the site. The application also allows them to model the subsurface geometry of suspected sliding surfaces which play a key role in the assessment of the rockslide and its prediction. This interactive modelling uses a constrained Delaunay triangulation method. This cross-platform application is designed to run on a typical desktop PC and does not contain any operating system specific components. While there are some caveats, OSG has provided us with a flexible and cost-effective high-level platform for the development of our geovisualization application.

Key words: 3D geovisualization, geometric modelling, rockslide, OpenSceneGraph

1 Introduction

Åknes is a headland (promontory) situated above the entrance of the Geiranger fjord (Fig. 1). This fjord is part of the world heritage list and attracts several hundred of thousands visitors per year (many of them onboard cruise ships). As the Åknes slope exhibits continuous creep of the rock mass which could lead to a massive rock slide, a comprehensive mapping program was initiated in 2004 (Derron et al. 2005, Kveldsvik et al. 2006) making Åknes one of the most investigated and instrumented slopes in the world (Ganerød et al. 2008). A large-volume rock slide from the slope into the nearby fjord could lead to a massive tsunami that would strike several inhabited villages (Blikra et al. 2005) and could endanger nearby cruise ships. A warning system has been established to allow the evacuation of the surrounding areas. The Åknes rock slide forms part of a steep mountain slope on the west side of the fjord. The slope rises from the fjord to about 1500 meters above sea level and is 30-40 degrees steep. Current investigations define the unstable parts of the slope as starting at a 800 m long back scarp zone (across slope) at roughly 800-900 m above sea level and extending for roughly 1000 m down the slope (Ganerød et al. 2008). Preliminary estimates put the overall volume of the potential rock slide at up to 80 million m^3 (Rønning et al. 2006), 30 times the volume of the Great Pyramid of Giza (2.6 million m^3, Edwards 2003).

Our work is part of a larger effort to analyse the many data sets obtained from the Åknes site and to determine the rock slide's 3D geometry and its material. This effort will help to estimate the potential volume of the unstable rocks and to predict the consequences of a potentially massive slide event. Surface data sets include; high resolution elevation data from Light Detection and Ranging (LIDAR), real-time Global Positioning System (GPS) data of critical locations along the slope and bathymetry data for the fjord's seabed, sub-surface data sets include geophysical surveys (geo-electrical resistivity, ground penetrating radar, seismic) and borehole measurements. Data from GPS, extensometers and interferometric Synthetic-Aperture Radar (inSAR) measurements are updated over time and thus add a temporal dimension to the research. Visualising these different types of surface and subsurface data sets as 3D ensembles helps geoscientists to validate incoming field measurements, to formulate and answer spatial questions, and to develop an idea of the "big picture". The accurate determination of Åknes' sliding surface(s) is of paramount importance. The extent and nature of these sliding surfaces, such as their slope and aspect and the volume of the rock units they delineate, are vital inputs for the numerical modelling of the slide dynamics and for the tsunami modelling, which, in turn are used to predict the consequences of different sliding scenarios.

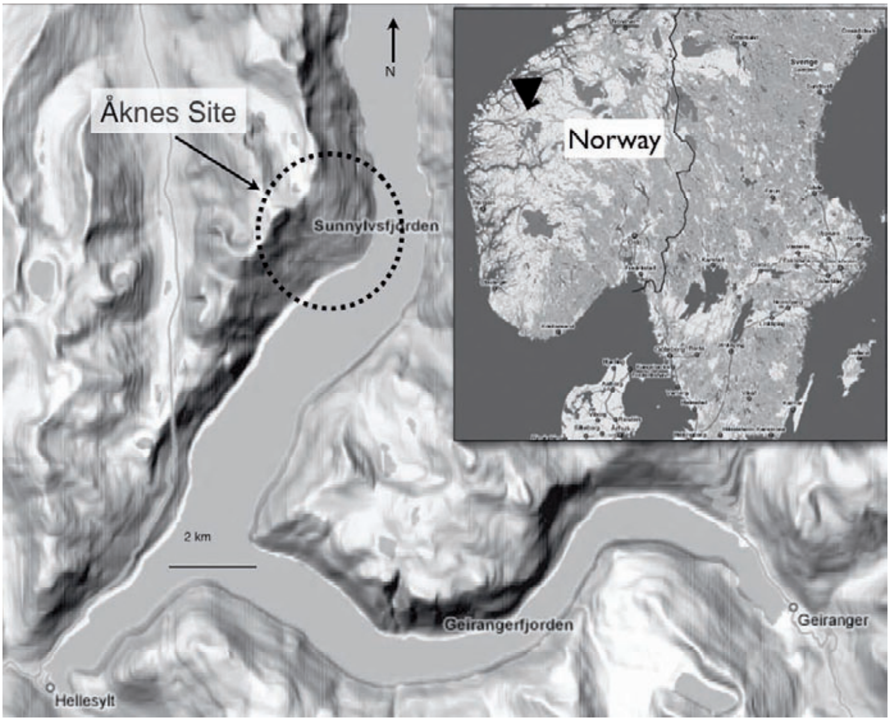


Fig. 1. Location of the Åknes site

While “off the shelf” software has been used for parts of the project, it has been difficult to find existing software that supports the requirements of the Åknes project. Instead, this custom software application has been created together with the project’s geoscientist, to adapt to the project’s needs and to its evolving data. Our application uses a high-level 3D graphics scene-graph architecture (OpenSceneGraph) and leverages many well known principles and techniques of 3D geovisualization, to address the specific visualization and modelling needs of the Åknes project. The application presents the currently available 3D data to the geoscientist in an easy-to-understand form and thus provides an intuitive overview of the spatial situation at the Åknes site. Because of the vital role of the sliding surfaces, the system also allows the geoscientists to interactively model (sculpt) the geometry of the main sliding surface within the spatial context of the other 3D data sets. This interactive visualization and modelling application can be run on common personal computer hardware. The rest of this paper describes the use of OpenSceneGraph in this application, the Åknes site's 3D model and its visualization, the interactive modelling of the sliding surface's geometry and closes with a discussion and summary.

2 OpenSceneGraph

OpenSceneGraph (OSG) was chosen as a basis for this 3D application (OSG, 2007). OpenSceneGraph is a freely available, open source application programming interface (API) used to develop 3D graphics applications including applications for geovisualization, e.g. (Kada et al. 2003), (ossimPlanet 2007) and (Sherman et al. 2007). OpenSceneGraph uses a high level scene graph structure and Open Graphics Library (OpenGL) rendering to provide a full visualization framework. The scene graph architecture, a collection of nodes organised as a hierarchical tree data structure, is used to efficiently manage and render 3D data. The OSG project started almost a decade ago and it is currently used by more than 1600 developers to develop 3D applications (Martz 2007). OpenSceneGraph is written in C++ and uses OpenGL, the de facto standard rendering API for cross-platform computer graphics. OpenSceneGraph supports the standard template library (STL) and combines high performance graphics with object oriented (OO) programming. Besides providing high level access to computer graphics (Fig. 2), OSG supports smart pointers and other modern programming functionality.

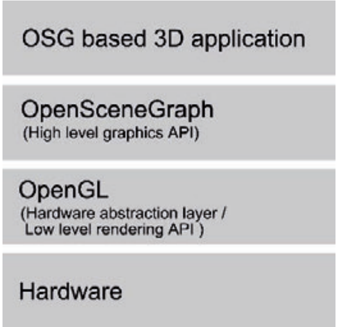

OSG based 3D application
OpenSceneGraph
(High level graphics API)
OpenGL
(Hardware abstraction layer /
Low level rendering API)
Hardware

Fig. 2. A cross-platform 3D application uses the high level OSG scene graph API, which uses OpenGL as the rendering engine

The nodes of a scene graph (Fig.3) represent a tree structure that connects the scene's physical and abstract objects. In general, rendering this scene means traversing this tree from the root and perform a function at each node encountered that possibly carries over to the nodes “beneath”, effectively compositing an image (frame) from the current state of the tree’s nodes. When navigating, the scene graph is traversed roughly 30 times per second and creates the impression of a fluid response to the user’s interactions. Types of nodes used in this project include: geometry nodes (to define a 3D geometric model as a collection of triangles), transformation-nodes (to move/rotate/scale the model in 3D space), lighting nodes (representing light sources that illuminate the model) and camera nodes (to “record” the view of the model seen on the monitor), group-nodes (to affect multiple nodes as a group) and switch nodes (selectively affect only a subset of the nodes).

Compared to using OpenGL alone for application development, OSG offers a number of convenient high level features and performs several levels of optimization automatically - for example, OSG automatically discards parts of the scene that are currently not visible (co-called culling). OpenSceneGraph’s architecture allows the programmer to work with a wide range of 3D data types; its object oriented nature encourages the re-use of already tried and true methods. Other OSG functionality useful for visualization includes interactive object selection (picking via mouse or a 3D ray) and stereo viewing. OpenSceneGraph supports the import/export of 3D models in many 3D file formats, for example VRML (.wrl), ESRI Shape (.shp), 3D Studio Max (.3ds) and Alias Wavefront (.obj) for data interchange with other 3D applications. Although, OSG is a general purpose graphics API there are some features especially designed for geospatial data such as the osgTerrain module for generating paged (tiled) terrain databases. However, using OSG for our project made us also aware of its limitations and shortcomings, some of which will be addressed in the discussion section.

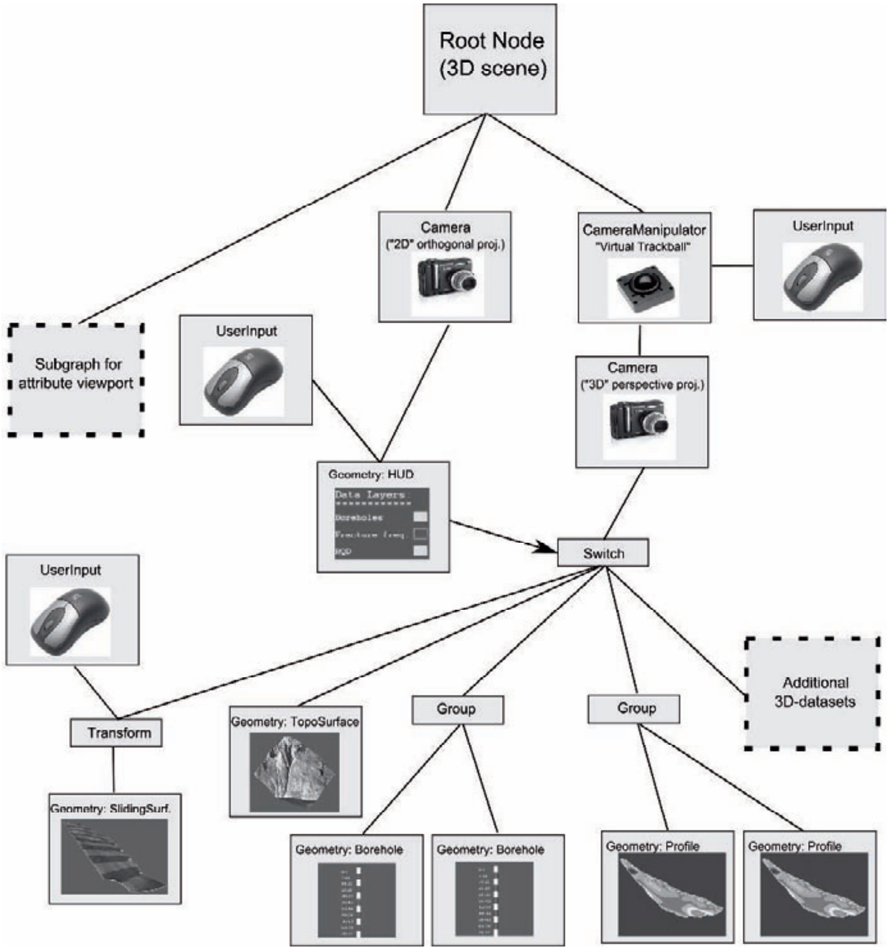
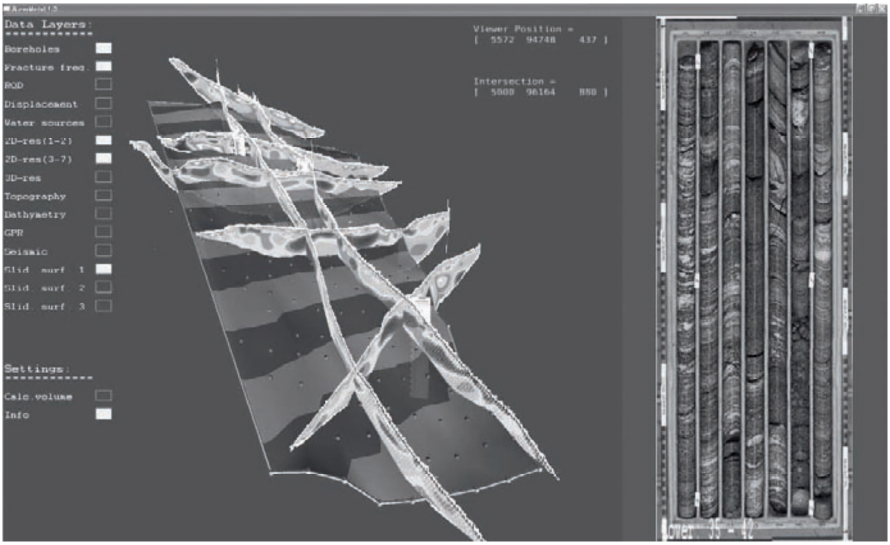


Fig. 3. Simplified subset of our application's scene graph. The scene graph provides a traversal (rendering) tree for nodes with specific functions to create a user-driven view of the 3D data sets. The switch node, selecting the currently visible data sets, is controlled by interaction with the head-up display (HUD)

3 Design and implementation of the 3D Åknes model.

Our application was developed using OpenSceneGraph (version 2.0) and Microsoft Visual Studio (version 8.0) running on the Windows XP operating system. The visualization application supports several forms of stereo viewing, such as active stereo (shutter glasses) and passive stereo (polarized projectors/glasses). Based on input and requirements from a group of geoscientists (geologists, geophysicists and geotechnical engineers) affiliated with the Åknes project, this geovisualization application was designed as a stand alone application to view and interact with the 3D model of the site on a common Windows PC. The 3D model of the site is organized as a hierarchical scene graph using a number of OSG node classes. Fig. 3 shows a simplified subset of the models scene graph. The main window displays one or more of the available 3D data sets in a perspective projection whilst an auxiliary 2D window displays additional 2D information such as drill core images (Fig. 4). The 3D data sets can be selected (switched on and off) by clicking on a set of check boxes with the mouse, currently around 30 different data sets. Because OSG itself does not integrate standard interface components such as menus and buttons etc., check boxes were implemented as a head-up display (HUD), which draws text and symbols in front of the 3D scene as "overlay graphics" (Fig. 4).



Data Layers:
============
Boreholes ☑
Fracture freq. ☑
RQD □
Displacement □
Water sources □
2D-res(1-2) ☑
2D-res(3-7) ☑
3D-res □
Topography □
Bathymetry □
GPR □
Seismic □
Slid. surf. 1 ☑
Slid. surf. 2 □
Slid. surf. 3 □

Settings:
============
Calc.volume □
Info ☑

Viewer Position =
[5572 94748 437]

Intersection =
[5000 96164 850]

Fig. 4. Graphical User Interface. Left viewport, displaying the 3D model, has perspective viewing projection whilst the right viewport, displaying additional attribute information, has orthogonal viewing projection

A predefined OSG trackball camera manipulator class is used for navigation in the 3D scene (main viewport). This provides basic navigation functions, such as panning, rotation and zoom, by mimicking a virtual trackball via the mouse. For example; left mouse button dragging will rotate the 3D model around its origin, while middle mouse button dragging pans (moves) around and right mouse button dragging will induce zooming (exo-centric navigation). To help the viewers to recover their bearings, we implemented an improved option for resetting the camera's position and attitude. In this improved reset, the position and the horizontal components of the viewing direction are maintained but the vertical view component is set to zero, that is the pitch and the roll are set to zero whilst the heading and the position are kept. This option is useful when moving vertically along objects in the model to examine geological structures and inspect data values. The navigation functions can also be used to create a fly-through, an animation that provides a continuous bird's eye view to the model and is well suited to communicate an overview of the site. Fig. 5 shows the topography and bathymetry of the Åknes area (the chessboard pattern was added to illustrate the rough extent of the Åknes site and can be switched off).

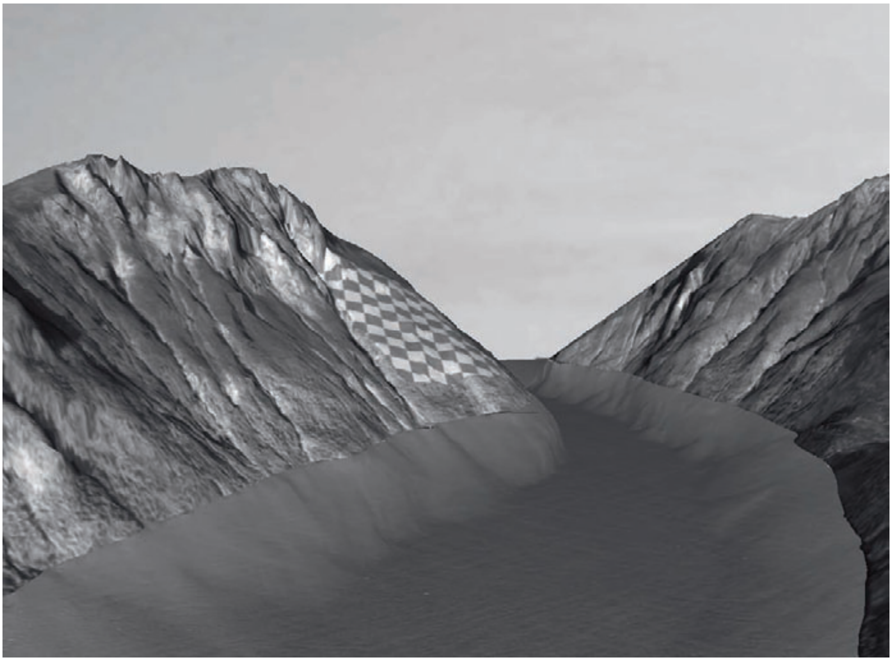


Fig. 5. Overview of the Åknes site. (The slide area is indicated with a chessboard pattern)

4 Methods used in the preparation and visualization of the 3D model

Since 2004, an extensive survey campaign of characterizing the geoscientific (surface and subsurface) conditions has been conducted at the Åknes site. This includes maps from field observations (geological, hydrological and geomorphological maps) and data from aerial photography, airborne and terrestrial LIDAR, geo-electrical resistivity, refraction seismic, ground-penetrating radar, various types of borehole measurements and bathymetric measurements. Several different types of permanent field stations were established to monitor rock changes throughout the Åknes site, these include GPS, inclinometers, extensometers, permanent radar reflectors (inSAR scatters), reflectors for total stations, climate sensors (precipitation, temperature) as well as permanent seismic installations. We integrated many of these different 3D data sets into the OSG 3D model of the Åknes site.

We used OSG Delaunay triangulations to represent 2.5D surface (elevation) data of the site's topography and bathymetry (Fig. 5), and its subsurface sliding surfaces (Fig. 4, 7, 8, 10 and 11). A triangulated irregular network (TIN) model represents both a contiguous mesh of non-overlapping triangular facets and a data structure: "A TIN model can be regarded both as a terrain model as well as a data structure. It is a model because the space-filling triangular planar facets determine a value for the surface everywhere, but it also has a specific structure in which the points, triangles and topologies are stored." (Kumler 1994). A Delaunay triangulation is one common way to create a TIN model, such a Delaunay type TIN is made from "*triangles where no points in the network are enclosed by the circumscribing circles of any triangle*" (Midtbø 1993). The Delaunay triangulation's solely mathematical criterion forces the creation of "well-shaped" triangles with edges that are not radically different in length (i.e. are not "skinny"); this has proven to be beneficial in many real world and theoretical applications. To further improve the shape of a triangulated surface, additional constraints can be imposed during its creation. For example, so-called break lines can be added which require the TIN to honour the geometry of these lines (i.e. none of its triangles can cut across a break line), however, such triangles will not necessarily fulfil the Delaunay criterion. Break lines are useful to model abrupt changes in the topography or to embed man-made objects, such as roads, into the terrain.

The site's raw topography and bathymetry data (measured by LIDAR and multi beam echo sounder, respectively) were given as geo-referenced points (x, y, z locations on the Earth's surface). The OSG surface model is based on these points. The x and y coordinates are given in UTM coordinates, the z coordinate (elevation) is given with respect to the geoid. Both raw data sets are very detailed and contain a large number of data points, roughly 50 million points for the area shown in Fig. 5 at typically 1 to 2 points per square meter. To achieve a real time visualization of the 3D model (i.e. more than 30 frames per second) on modest hardware, we had to reduce the 3D model's number of points and the size and total amount of texture ("images") draped over the TIN. OpenSceneGraph provides a "DelaunayTriangulator" node which can be customized using its "DelaunayConstraints". When creating the OSG surface models from this original data, we used an adaptive triangulation approach (Heller 1990, Nordvik and Midtbø 2007) which intelligently lowers the number of vertices and triangles while keeping the mesh's quality adequate. Adaptive triangulation changes the "resolution" of the triangle mesh depending on the raw data's elevation. In areas with little change of slope, a relatively coarse mesh sufficiently represents the topography, in more complex areas; a larger number of triangles are used to archive an acceptable level of approximation error. This method could also be used to create multiple surfaces of different quality for the same spatial extent, which represent different levels of details (LOD) of this area.

To create a photo-realistic appearance, a high resolution ortho-photo was draped over the geometric model of the site's topographic surface. In the original ortho-photo 1 pixel corresponds to 12.5cm on the ground, roughly 15,000 x 15,000 pixels to cover the site. Although even commodity graphics adapters can now store many 100s of megabytes of total texture images, single texture images are typically limited to sizes from 2048 x 2048 pixels to 8192 x 8192 pixels. Larger texture images have to be either rescaled or split into smaller images (tiles) before they can be used. However, a reduction in resolution is undesirable; draping the split texture images on rectangular tiles requires calculating and inserting new "helper vertices" and would create artificial break lines along the straight borders. Using OSG's multi-texture facilities allowed us to split the large texture image into several smaller textures "tiles" and to blend these tiles together without breaking up the TIN internally. Although each texture tile is mapped to the entire mesh, valid texture coordinates ($s \in [0,1]$, $t \in [0,1]$) are only assigned to vertices within the area corresponding to that specific tile. Vertices outside the tile's area are given texture coordinates outside the valid 0.0-1.0 range resulting in a RGB colour value of (0,0,0) (i.e. the colour is clamped to black). After this procedure is completed for all tiles, the final colour values for the TIN are calculated by summarising all separate texture tiles. Although this method requires additional work in terms of texture coordinates, the final result is a mosaic corresponding to the initial oversized texture (Fig. 6) where only pixels within each tile will contribute to the overall texture colour while the black pixels add nothing (0,0,0).

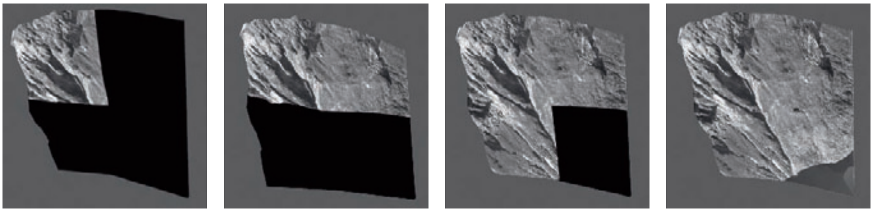


Fig. 6. Assembling a high resolution texture by adding four textures together using OSG's multi-texturing capability. From left to right, each images shows the addition of a new texture (quarter)

We visualized data from various geophysical surveys, including 2D and 3D inversions of geo-electric resistivity measurements and ground-penetrating radar. We show the 2D resistivity inversions as series of 2D textured, vertical slices at the correct spatial location within the 3D model, so-called fence diagrams. The 3D resistivity inversion was transformed into a 3D texture (a 3D texture array); this automatically "paints" any arbitrary geometry placed inside this volume with the resistivity values. Besides allowing the user to move vertical and horizontal sections through the volume, we also paint the 3D resistivity volume on the geometry (TIN) of the sliding surface (see next section). As the TIN of the topography may obscure the subsurface data beneath it, we use an OpenGL clip plane to temporarily cut away parts of the TIN. Fig. 7 shows the 3D textured sliding surface (TIN) and a horizontal section (profile) moved to two different positions.

Core samples of seven boreholes in three different locations of the slide area provide accurate information about the subsurface along these holes. After the drilling, the cores themselves were photographed and a large number of parameters were measured either in the drill holes or from the drill core samples (see Rønning et al. 2006 and Ganerød et al. 2007 for a complete list of measured parameters). In the 3D OSG model, the boreholes themselves are represented as sets of black-and-white segments (cylinders); each segment represents a 7 m stretch of core. Plots of the measured parameters are presented as textured "billboards" (panels) next to the boreholes in the 3D model. Billboards are "smart" rectangles which automatically rotate towards the viewer along the vertical axis and thus avoid situations where the viewer cannot see the front of the panel. Clicking on a borehole's cylinders with the mouse shows high-resolution images of the rocks found along each segment of the borehole (core images) in a separate part of the application (right in Fig 8).

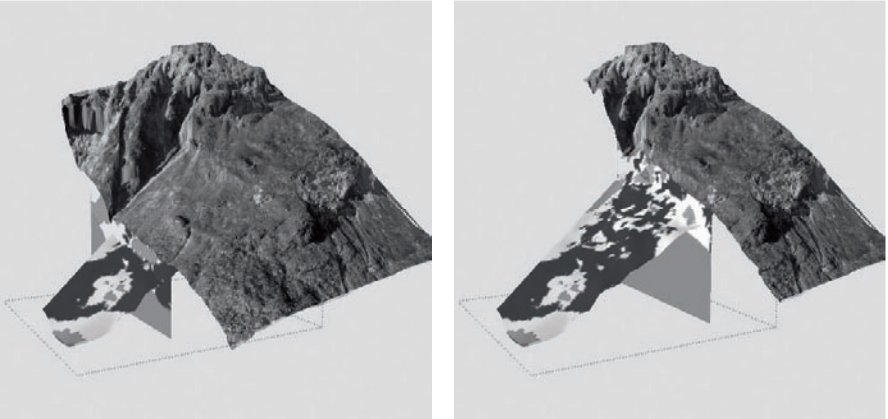


Fig. 7. Using 3D texture to visualize the 3D resistivity beneath the topography. Left and right figure show two different cross sections set by the user, the topography is clipped along the cross section

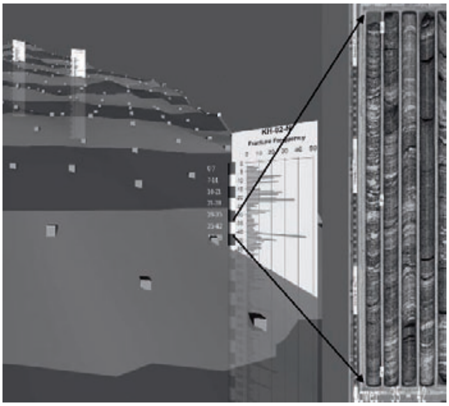


Fig. 8. A semi transparent sliding surface and boreholes with billboards containing fracture frequency plots displayed together with a core image

During field work campaigns the locations of water sources within the slide area were noted and their properties were established. These water sources are represented as animated OSG particle generators in the 3D model. The properties of the particle generators are set to mimic the properties of these actual water sources; the size and frequency of the particles reflect the amount of water, the height of the fountain reflects the pressure. While certainly not realistic, these fountains serve as symbolic, yet quantitative visualizations of the location of these water sources and their properties (Fig. 9, left). The parts of the Åknes site that are suspected to form the sliding rock units are monitored for creep by extensometers, inSAR, GPS and total stations. The right section of Fig. 9 shows a set of displacement vectors derived from GPS measurements.

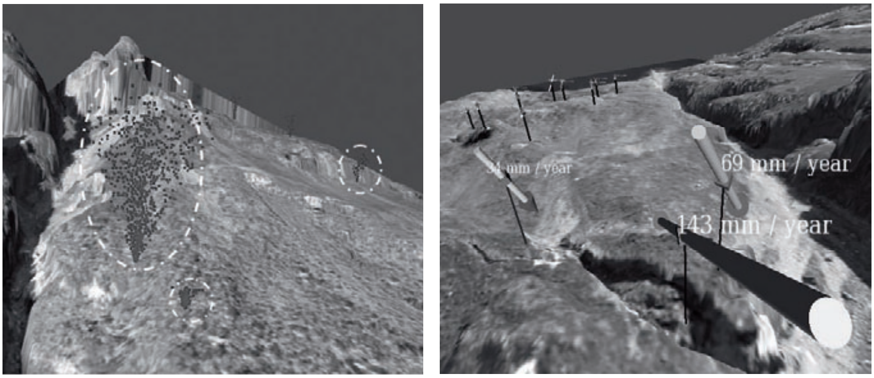


Fig. 9. Left figure: Water sources are represented as particle generators. Right figure: Vectors of the displacement at certain points of the slope over time (measured via GPS) are represented as 3D arrows in the model

5 Interactive modelling of the sliding surfaces

Besides offering a 3D visualization of the site's data sets, our application uses a constrained Delaunay triangulation to implement the interactive modelling of the sliding surfaces i.e., surfaces that, in 3D, delineate the unstable rock units from the solid bedrock below. To start this modelling process, an initial sliding surface is created at a constant depth beneath the site's topographic surface. Each TIN node is used as an editable "control point" (shown as small cubes in Fig. 8 and 10), that can be manually adjusted in x, y and z directions via the keyboard. Similarly, groups of control points can be selected by the mouse and shifted along the three coordinate axes, as can the entire surface mesh. Other 3D data sets (e.g. drill-hole data or resistivity data) can be displayed to provide the needed 3D spatial context for defining the shape of the sliding surface. A horizontal mesh resolution of 50 m was considered sufficiently detailed to capture the initial shape of the surface, however the user can interactively add more control when and where more subsurface information becomes available, e.g., along the profiles and in the vicinity of boreholes. After the user has moved, added or deleted vertices, the surface is re-triangulated in real time to show the new shape and how it fits into the surrounding data (Fig. 10). This lets the user explore what-if scenarios and to compare different solutions. In general, a “good” surface needs to fit not only into the surrounding information but also into a mental model of the overall geological and geophysical “situation” of the site. It should be noted that this geometric modelling is not as simple as forcing the shape of the surface to follow the displayed data sets (boreholes, fence diagrams, etc.). As the scale and uncertainty of the displayed data sets differ considerably, it is quite possible that data sets seem to contradict each other at certain locations. While it is still up to the skill and knowledge of the geoscientists to create a geometric model as a personal interpretation of the available data, interactive 3D visualization and modelling tools automatically provide the users with visual context and thus allow them to better concentrate on the essential tasks.

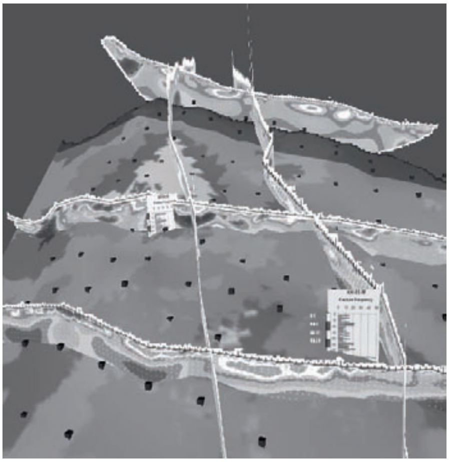


Fig. 10. Geometric modelling of a sliding surface (TIN) by moving control points, it is coloured by a 3D inverted resistivity data (3D texture); also shown: 2D inversion (fence diagrams) and borehole data

The boundary of the sliding surface is represented as a constrained polyline which enables it to form a non-convex hull (unlike the unconstrained Delaunay triangulation’s boundary, which will always form a convex hull). The sliding surface and parts of the topographic surface are connected together by embedding a “projection” of the boundary polygon into the topography TIN above (fat poly-line in Fig. 11). Note that the vertices of this upper boundary can also be moved laterally and are not simply vertical extrusions of the lower boundary polygon. This vertex-to-vertex connection defines a 3D body between the sliding surface and the corresponding part of the topography above it. The body's volume is calculated as the difference of the volume between the upper surface (topography) and a horizontal reference plane; and the volume between the lower surface (sliding surface) and the same reference plane.

As inserting more than one new *boundary* control points at once could lead to unintended switch backs (i.e. the boundary line crossing itself), we re-triangulate both surfaces after a boundary control point is inserted or deleted. When the modelling process of a sliding surface is completed, it can be exported to a file for further processing with external software.

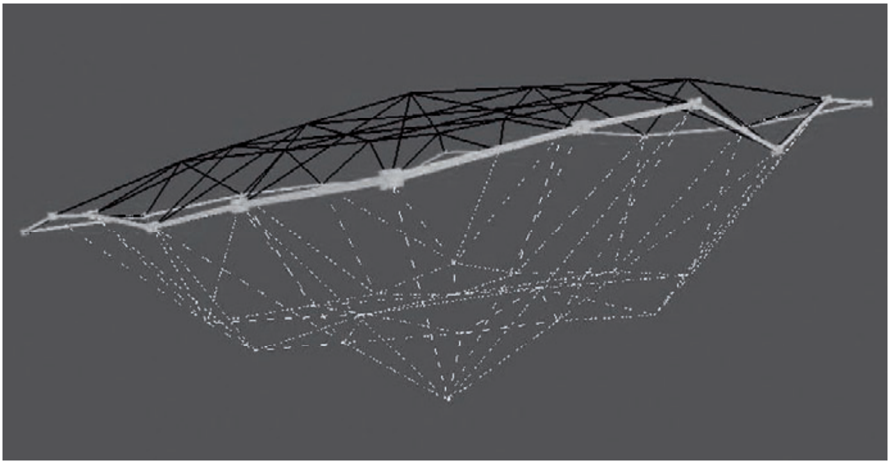


Fig. 11. Construction a 3D body from two TIN surfaces (bottom: sliding surface, top: topography), the surfaces' constraint boundary (poly-line) are connected vertically

6 Discussion and summary

We have created a visualization and modelling application that solves the specific needs of the geoscientists working on the Åknes site, including the creation of a geometric model of the sliding surfaces underlying the site. Challenges included the complexity, volume, heterogeneity of the site's data sets, the fact that the data collection is ongoing and the number and diversity of groups involved in the Åknes project. While each group may use other software packages, such as ER Mapper, ArcGIS, QT Modeler and Plaxis, to fulfil a specific functionality, our application aims to provide a "communal" visualization solution that is easy to learn and can be used by all geoscientists involved.

Unlike regular grids, TIN models are capable to adapt to the underlying scattered data set (e.g. LIDAR data) and can efficiently capture both smooth and abrupt changes in height. As TIN models are based on the same underlying primitives (triangles) as OpenGL, their display can be optimized by the renderer. For the application to work at interactive rates on an ordinary PC, we used a *constrained 2.5D TIN* surface approach as a flexible way of reducing the size of the original surface data sets by intelligently adapting the TIN's resolution. This constrained TIN approach is also used in the interactive geometric modelling of the all-important sliding surfaces and allows geoscientists to shape a surface within the context of surrounding 3D data sets. While the use of Delaunay constraints enables us to integrate break lines, holes and vertical sections in the surfaces, there are fundamental limitations to such a 2.5D approach, most notably its fundamental inability to capture more complexly shaped (folded) surfaces with multiple z values, i.e. overhangs.

OpenSceneGraph provided us with the flexibility to combine different types of data sets, such as elevation data, 3D grid data, 2D vertical sections and 1D borehole data, into high-level 3D objects that the user can interact with. OpenSceneGraph also provided a flexible way for 3D navigation. While the open source nature provides everybody with OSG's source code, its documentation is still of variable quality and occasionally absent; this may provide a hurdle to newcomers who are not used to learning from code examples. The OSG version used for this project (version 2.0) does not provide direct support for parametric surfaces (2D/3D splines), 3D triangulations (tetrahedral meshes) or 2D/3D iso-surfaces (contours). Despite these caveats, our experience points to OSG as a flexible and cost-effective high-level platform for the development of cross-platform geovisualization applications. Because OSG is an open-source (non-commercial) system, our application can be freely distributed to all scientists involved in the project and can be further extended as the need for additional functionality arises. For example, while our application can already work in stereo and make use of a large single projection screen, it can be further extended via virtual reality (VR) toolkits (such as VR Juggler) to work on multiple stereo screens (e.g. a CAVE system) and to support 3D tracking and 3D input devices. While our application is developed on a Windows desktop computer, it uses only OSG and OpenGL libraries and can be adapted to multi-node (cluster) systems and other operating systems. We have experimented with such a VR-version of our application, using a four-sided CAVE and a wireless gamepad, and envision its use for collaborative visualization and modelling.

Acknowledgements

The authors would like to acknowledge a number of scientists affiliated with the International Centre for Geohazards, the Geological Survey of Norway and the Norwegian University of Science and Technology for their contribution to this project either by exchange of ideas or by providing data. The first author wishes to thank the faculty and the staff at the Virtual Reality Applications Center at Iowa State University of Science and Technology, USA, for having him as a visiting scholar during the initial stage of this project.

References

Blikra LH, Braathen A, Derron MH, Kveldsvik V, Grøneng G, Dalsegg E, Elvebakk H (2005) Rockslope failure at Åknesremna, Stranda, western Norway. Report by the Geological Survey of Norway

Derron MH, Blikra LH, Jaboyedoff M (2005) High resolution digital elevation model analysis for landslide hazard assessment (Åkerneset, Norway). In: Senneset k, Flaate K, Larsen JO (eds) Proceedings of the 11th International Conference and Field Trip on Landslides (ICFL). 1st-10th September 2005, Norway. Taylor & Francis Group, London

Edwards JF (2003) Building the Great Pyramid - Probable Construction Methods Employed at Giza. Technology and Culture. The Johns Hopkins University Press, Vol.44, No. 2, pp 340-354

Ganerød GV, Grøneng G, Rønning JS, Dalsegg E, Elvebakk , Tønnesen JF, Kveldsvik V, Eiken T, Blikra LH, Braathen A (2008, in press) Geological Model of the Åknes Rockslide, western Norway. Engineering Geology

Ganerød GV, Grøneng G, Aardal IB, Kveldsvik V (2007) Logging of drill cores from seven boreholes at Åknes, Stranda municipality, Møre and Romsdal County. Report 2007.020 by the Geological Survey of Norway

Heller M (1990) Triangulation algorithms for adaptive terrain modeling. In: Brassel K, Kishimoto H (eds) Proceedings of the 4th International Symposium on Spatial Data Handling, International Geographical Union, Zurich, Vol. 1,pp 163–174

Kada M, Roettger S, Weiss K, Ertl T, Fritsch D (2003) Real-time visualization of urban landscapes using open-source software. In: Proceedings of the 24th Asian Conference on Remote Sensing & Internatioanl Symposium on Remote Sensing, 3rd-7th November, 2003, Busan, Korea.

Kumler MP (1994) An Intensive Comparison of Triangulated Irregular Networks and Digital Elevation Models. Cartographica, Monograph 45, Vol. 31, No. 2, pp 1-49

Kveldsvik V, Eiken T, Ganerød GV, Grøneng G, Ragvin N (2006) Evaluation of movement data and ground conditions for the Åknes rock slide. The

International Symposium on Stability of Rock Slopes. The African Institute of Mining and Metallurgy. April 2006, pp 279-299

Martz P (2007) OpenSceneGraph Quick Start Guide. Skew Matrix Software LLC

Midtbø T (1993). Spatial Modelling by Delaunay Networks of Two and Three Dimensions, Doctoral thesis No. 23, 1993, Norwegian Institute of Technology

Nordvik T, Midtbø T (2007) Application of Data Reduction Methods in Dynamic TIN Models to Topographic LIDAR Data. In: Bjørke JT, Tveite H (eds) Proceedings of the 11th Scandinavian Research Conference on Geographical Information Science, 5th - 7th September, 2007, Ås, Norway, pp 111-128

OSG (2007) OpenSceneGraph website (Cited 26 November 2007) URL: http://www.openscenegraph.org/projects/osg

ossimPlanet (2007) ossimPlanet website (Cited 27 November 2007). URL: http://www.ossim.org/OSSIM/ossimPlanet.html

Rønning JS, Dalsegg E, Elvebakk H, Ganrød G, Tønnesen JF (2006) Geofysiske målinger Åknes og Tafjord. Report 2006.002 by the Geological Survey of Norway, (Norwegian text)

Sherman WR, Penick MA, Su S, Brown TJ, Harris Jr FC (2007) VR Fire: an Immersive Visualization Experience for Wildfire Spread Analysis. IEEE Virtual Reality Conference, 10th – 14th March, 2007, Charlotte, North Carolina, USA, pp 243-246

FieldGML: An Alternative Representation For Fields

Hugo Ledoux

Delft University of Technology (OTB—section GIS Technology)
Jaffalaan 9, 2628BX Delft, the Netherlands
h.ledoux@tudelft.nl

Abstract

While we can affirm that the representation, storage and exchange of two-dimensional objects (vector data) in GIS is solved (at least if we consider the *de facto* standards *shapefile* and GML), the same cannot be said for fields. Among the GIS community, most people assume that fields are synonymous with raster structures, and thus only representations for these are being used in practice (many formats exist) and have been standardised. In this paper, I present a new GML-based representation for fields in 2D and 3D, one that permits us to represent not only rasters, but also fields in any other forms. This is achieved by storing the original samples of the field, alongside the interpolation method used to reconstruct the field. The solution, called *FieldGML*, is based on current standards, is flexible, extensible and is also more appropriate than raster structures to model the kind of datasets found in GIS-related applications.

1 Introduction

There exist two contrasting conceptualisations of space: the *object* and the *field* views (Couclelis, 1992; Goodchild, 1992; Peuquet, 1984). In a nutshell, the former view considers space as being 'empty' and populated with discrete entities embedded in space, while the latter considers the space as being continuous, and every location in space has a certain property (there is *something* at every location). In the former model, entities can be for example roads, cups of tea, churches, etc., and they have certain properties; in the latter, they are formed by clusters of properties. When one wants to represent and store a certain piece of space in a computer, the field-view approach is much more problematic than its counterpart. The problems are most likely caused by the fact that the definition of a field itself changes from discipline to discipline, and that the issues can be seen from a philosophic, conceptual or implementation point of view (Peuquet et al., 1999). There is also much confusion among users between spatial models, data structures, and spatial concepts (Frank, 1992). While in the GIS jargon object- and field-views of space are often synonymous with respectively vector and raster models, Goodchild (1992), among others, explains that this is simply false as both views can be stored with either model. Put on top of that that fields are by definition something continuous—and that computers are discrete machines—and one can start understanding the confusion among users. (More information about fields and their representations is available in Section 3.)

In recent years, with the multitude of formats available, practitioners have turned to defining and using standards (e.g. those of the Open Geospatial Consortium (OGC), and of the International Organization for Standardization (ISO)) to facilitate the storage and exchange of GIS data. While the current standards are somewhat successful and promising for objects (with the Geography Markup Language (GML) leading the way), their use for fields are very scarce, and are mostly limited to raster solutions. However, as argued in Section 4, rasters are technically and theoretically restrictive and therefore alternative solutions should be sought.

In this paper, instead of using solely raster formats to represent fields, I propose an alternative representation called *FieldGML*. As described in Section 5, this generic solution is based on current standards (i.e. GML), and permits us to efficiently store and exchange field-based geographic information. The main idea behind this representation is that instead of storing explicitly grids or tessellations, we store the data that were collected to study the field (the samples), and we also store the interpolation method that will permit us to reconstruct the field in a computer. I also argue in the following that FieldGML offers a better representation than current ones because: (i) it takes into account the nature of datasets as found in GIS-related applications; (ii) it is valid for fields in 2D and 3D, but can be readily extended to higher-dimensions; and (iii) it is flexible in the sense that different types of fields can be stored (scattered points, tessellations, tetrahedralizations, voxels, etc.). I also present in Section 5 a prototype that was developed to create FieldGML files from already existing fields, and also to transform FieldGML files into different formats and representations that are being used by commercial GISs.

2 Related Work

From a "standards" point of view, different XML-based languages (eXtensible Markup Language) have been proposed. First of all, there is the more general-purpose GML that implements many of the ISO/OGC standards for fields, but not all of them. Note that the definitions of these standards can be found in Section 4, and their implementation with GML is discussed in Section 5. Based on GML/XML, there are different languages to model fields. For instance, Nativi et al. (2005) propose the *NcML-GML*, which permits us to store with GML the metadata associated with netCDF files (this is a multi-dimensional raster format described in the next section). Also, Woolf and Lowe (2007) propose the *Climate Science Modelling Language* (CSML), which is used to represent all the different kinds of climate data (often fields) and their relevant information. The particularity of CSML is that, for the sake of simplicity and performance, the authors chose to use only parts of the standards: they offer a GML-based 'wrapper' around legacy formats to simplify exchange, but they are still using the legacy file for applications (these legacy files are all raster-based). Furthermore, the *Geoscience Markup Language* (GeoSciML) can be used to store any kind of information related to geology (Sen and Duffy, 2005). When fields are involved, they are usually stored in raster formats, but GeoSciML also allows the storage of the observations that were collected (interpolation methods are however not discussed).

From a GIScience point of view, different alternatives to the ubiquitous rasters have been proposed over the years, starting with tessellations into triangles (Mark, 1975; Peucker, 1978). Kemp (1993) proposes different alternatives to store 2D fields, and shows how to convert them from one representation to another when needed (for analysis). Gold and Edwards (1992), and Ledoux and Gold (2006), among others, have also discussed the use of the Voronoi diagram (in 2D and 3D) as an interesting alternative to raster-based approaches. In a proposition that is similar to the one in this paper (at least conceptually), Haklay (2004) proposes, in an attempt to model and manipulate 2D fields, to store only the samples collected, and the parameters of interpolation functions.

FieldGML is also conceptually very similar to the concept of *virtual data set* (VDS) (Stephan et al., 1993; Včkovski and Bucher, 1996; Včkovski, 1998). A VDS is a dataset "enhanced" with a set of methods that are used to access, manipulate or transform the data—it is an object in the object-oriented sense of the term. The term "virtual" means that different representations of a dataset can be generated for different users/applications. In the context of fields, that means that the samples of a field are stored, and also that interpolation methods to generate different representations of that field are available (pixel size, format, data model, etc.). It is implemented as a Java class where an interface is defined.

VDSs were introduced around 15 years ago as a solution to the interoperability of GISs and to improve the quality of datasets used in GIS. The whole concept of interoperability through VDS was based on the idea that "data exchange is not specified by a standardized data structure (e.g. a physical file format) but a set of interfaces" (Včkovski, 1998, p.54). If we fast-forward to 2008, we now have widely-accepted GIS-related standards (see Section 4) and even a *de facto* language (GML). These standards have taken a different approach to interoperability since all datasets are coded with the same language, which clearly contrasts with VDS where one could store the datasets in his own format as long as he/she implemented the interface. FieldGML can thus be seen as implementation of the conceptual ideas of VDS in a 2008 context where GML is synonymous with interoperability in the GIS world.

3 Fields and Their Representations

This section gives a brief overview of what fields are, from the point of view of GISscience.

3.1 Definition of a Field

A field is a concept rather difficult to define because it is not tangible and not part of our intuitive knowledge. It is easy for us to see and describe entities such as houses or chairs, but, although we can imagine fields, they are somewhat an abstract concept. The consequences of that are firstly that formalising a field is difficult, and secondly that many definitions exist in different disciplines (Peuquet et al., 1999). The definition usually used in a GIScience context is borrowed and adapted from physics. Physicists in the 19th century developed the concept of a *force field* to model the magnetic or the gravitational force, where a force (a vector with an orientation and a length) has a value everywhere in space, and changes from location to location. For most GIS applications, the vector assigned to each point of the Euclidean space is replaced by a scalar value, and we obtain *scalar fields* (it is assumed in the following that all fields are of that type).

Because each location in space possesses a value, a field must be represented mathematically. It is a model of the spatial variation of a given attribute a over a spatial domain, and it is modelled by a function, from $\mathbb{R}^d$ to $\mathbb{R}$ in a d-dimensional Euclidean space, mapping the location to the value of a, thus

$$a = f(location).$$

The function can theoretically have any number of independent variables (i.e. the spatial domain can have any dimensions), but in the context of geographical phenomena the function is usually bivariate (x, y) or trivariate (x, y, z). Notice that the domain can also incorporate time as an extra dimension, and thus we have $a = f(location, time)$.

3.2 Representation in Computers

The representation of a field in a computer faces many problems. Firstly, fields are continuous functions, and, by contrast, computers are discrete machines. Secondly, it should be stressed out that we never have access to a 'complete representation' of a geographical phenomenon. Indeed, to obtain information about a given phenomenon, one must sample it, and reconstruct the field from these samples[1]. In the context of GIS-related applications (e.g. modelling of elevation, geosciences, geology, hydrology, bathymetry, etc.), this collection of samples is hindered by the fact that unlike disciplines like medicine or engineering, we seldom have direct access to the whole object (think of collecting data underground, or at sea for instance). And even if we have complete access to the object, it is often too expensive to sample the object everywhere.

In short, to represent a field in a computer (i.e. to be able to model a continuous phenomenon), we need to:

1. have a set of samples for the given fields—they are the "ground truth" of a field. The samples are usually point-based, but other forms can also exist (for instance an image obtained with remote sensing).
2. define a set of rules to obtain the values of the attribute studied, at any location. This operation is referred to as spatial interpolation.

4 Standards and Formats to Represent Fields

4.1 GIS Standards

There are two "levels" of geographic information standards: abstract and implementation specifications. The former defines a conceptual architecture (or reference model) for different aspects related to the storage and exchange of information; and the latter are at a lower-level, i.e. they define an interface to access the properties and methods of classes defined in the abstract specifications.

Abstract specifications

In the case of fields, two documents exist: the 'Schema for coverage geometry and functions' (ISO, 2005), and the OGC document with the same title (OGC, 2007b). Notice here that fields are referred to as 'coverages' in these documents; both terms are synonymous and used interchangeably in the following. Both documents have the same content. A coverage is considered a *feature*[2], like is any geographic object in the ISO/OGC documents. So while each geographic object in a representation of a field is a feature, the field as a whole is a feature too.

The formal definition of "coverage" is the following (and its principal classes are shown in Figure 1):

[1] Even if a sensor is used to collect samples, the result (e.g. an image with pixels) is not a complete representation since each pixel usually averages the value of the studied phenomenon over the pixel area, or each pixel represents the value located in the middle of the pixel.

[2] A feature is an abstraction of a real world phenomenon; it is a geographic feature if it is associated with a location relative to the Earth (ISO, 2003).

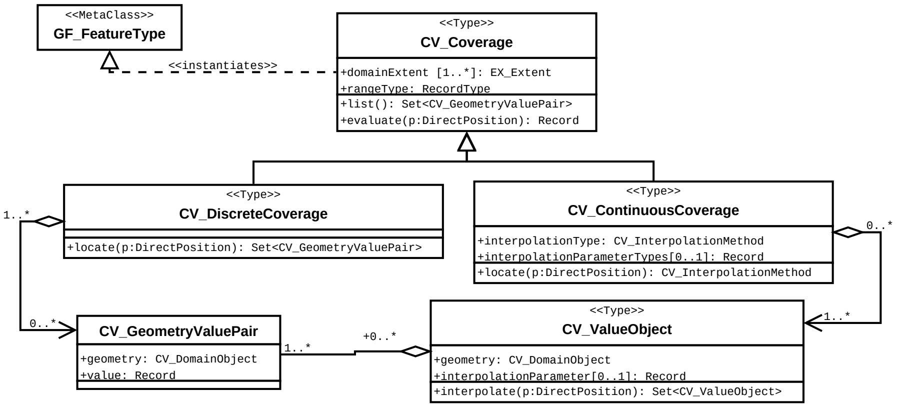

<<MetaClass>>
GF_FeatureType
<<instantiates>>
<<Type>>
CV_Coverage
+domainExtent [1..*]: EX_Extent
+rangeType: RecordType
+list(): Set<CV_GeometryValuePair>
+evaluate(p:DirectPosition): Record
1..*
<<Type>>
CV_DiscreteCoverage
+locate(p:DirectPosition): Set<CV_GeometryValuePair>
<<Type>>
CV_ContinuousCoverage
+interpolationType: CV_InterpolationMethod
+interpolationParameterTypes[0..1]: Record
+locate(p:DirectPosition): CV_InterpolationMethod
0..*
0..*
CV_GeometryValuePair
+geometry: CV_DomainObject
+value: Record
1..*
+0..*
<<Type>>
CV_ValueObject
+geometry: CV_DomainObject
+interpolationParameter[0..1]: Record
+interpolate(p:DirectPosition): Set<CV_ValueObject>
1..*

Fig. 1. UML diagram for the main classes of an ISO/OGC coverage. (Figure after ISO (2005))

> A coverage is a feature that acts as a function to return values from its range for any direct position within its spatial, temporal or spatiotemporal domain. [...] [it] has multiple values for each attribute type, where each direct position within the geometric representation of the feature has a single value for each attribute type.

Notice that a ISO/OGC coverage can have many different attribute types, but that this is not relevant here, and we simply assume that one coverage is for one attribute type (let that be the temperature of the air, the elevation of a terrain, the density of the population, etc.)

The coverage type is divided into two distinct but closely related subtypes:

Continuous Coverage: coverage that returns different values for the same feature attribute at different direct positions within a single spatial object, temporal object or spatiotemporal object in its domain.

Discrete Coverage: coverage that returns the same feature attribute values for every direct position within any single spatial object, temporal or spatiotemporal object in its domain.

The definition of a continuous coverage is more or less equivalent to that of the general coverage type. The definition refers to the fact that interpolation is used to obtain the attribute value at a given location x. The terms "within a single object" can be misleading, but means that interpolation is always performed with a function defined over one geometric object (e.g. a polygon in 2D); if no object is present at a location x (possible according to the definition of a coverage) then no value at x is returned. The latter type, the discrete coverage, seems to exist only because "a coverage can be derived from a collection of discrete features with common attributes" (ISO, 2005). As explained in Section 3.2, this is true (the samples), provided that we have a set of rules to reconstruct the coverage at every location, but this is not the case in the ISO/OGC documents. It is also stated that "a discrete coverage has a domain that consists of a finite collection of geometric objects and the direct positions contained in those geometric objects". The problem here is that these geometric objects do not have to fully partition the domain, i.e. according to that definition a set of unconnected lines and/or polygons (in which each object has a value attached to it) is considered a coverage. Even worse, the objects are permitted to overlap, which means not only do we have locations without any answer, but that there can be more than one answer at one given location! This might be useful for some applications—I am however not aware of any—but none are mentioned in the documents.

Another shortcoming is the list of interpolation methods discussed in the ISO/OGC documents is very restricted. Many interpolation methods are simply ignored, and if one wanted to use them it would be very difficult to integrate them in the coverage framework. Inverse-distance to a power (IDW) and Kriging are for instance not listed, and for many subtypes (such as *CV_TIN-Coverage*, to store triangulated irregular networks (TINs)) only one type of interpolation is possible within each piece (which is restrictive in practice).

Also, the ISO/OGC documents state that the concepts are valid not only for the 2D case, but also for three and higher dimensions. The problem is that it is only a statement weakly backed up by a few types in 3D and no explanations of interpolation methods in 3D are given.

In brief, the abstract standards for coverages do avoid the distinction between raster and vector, but by creating two types for which the differences are rather blurred and subtle, they probably also contribute to the confusion that already exists about fields.

Implementation Specifications

To my knowledge, the only implementation of the ISO/OGC abstract specifications is that of GML. It is an XML-based modelling language developed to facilitate the exchange of geographic data, and has been fairly successful in recent years. While a GML file is verbose (and thus files can become enormous), there are many advantages to using it. Lake (2000) mentions, among others: (i) it is self-descriptive, (ii) it can be processed with already existing XML software, (iii) there are mechanisms to store metadata, and (iv) data integrity can be verified with the help of *schemas*. The reader is referred to Lu et al. (2007) and OGC (2007a) to learn more.

As of GML version 3.2, only the *CV_DiscreteCoverage* types have been implemented: there are GML schemas for all subtypes of *CV_DiscreteCoverage*, and also for grids (*CV_Grid*). That results in a representation that does not necessarily cover the whole spatial domain, and no mechanisms are present to estimate the value of an attribute where there are no spatial objects, or a default and simplistic method is assumed. Using simplistic interpolation methods, or the wrong parameters for a method, is dangerous as many researchers have highlighted (see Watson (1992) for instance).

Not all abstract classes were implemented in GML, *CV_GeometryValuePair* is for instance not present, and was replaced by an implementation that follows closely the conceptual distinction between the spatial domain and the range (the attribute modelled). The resulting XML file has to have three separate types: the domain, the range and another one for mapping these two correctly. As Cox (2007) explains, although this is conceptually valid, it also hinders the use of these standards in practice because of the difficulties of processing large files, of updating files, etc.

4.2 Formats Used in Practice

Among GIS practitioners, fields are being used almost exclusively in 2D, while in the geoscience community 3D and higher-dimensional fields are extensively used. Note that the dimensions in oceanographic/atmospheric coverages are not necessarily spatial dimensions, as any parameters (e.g. temperature of the air, or density of water) can be considered a dimension.

As mentioned before, within the GIS community, coverages are more or less synonymous with grids, although it must be said that TINs are also widely used for modelling terrain elevation. There exist many different formats for 2D grids, but they can be easily all converted to one another.

In geoscience, netCDF[3] seems to be the *de facto* standard to exchange datasets, although other similar formats, such as HDF5[4], are also popular. These formats are raster-based, and permit users to use n-dimensional grids, with different spacing for different dimensions. They are binary and spatially structured, which means that parts of a dataset can be efficiently retrieved and processed. The use of other representations (e.g. tetrahedralizations or arbitrary polyhedra) is very rare and mostly limited to the academic community.

4.3 The Dangers of Using Raster Formats

As argued by many over the years, using raster structures has many drawbacks (Gold and Edwards, 1992; Kemp, 1993; Haklay, 2004; Ledoux and Gold, 2006). Firstly, as Fisher (1997) points out, the use of pixels as the main element for storing and analysing geographical data is not optimal. The problems most often cited are: (i) the meaning of a grid is unclear (are the values at the centre of each pixel, or at the intersections of grid lines?), (ii) the size of a grid (for fine resolutions, grids can become huge), (iii) the fact that the space is arbitrarily tessellated without taking into consideration the objects embedded in that space. Secondly, the use of grids in GIS/geoscience applications has wider repercussions since we can assume in most cases that a given grid was constructed from a set of point samples. Converting samples to grids

[3] `http://www.unidata.ucar.edu/software/netcdf/`

[4] http://www.hdfgroup.com

is dangerous because the original samples, which could be meaningful points such as the summits, valleys or ridges or a terrain, are not present in the resulting grid. The importance of the original samples for a field is such that they have even been dubbed the *meta-field* by Kemp and Včkovski (1998). It should also be said that when a user only has access to a grid, he often does not know how it was constructed and what interpolation method was used, unless metadata are available. Notice that all the previous statements are also valid in 3D (a pixel becomes a voxel).

5 FieldGML: The Field Geography Markup Language

Because of the current standards' shortcomings and weaknesses, as highlighted in the previous section, I propose an alternative to represent fields: FieldGML. It is an XML-based language based on GML, and it permits us to represent fields in 2D and 3D, although conceptually it can be easily extended to higher dimensions. Unlike current standards where there is a distinction between discrete and continuous fields/coverages, I argue in this paper that a field should always have one—and only one!—value for a given attribute at every location in the spatial domain (be this domain the surface of the Earth, a 3D volume, or even a 4D spatio-temporal hypercube). The concept of discrete coverage can be then removed, as it is misleading and it creates confusion among users.

5.1 A Field = Samples + Interpolation Rules

The principal idea behind FieldGML is that two things are needed to have a coverage: (i) a set of samples of the phenomenon, and (ii) an interpolation function to reconstruct the continuity of the phenomenon studied.

Samples.

By that it is meant what is referred to as 'discrete coverage' in ISO/OGC terms. It is any data that were collected to study the phenomenon:

1. a set of scattered points in 2D or 3D.
2. a set of lines, e.g. contour lines coming from a topographic map.
3. a set of scattered polygons to which one value is attached. Although this case is possible, I am not aware of any interpolation method that would take a set of polygons as input. It is nevertheless always possible to discretise each polygon into a set of points. Polyhedra in 3D are also considered samples.
4. a raster image coming from remote sensing or photogrammetry where the value of each pixel represents the temperature of the sea for instance.

Observe that a set of samples is simply a "normal" vector file or a grid (as defined in other ISO/OGC standards, e.g. in ISO (2003)), where each object is assigned a value for a common attribute.

Interpolation Method.

The set of rules used to reconstruct the field from samples can take many forms. Interpolation methods are rather difficult to categorise because they are based on different paradigms, and some methods fall into more than one category. No attempts will be made here to introduce categories (see Mitas and Mitasova (1999) and Watson (1992) for that), but what should be kept in mind is that although several different interpolation methods are used in GIS and that several publications advocate the use of "better" methods, the current standards, while discussing a few methods, give no importance to interpolation and do not permit the use of many of the known methods.

Storing explicitly the interpolation method, as FieldGML is doing, is efficient in practice as only a few parameters have to be stored. Finding the appropriate values for interpolation parameters is a difficult and time-consuming task, as the user must have a good understanding of the spatial distribution of the objects in the set of samples, and of the details of the method. A vivid example is Kriging (Oliver and Webster, 1990), with which experienced users can obtain very good results, but which also leaves newcomers clueless with its many parameters and options. Using Kriging with the appropriate parameters leads to a result that has statistically minimum variance, however, simply using the default values for the parameters will most likely lead to unreliable results. Thus, if we leave the job of modelling the datasets and deriving the interpolation parameters to specialists, the users would not have to worry about these anymore. This is one of FieldGML's main benefits.

5.2 Abstract and Implementation Specifications

Figure 2 shows the class diagram of FieldGML (for the main classes). To ensure that a FieldGML file respects the rules in the model (and that it is therefore 'valid'), a GML application schema has also been developed. What follows is an overview of the engineering decisions that were taken in order to develop FieldGML and its schema. I tried to use GML types as much as possible, but for practical reasons (e.g. simplicity of implementation and performances for processing files) new types also had to be defined. The full application schema is not described here (for lack of space), but it can be obtained on the website of FieldGML[5].

The first thing to notice is that a FieldGML type *Field* inherits directly from a GML feature, which means that it can use all the mechanisms already defined by the OGC to deal with metadata.

An important decision that was taken was not to use directly the GML implementation of *CV_DiscreteCoverage* for the set of samples, for the reasons described in Section 4.1. Instead, four new types were created: (i) scattered points (notice here that even if points are regularly spaced, this type

[5] www.gdmc.nl/ledoux/fieldgml.html

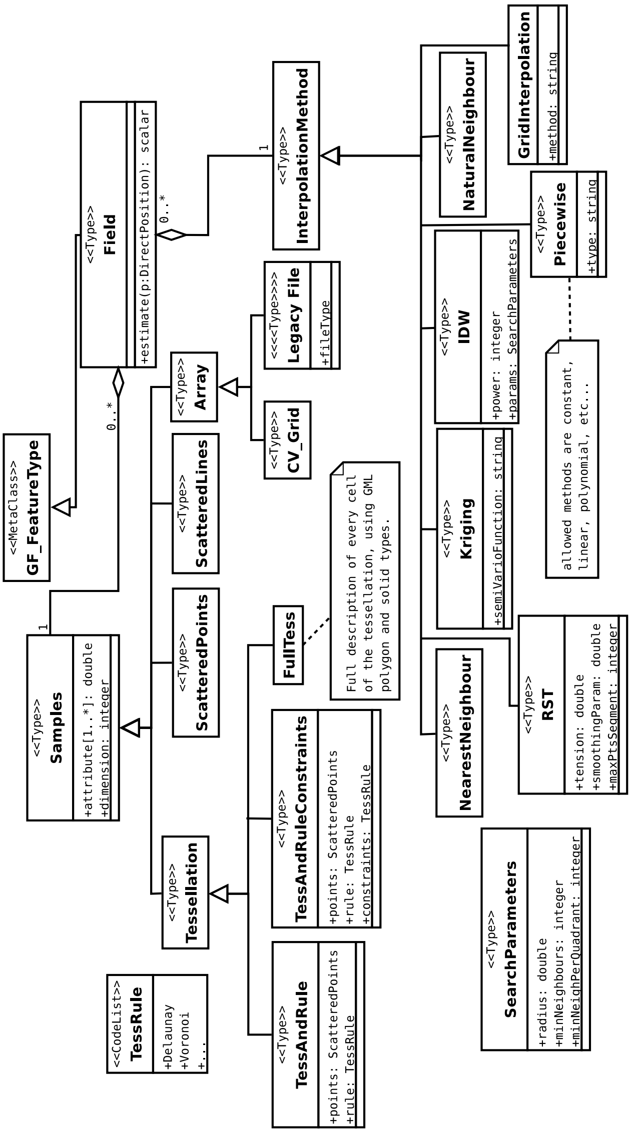


Fig. 2. UML diagram for the main classes of FieldGML.

can still be used); (ii) scattered lines; (iii) full tessellations; and (iv) arrays, which includes all the raster-based types. All these types inherit from `gml:AbstractGeometryType` (which means that mechanisms defined by GML for reference systems can be used), and GML types were used where possible (e.g. `gml:MultiPointType` for the scattered points). In addition, these types were extended so that an attribute (a scalar value) is attached to each object;

a version of the *CV_GeometryPairValue* was implemented, as in Cox (2007). Also, it should be noticed here that if a tessellation is needed for the interpolation (e.g. Delaunay triangulation for a piecewise interpolation) this structure need not be persistent: only the samples can be stored, and it is calculated on the fly. Constraints that a triangulation must follow can also be stored, but if it is impossible to define rules to automatically construct a tessellation (there has been human intervention in the construction) then the full tessellation must be stored. The type *FullTess* represents this full description (each cell is described); the GML types `gml:MultiSurface` and `gml:MultiSolid` can be used for that, albeit they are rather non-efficient in practice.

As is the case for GeoSciML (Woolf and Lowe, 2007), it was decided that 'legacy files' (i.e. raster formats used in commercial GISs) could be used directly without having to convert them to GML types (which are non-efficient and cumbersome to use in practice). It is however still possible to use *CV_Grid* as defined in ISO/OGC standards and implemented in GML (as a discrete coverage). Legacy files are simply referenced to by a pointer; the metadata about the file (georeferencing, pixel size, etc.) have however to be stored in the FieldGML file with GML types and/or attributes.

Interpolation methods play an important role in the FieldGML model, and several methods used in the GIS world have been listed. The list of methods in Figure 2 is by no means exhaustive as other ones can be easily be added if needed (which is a big advantage over current standards). It should also be noticed that all the methods listed are perfectly valid in 2D and 3D.

While space constraints does not permit to discuss the details of these methods, the manual of FieldGML will describe them and discuss what the parameters imply, and that for interpolation in 2D and 3D.

A few examples of the methods available in FieldGML:

Piecewise: a function is defined over each cell of a tessellation (usually constant or linear).

IDW: as described in Shepard (1968), it requires different parameters to define which points are involved in the interpolation at a given location (different criteria can be used), and also the power must be defined.

Kriging: while the modelling of a dataset is a difficult and time-consuming task, the output of the modelling (a function characterising the dependence between the attributes of any two samples that are at a given distance from each other) can be simply stored as a string. The parameters and the functions as defined in the program *gstat* (Pebesma and Wesseling, 1998) are used.

Natural neighbour: the basic method (Sibson, 1981) does not need any user-defined parameters, but it is possible to obtain a smoother interpolation if certain parameters are used (see Watson (1992)).

RST—regularized spline with tension: this method is available in the open-source GIS GRASS, and by storing a few parameters a field can be reconstructed from a set of samples (Mitasova and Mitas, 1993).

Grid interpolation: while a grid can be seen as a special case of scattered points, different methods optimised for grids have been developed. FieldGML implements a few of them, for instance bilinear and biquadratic. See Kidner et al. (1999) for a discussion in 2D, but these methods trivially generalise to higher dimensions.

In brief, when fields are represented with FieldGML, any kinds of fields can be defined. Observe also that all the ISO/OGC (sub)types can be mapped to a samples/interpolation in FieldGML (so there are no needs to define explicitly subtypes). For example, the *CV_TINCoverage* is based on a set of points, and the interpolation is a piecewise function (linear function inside each Delaunay triangle). Also, notice that even if a grid is the set of samples for a field, an interpolation method must be also defined (it can be for instance constant or bilinear inside each cell).

5.3 Prototype

To convert back and forth between FieldGML representations and the formats used in GIS and geoscience applications, a prototype was built. Currently it permits users to read a FieldGML file and output to different formats, and it is also possible to create a FieldGML file when a set of samples is already available. To output to a format used by commercial GISs, the user has to choose the spatial extent, the resolution of the grids (only grids are possible right now, although triangulation could be implemented in the future), and the format. The possible grid formats currently supported are the ones in the GDAL library[6] (in 2D), and netCDF (in 3D).

The prototype was developed with the Python programming language, and uses only open-source software. The interpolation methods described in the precedent section were implemented or their libraries were linked to the prototype. For instance, the program *gstat* (Pebesma and Wesseling, 1998) was used for Kriging, GRASS for RST, and CGAL[7] to create triangulations in 2D and 3D.

5.4 Discussion Over the Implementation

At this moment, the interpolation methods have been implemented in the prototype, but to favour interoperability, I plan to make use of the newly adopted OGC standards about web processing service (WPS) (OGC, 2007c), which defines how GIS operations can be performed over the Internet. The methods used by FieldGML would simply be available on a server, and a user would upload his FieldGML file, specify what representation is needed, and then he/she would get the file.

[6] The Geospatial Data Abstraction Library: `www.gdal.org`

[7] The computational geometry algorithms library: `www.cgal.org`

Also, it is interesting to observe that while FieldGML and VDS have very similar conceptual ideas, the implementations are totally different because of the way interoperability is tackled. The VDS approach was about having proprietary formats not directly accessible to users, who had to access data through common interfaces. While theoretically very sound, this was not the choice the GIS community picked, and now instead we have one language (GML) that can be used to represent any geographical dataset. While probably less efficient (GML is very verbose and complex), it offers more flexibility as anyone can read a FieldGML file and extract the original samples, while in the case of VDS you would have to have a piece of software implementing the interface. With the original dataset, the user can then choose another interpolation method, if needed.

6 Conclusions

The ultimate goal of a digital field representation is to reconstruct in a computer the continuity of a studied phenomenon, i.e. to be able to accurately estimate, or calculate, the value of the phenomenon at any location in a spatial domain. As an alternative to current standards for fields (i.e. the discrete coverage as implemented in GML), what has been proposed in this paper, a GML-based representation, is admittedly rather simple from a theoretical point of view, but yet it permits us to model every situation (and that in two and three dimensions), and it uses the types already defined in current standards (thus it is a step in the direction of interoperability). It is also more adapted than raster structures to the kind of datasets found in GIS-related applications, because it permits us to always keep the original data that were collected to study a phenomenon, and simply generate new representations that are adapted to a particular application. FieldGML was also designed with flexibility in mind, so that other interpolation methods and sample forms can be added.

While the use of FieldGML requires a rethinking from people who produce fields, the users need not be affected. Indeed, a potential user of FieldGML would simply obtain a field in the form of a FieldGML file, select the format and resolution of the output file, and carry on with his/her work as before. But when he/she would need to exchange the field with someone else, the shortcomings of raster structures would not arise.

Future works include the implementation of a WPS and the processing of very large datasets (which are common in GIS-related applications, especially in three dimensions). I also plan to extent FieldGML so that dynamic fields, and fields having a nominal scale of measurement, are handled.

Acknowledgments

This research was financially supported by the project “Internationale geostandaarden” (RGI-116) in the Netherlands, and by the EU project HUM-

BOLDT. I would also like to thank my colleagues Marian de Vries and Wilko Quak, and the excellent comments from the three anonymous reviewers, especially Mr or Mrs "Reviewer 3" who pointed out Andrej Včkovski's work on VDS.

References

Couclelis H (1992) People manipulate objects (but cultivate fields): Beyond the raster-vector debate in GIS. In AU Frank, I Campari, and U Formentini, editors, *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*, volume 639 of *Lecture Notes in Computer Science*, pages 65–77. Springer-Verlag.

Cox S (2007) GML encoding of discrete coverages (interleaved pattern). Open Geospatial Consortium inc. Document 06-188r1, version 0.2.0.

Fisher PF (1997) The pixel: A snare and a delusion. International Journal of Remote Sensing, 18(3):679–685.

Frank AU (1992) Spatial concepts, geometric data models, and geometric data structures. Computers & Geosciences, 18(4):409–417.

Gold CM and Edwards G (1992) The Voronoi spatial model: Two- and three-dimensional applications in image analysis. ITC Journal, 1:11–19.

Goodchild MF (1992) Geographical data modeling. Computers & Geosciences, 18(4):401–408.

Haklay M (2004) Map Calculus in GIS: A proposal and demonstration. International Journal of Geographical Information Science, 18(2):107–125.

ISO (2003) ISO 19107: Geographic information—Spatial schema. International Organization for Standardization.

ISO (2005) ISO 19123: Geographic information—Schema for coverage geometry and functions. International Organization for Standardization.

Kemp KK (1993) Environmental modeling with GIS: A strategy for dealing with spatial continuity. Technical Report 93-3, National Center for Geographic Information and Analysis, University of California, Santa Barbara, USA.

Kemp KK and Včkovski A (1998) Towards an ontology of fields. In *Proceedings 3rd International Conference on GeoComputation*. Bristol, UK.

Kidner D, Dorey M, and Smith D (1999) What's the point? Interpolation and extrapolation with a regular grid DEM. In *Proceedings 4th International Conference on GeoComputation*. Mary Washington College Fredericksburg, Virginia, USA.

Lake R (2000) Introduction to GML: Geography Markup Language. In *Proceedings W3C Workshop on Position Dependent Information Services*. Sophia Antipolis, France. Available at `http://www.w3.org/Mobile/posdep/GMLIntroduction.html`.

Ledoux H and Gold CM (2006) A Voronoi-based map algebra. In A Reidl, W Kainz, and G Elmes, editors, *Progress in Spatial Data Handling—12th International Symposium on Spatial Data Handling*, pages 117–131. Springer.

Lu CT, Dos Santos RF, Sripada LN, and Kou Y (2007) Advances in GML for Geospatial Applications. GeoInformatica, 11:131–157.

Mark DM (1975) Computer analysis of topography: A comparison of terrain storage methods. Geografiska Annaler, 57A(3–4):179–188.

Mitas L and Mitasova H (1999) Spatial interpolation. In PA Longley, MF Goodchild, DJ Maguire, and DW Rhind, editors, *Geographical Information Systems*, pages 481–492. John Wiley & Sons, second edition.

Mitasova H and Mitas L (1993) Interpolation by regularized spline with tension: I. Theory and implementation. Mathematical Geology, 25:641–655.

Nativi S, Caron J, Davies E, and Domenico B (2005) Design and implementation of netCDF markup language (NcML) and its GML-based extension (NcML-GML). Computers & Geosciences, 31(9):1104–1118.

OGC (2007a) Geography Markup Language (GML) Encoding Standard. Open Geospatial Consortium inc. Document 07-036, version 3.2.1.

OGC (2007b) Topic 6: Schema for coverage geometry and functions. Open Geospatial Consortium inc. Document 07-011, version 7.0.

OGC (2007c) Web Processing Service. Open Geospatial Consortium inc. Document 05-007r7, version 1.0.0.

Oliver MA and Webster R (1990) Kriging: A method of interpolation for geographical information systems. International Journal of Geographical Information Systems, 4:313–332.

Pebesma EJ and Wesseling CG (1998) Gstat: a program for geostatistical modelling, prediction and simulation. Computers & Geosciences, 24(1):17–31.

Peucker TK (1978) Data structures for digital terrain models: Discussion and comparison. In *Harvard Papers on Geographic Information Systems.* Harvard University Press.

Peuquet DJ (1984) A conceptual framework and comparison of spatial data models. Cartographica, 21(4):66–113.

Peuquet DJ, Smith B, and Brogaard B (1999) The ontology of fields: Report of a specialist meeting held under the auspices of the VARENIUS project. Technical report, National Center for Geographic Information and Analysis, Santa Barbara, USA.

Sen M and Duffy T (2005) GeoSciML: Development of a generic geoscience markup language. Computers & Geosciences, 31(9):1095–1103.

Shepard D (1968) A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings 23rd ACM National Conference*, pages 517—524.

Sibson R (1981) A brief description of natural neighbour interpolation. In V Barnett, editor, *Interpreting Multivariate Data*, pages 21–36. Wiley, New York, USA.

Stephan EM, Včkovski A, and Bucher F (1993) Virtual Data Set: An Approach for the Integration of Incompatible Data. In *Proceedings AutoCarto 11 Conference*, pages 93–102. Minneapolis, USA.

Včkovski A (1998) *Interoperable and Distributed Processing in GIS.* Taylor & Francis.

Včkovski A and Bucher F (1996) Virtual Data Sets—Smart Data for Environmental Applications. In *Proceedings 3rd International Conference/Workshop on Integrating GIS and Environmental Modeling.* Santa Fe, USA.

Watson DF (1992) *Contouring: A guide to the analysis and display of spatial data.* Pergamon Press, Oxford, UK.

Woolf A and Lowe D (2007) Climate Science Modelling Language Version 2—User's Manual. `http://ndg.badc.rl.ac.uk/csml/`.

Marine GIS: Progress in 3D Visualization for Dynamic GIS

Rafal Goralski, Christopher Gold

University of Glamorgan, Faculty of Advanced Technology
CF37 1DL Pontypridd, Wales UK
email: {rigorals; cmgold}@glam.ac.uk

Abstract. The paper presents the advancements in our work on 3D visualization for GIS. In (Gold et al 2004) we presented the idea of a dynamic GIS system that uses 3D graphics for data visualization. The system presented there – the Marine GIS – was at a preliminary stage, and was built mainly to show the potential of 3D gaming approach for GIS. Our work since then allowed us for much clearer, better and more comprehensive understanding of the needs and requirements for such a system, and consequently of the appropriate system design, together with additional real-life experience. During the implementation work we encountered many problems that we would not have expected before. This paper is a report of the process of the implementation of a real-time, dynamic 3D GIS for maritime purposes, and includes the final requirements, system design, description of some of the problems and the ways used to solve them. We believe that by sharing our experience we can help other researchers working towards building 3D GIS systems and tools, and that it is a contribution towards the future development of GIS.

Keywords: GIS, Dynamic, 3D, Visualization, Marine.

1 Introduction

Traditional land-based GIS systems are well known and widely used and appreciated. However there are areas where application of these is not practical and does not give satisfactory results. That is true for marine applications, where objects are far from being static and changes are happening in more than just two dimensions. Some marine applications such as analysis for Oceanography require real 3D data structures. Others like maritime navigation can be based on the simpler 2D equivalents. In both cases the common observation remains valid – they would benefit from the use of kinetic methods and data structures. The knowledge of the distinctness of the marine requirements led to emergence of a new specialized family of GIS systems, so called Marine GISs. An example application of Marine GISs is the use of chart plotters onboard ships and pleasure boats for maritime navigation and safety.

Maritime safety has a huge impact on the world economy and our everyday lives. The vast majority of world cargo is carried by sea by huge container-carriers. Cargo ships carrying hazardous loads pose serious threats to the environment. The disaster and damage caused in the event of major sea collisions can be difficult and costly to deal with (Ward et al 1999). Studies on the classification and reasons for marine accidents show that human error is the most common cause of problems (Talley 2006). To address this researchers like Goulielmos and Tzannatos proposed an information management and decision support system as early as (1997). Since then sea authorities have introduced many standards and new technologies that support navigation and communication, e.g. ECDIS - Electronic Chart Display and Information Systems (IMO 2004; Ward et al 1999) and AIS – Automatic Identification System (IMO 2004). ECIDS is an interactive 2D map that uses ENC (Electronic Navigational Charts) data and is the only marine law approved equivalent to official paper nautical charts.

In our study we took a broad look at the problem of maritime safety in order to identify the areas where application of a Marine GIS could lead to the most significant improvement. Then a new type of GIS system for maritime navigation safety was proposed. The idea starts from ECDIS, but the system takes advantage of the newest developments in computer graphics and GIS technologies, and is aimed at tackling the main cause of marine accidents – human errors – by providing navigational aid and decision support to mariners. The goal of our project is to produce and demonstrate an on-board 3D real-time decision support system for marine navigation that will show a possible future of ECDIS.

2 System requirements

The current stage of our project involved an extensive business analysis to be performed in order to determine the required functionality of the target system. As a result of this process, that included analysis of related software and publications, numerous meetings with various specialists from the marine research, authorities and industry, and gathering feedback to our presentations, we decided to implement the following functionalities, presented here as a requirements list. This is also the list of differences / improvements to the preliminary version of Marine GIS. The majority of the points on the list have been already implemented but some have yet to be completed.

- *ENC data.* ENC charts (IHO 2000) are the only legal electronic data format that can be used as a replacement of official paper charts. It is also the format of data used by ECDIS. As such if our system is to be considered a real navigational aid and potential path of ECDIS evolution, it has to work with this format of data.
- *Real world geographical coordinates.* The system had to be based on the real world geographical coordinates. These are also used as the internal format in which location of terrain points, ships and navigational objects are stored. The geographical coordinates are dynamically converted to the model equivalents for drawing, which allows us to use different map projections.
- *Real-time and simulation.* The system had to allow for real-time display of the host ship and its surroundings, but also for recording and replay of the pre-recorded data. This allows its use as a real-time navigational aid as well as a tool for analysis of historical data.
- *Open standards for real-time data.* It had to be able to display the real-time data of ships from many sources, in order to broaden the potential range of the system applications.
- *On-board equipment integration.* As in ECDIS the on-board navigational aid has to integrate with ship's on-board equipment. This was done through implementation of the NMEA protocol and included GPS for the position of the host ship and AIS for positions of the surrounding ships. Open and elastic means of implementation allow for relatively easy integration with different instruments such as ARPA (radar), speedometer, etc.
- *Decision support – external systems integration.* In addition to the open standards for display of real-time data the system was fitted with open standards of communication with external decision support logic systems. This was done especially to allow integration of Tractable

Role-based Agent for Navigation Systems – TRANS – intelligent navigation system that helps involved parties to negotiate the rules of way automatically in real-time. TRANS has been developed at the French Naval Academy in Brest (Fournier et al 2003).

- *Improved User Interface.* The Graphical User Interface (GUI) had to be redesigned to allow real work on-board a ship. The interface had to accommodate all new functionalities, but was also simplified in order to ease its usage in cramped deck conditions, made more readable, and allow for customization to account preferences of a user.
- *Improved manipulation.* The manipulation mechanism also has been improved in order to make it more simple and robust. This includes introduction of limits of camera movements, translations and rotations so the displayed situation always remains visible. We also integrated an additional 3D controller, SpaceNavigator (Fig. 1) produced by Logitech and 3Dconnexion. The system can be manipulated using a mouse in two modes, one for rotation, zooming, translation of the scene (when the mouse cursor is hidden), and another for queries, selection, etc. Usage of additional 3D manipulation device, which is controlled by the left hand for manipulation of the scene while a mouse can be simultaneously used for navigational operations, simplifies the manipulation process and makes it more effective, natural and intuitive.

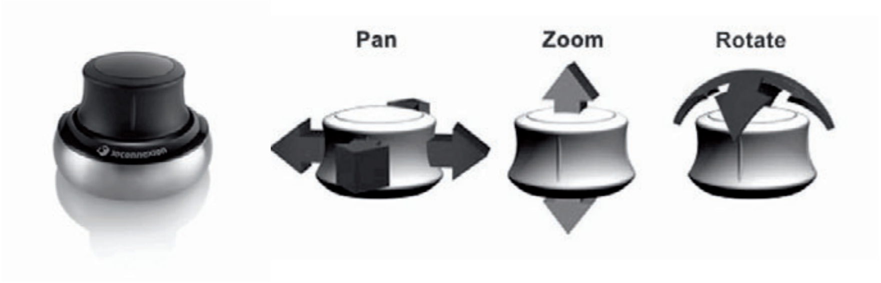


Fig. 1. 3D manipulator (courtesy of 3Dconnexion)

- *Custom updates.* The new system allows for custom updates of the chart by the user. It includes adding notes to any object, adding objects from predefined objects library (all standard navigational objects) as well as adding custom objects with properties and notes, which can be retrieved by a query.
- *Queries.* As in the preliminary version of the system every object can be queried in order to retrieve its properties and attached notes. The result of a query provides all relevant attributes of a navigational object

encoded in the ENC chart, or all AIS properties of a ship. Also custom objects can be queried.

- *Collision prediction/avoidance.* Collision prediction is performed based on the underlying kinetic Quad-Edge data structure (Gold 1999), and on calculation of the Closest Point of Approach (CPA) and Time of Closest Point of Approach (TCPA), with the use of simple physics of motion described below. Possible dangers are graphically illustrated to alert the navigator and provide him with details of the predicted danger. To make assessment of the situation easier, as in 2D ECDIS, directions and speeds of moving objects are illustrated with a speed vector visualization (the vector shows the direction of motion, and its length denotes the speed).
- *Navigation support.* Several navigation-related features were introduced. These include display of the compass and other host ship data, several options of measurement of angles and distances in general view as well as ship-related view modes – useful for heads-up display, and the possibility to assess the scene from any user-selected location.
- *Simple physics.* To allow for better simulation of ships' motion a library of simple physics was developed. It allows for simulation of various types of motion, including the uniform linear motion, uniform motion with turn, uniformly accelerated/retarded motion and a rotary motion.
- *Advanced animation.* The animation mechanism was completely redesigned to allow for constant speed of movement within the animated model. One of the problems of the previous version was that the speed of animation was dependant on the number of frames per second (FPS) that a machine could draw. This caused a lot of problems starting from different simulation speed on different machines, to variations of the speed depending on the CPU usage. The current mechanism allows for keeping constant speed of motion in spite of variations of the drawing speed. This allows for simulation and replay of pre-recorded data associated with real-world time, as well as for prediction of the future positions of objects that are updated in discrete periods of time, such as ships detected by the AIS transponder. Because the moving ships, depending on their speed, can be updated as rarely as every 3 minutes, the prediction allows for smooth animation and estimation of the current position of ships at any given time, based on their speed, direction and rate of turn (ROT). The prediction utilises the simple physics described above.
- *Better visualization.* Visualization of the terrain has been improved to allow easier orientation. Two main improvements are use of colour shading to illustrate the height / depth of the terrain, and the display of a

2D mini map in the right bottom corner of the screen. The default colours of the terrain were picked to match a typical navigational chart, and are produced automatically using GL Shading Language, without a need to prepare a dedicated texture. The colours can be customized by the user. The mini map displays a birds- perspective view of the whole area, with the position of the own ship and other ships in range. It can be switched on and off at any time.

- *Better symbology.* Further improvement to visualization is the introduction of new symbology, designed especially for 3D based on S57 Navigational Objects Catalogue and navigational charts standard symbols. The development has been done under our supervision by two cadets of the French Naval Academy in Brest, as a part of their Final Project (Busset and Fournier 2008). A wide set of the most relevant objects from the S57 Catalogue were picked and developed. The system's display engine was improved to allow for additional visualization tweaks to make sure that the objects are always easily readable. This includes rotation of the objects with the rotation of the scene, so they always face the observer, and scale adjustments. Also libraries of landmarks and buildings, as well as ship types for the AIS display, were developed (Fig. 2, Fig. 3).

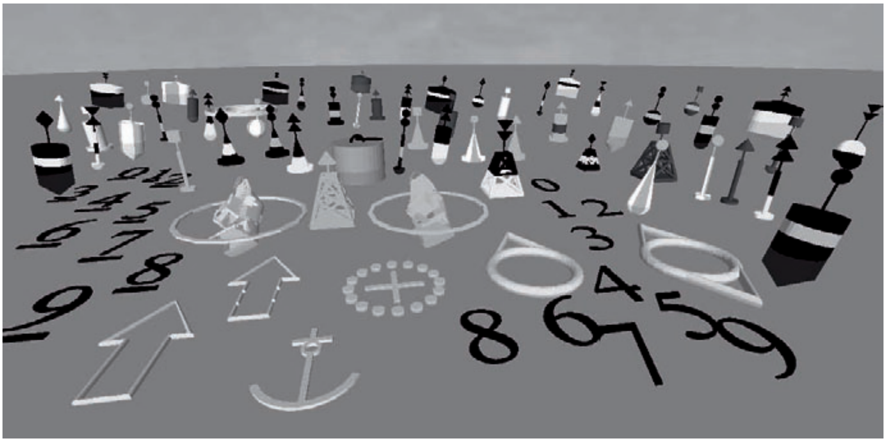


Fig. 2. A selection of Marine GIS S57 navigational objects

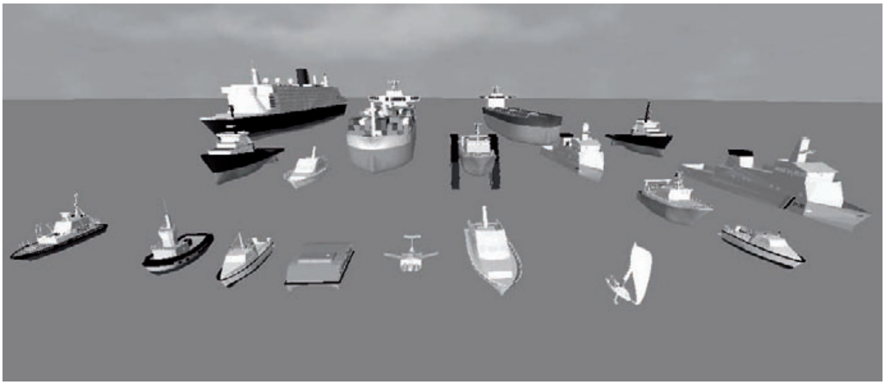


Fig. 3. Marine GIS ships library

3 System design

The list of requirements was presented in the previous section. In this section we will take a closer look at some of the described features from a more technical perspective. Some of the implementation details including related problems and solutions used are given.

3.1 GIS 3D Graphical Engine

The system is based on the graphical engine, also called the Scene Graph. The engine is responsible for storage, management and drawing of all graphical objects within the system. As described in (Gold et al 2004) we decided to use our own Graphical Engine, which was developed in Delphi using object-oriented design and OpenGL. Since that time the engine has been modified and extended, and redesigned to improve the efficiency and accommodate new features of the system and new capabilities of OpenGL 2.0, such as GLSL. We call it a GIS Graphical Engine because it is different both from the types of engines used for gaming and from CAD systems. It is less oriented to graphical display optimisation, speed and reality than the gaming scene graphs, and less to precision of manipulation and modelling than the CAD ones. Its strength is in flexible object-oriented design, relative simplicity and built in support for typical GIS operations, such as queries, work with spatial data, measurements, geographical projections, etc. The fact that the whole code has been developed by us gives us also a great level of understanding and control and power to share our work with other researchers without worrying about possible violation of licence agreements. The scene graph stores a tree of all objects within the scene (ships, navigational objects, terrain, water surface, sky, etc.) as well as cameras, lights with their relative positions within the hierarchy, and also OpenGL settings and shaders. It manages manipulation, selection, queries and animation mechanisms and accommodates the changes made to objects. Animation and physics mechanisms are built-in to allow for simulation of objects' motion. Constant speed of motion is kept based on the animation speed, as all objects are notified about the time that has passed since the last animation event, and move accordingly.

3.2 ENC Reader

This module is responsible for reading charts in the ENC format. It is based on the GDAL (Geospatial Data Abstraction Library) which is an open-source project managed by the Open Source Geospatial Foundation, and more specifically on its OGR Simple Features Library. We used the ANSI C DLL library provided for OGR and have written Delphi headers that allow the use of its functions within our system. The reader performs several operations related to the data input, such as reading navigational objects of a specified type, determining the area boundaries, creating a terrain model based on contour lines and soundings, etc. Because of problems with the stability of the ANSI C library and its frequent memory crashes we have to read all data in small portions when an ENC chart is loaded and store them in the system memory, rather then keeping the GDAL data source open and reading the data when needed. This attitude guarantees us the stability of our system and works well. We had hoped that the terrain model generated directly from the ENC chart would be good enough for direct use in the model, but despite several data optimisation mechanisms at read time built in the ENC Reader, for the reasons described in the 'Display of terrain' section it turned out to be not feasible at the current stage.

3.3 Geo converter

Geo Converter is an abstract class used for conversion of the geographical coordinates to model units (metres) when a model is being loaded. Several descendant classes implementing selected map projections have been developed. A particular type of descendant class used is loaded based on a chosen map projection using the polymorphism mechanism. This approach allows for implementation of various additional map projections and further extension of the system when needed.

3.4 AIS module

To provide real-time data of ships an interface between the ‘Marine GIS’ and the AIS has been developed. The integration was performed in two steps. First an external specialized library for reading AIS messages was designed and implemented using UML modelling tools and an elastic object-oriented design that allows for future expansion to other on-board NMEA devices. Then the library was incorporated into the ‘Marine GIS’ 3D interface. The library allows for real-time tracking and recording of the AIS data, as well as for its later playback for test and simulation purposes. Several safety features related to the AIS specifically were implemented and tested. The integration with the NMEA multiplexer allowed for incorporation of the GPS data of the observer’s own position.

3.5 Display of 3D models

The navigational objects, ships, landmarks and buildings have been modelled using an external tool in the 3DS format, and are kept in external files. The model files are associated with a navigational sign of a given type and attributes based on a configuration file. This approach allows for simple modification and extension of the models library. For optimisation reasons only one 3DS model of any single type is loaded (only when needed) and stored in the memory as an OpenGL call list. Instances are then drawn several times for all objects associated with it in all required positions.

3.6 Display of terrain

The effective reconstruction of a 3D model of terrain from 2D data is problematic. This is typical and very likely to be encountered by other researches working in the field of extension of a 2D GIS into 3D. The problem has been originally explained in (Dakowicz and Gold 2003).

The terrain model is constructed based on the underlying ENC chart. Several layers are used and intelligent re-sampling algorithms are used to produce data of the required density and quality. Then the original model (Fig. 4) is enriched with slope and valley reconstruction algorithms, based on the skeleton and crust properties of the Voronoi mesh used for storage of the model. However not everything can be done automatically. This is due to the original quality of the digitized chart data. The density of sampling varies, but that is solved by reading optimisation. Contour lines are often broken, and sometimes several contour lines of different levels are drawn at the same location. Manual model modification is needed in order to close all contour lines. The final enriched terrain model is stored as a file with simple points. The points are stored as geographical locations with associated height level in metres. When the model is loaded points are read and added to the Voronoi mesh using an incremental local update algorithm. The final model is shaded to resemble a typical chart using a GLSL shader provided with predefined colour values and minimal and maximal height. A simplified model used for display of the 2D mini map is generated automatically from the ENC file.

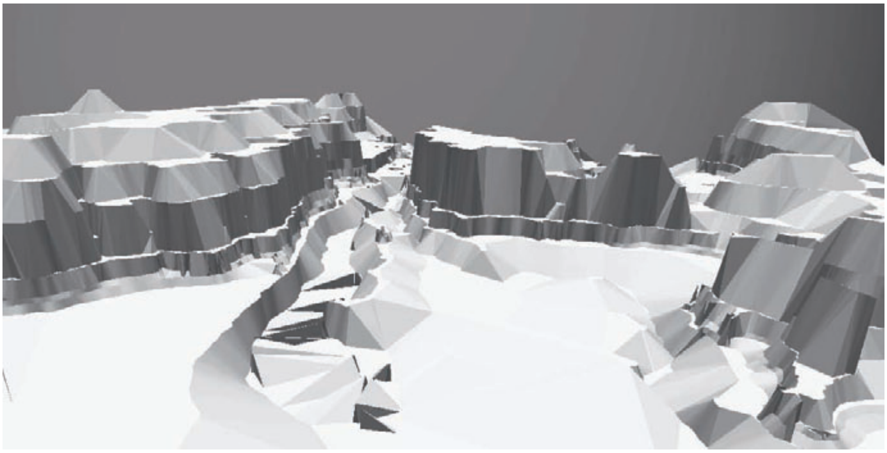


Fig. 4. Model of terrain generated directly from an ENC chart – needs additional semi-manual treatment for reconstruction of slopes and valleys

3.7 Data structures

The system uses two types of data: static and real-time/kinetic. Static data are obtained from ENC charts, and include the terrain model, landmarks and navigational signs. These can be modified and further extended by the user, including addition of custom objects and buildings. Real-time/kinetic data are provided by AIS transponder, or by a file with pre-recorded data of moving ships.

An appropriate data structure to maintain the real-time locations of the ships and other navigational objects had to be developed. We decided to use the “Quad-Edge” structure in which the points of Voronoi diagram (VD) represent marine vessels. As the ships move the positions of points are updated in real-time to reflect the current locations of ships. The moving-points algorithm (Gold 1999) is used for local updates of the structure.

Shoreline points were calculated from the intersection of the triangulated terrain model with the sea surface. These were also incorporated within the kinetic Voronoi diagram layer representing the ships, expressing the neighbourhood relations on the sea surface, and this was used for collision detection and avoidance. The kinetic Voronoi diagram layer of our system is shown on the Figure 5.

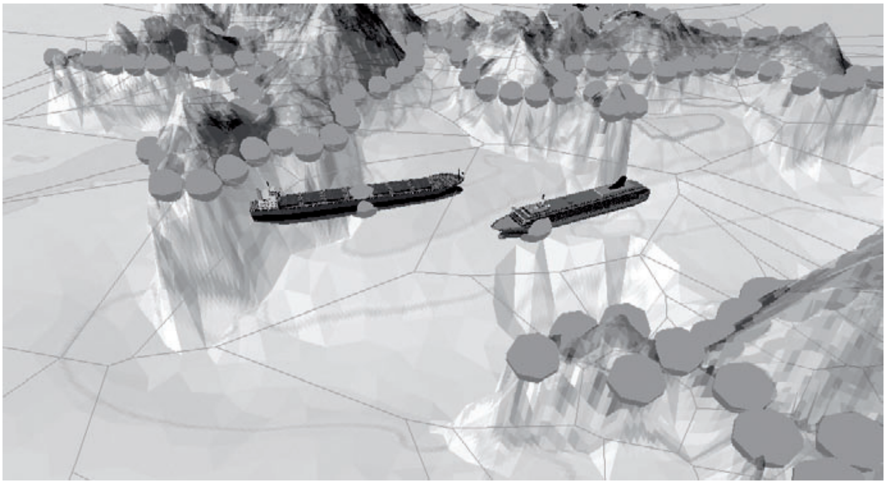


Fig. 5. The incorporated kinetic VD for maintenance of spatial relationships and collision detection

3.8 External data sources

The implementation of external data sources for both real-time data display and external decision support and navigation logic systems has been developed using the Listener / Connectors architecture. The system has been fitted with a listener that keeps an open socket and reads incoming messages. The messages have to have the correct format required by the Marine GIS. They can be delivered by many connectors that know the format. The connectors can be written for different external systems, using any technology, and their task is to translate communications between the external systems and Marine GIS. This has been illustrated on the Figure 6 in the next section.

4 Other applications

The Marine GIS at this stage is mainly a real-time navigation aid system. But with its AIS replay capabilities it can be also used for analysis of the pre-recorded AIS data, for example in a situation of collision. The system with its open design can also be used as a platform for display of different sources of external data. Currently we are working towards two applications, one for display of AIS data broadcast on the Internet, another for effective and attractive real-time visualization of regattas which take place every spring at the French Naval Academy, l'Ecole Navale in Brest. Figure 6 illustrates the usage of dedicated external connectors to accommodate these applications.

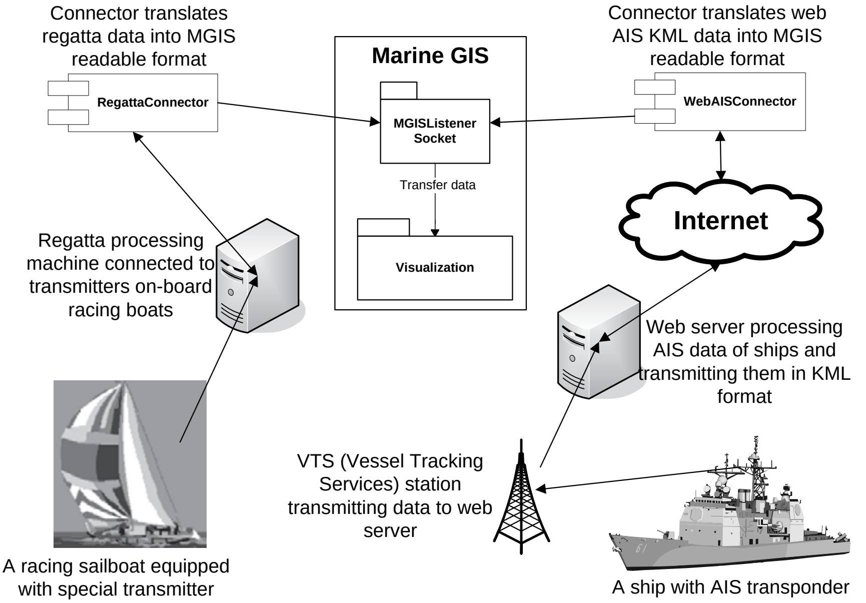

Connector translates regatta data into MGIS readable format
RegattaConnector
Marine GIS
MGISListener Socket
Transfer data
Visualization
Connector translates web AIS KML data into MGIS readable format
WebAISConnector
Internet
Regatta processing machine connected to transmitters on-board racing boats
Web server processing AIS data of ships and transmitting them in KML format
VTS (Vessel Tracking Services) station transmitting data to web server
A racing sailboat equipped with special transmitter
A ship with AIS transponder

Fig. 6. Use of Listener and Connectors for integration with external data sources

Implementation of additional simulation capabilities would allow for its use for navigator training. With implementation of multiplayer network communication it could be used as an intelligent educational game where several navigators could practice together leading different ships.

5 Future work and possibilities

Our goal at the current stage of the project is to produce a system that will be usable and convincing as a proof-of-concept. We do not aim at implementing all functionalities of a commercial quality product. However we are aware of numerous improvements that could be introduced. Below is our “wish list” for functionalities of the final system.

- *Incorporation of land data.* ENC charts do not contain enough data of the surrounding land, especially further from the sea. The land coverage is very limited. It would be good to incorporate some additional source of data for the reconstruction of land.
- *Animation of navigational lights.* Implementation of the patterns of navigational lights would be a very useful feature and a great help in the night time conditions, allowing navigators to match the lights they observe from the deck with the lights within the system. Improvement of the heads-up display for the night or bad weather conditions would be useful.
- *Improvement of collision detection algorithms.* Current algorithms are aimed at demonstration of the idea. These could be improved with more advanced prediction and collision avoidance logic.
- *Automatic terrain model generation.* Although it would be not easy, it should be possible to allow for fully automatic good quality terrain model generation directly from the ENC chart. That would be necessary for the system to be usable at any place on the globe, where ENC coverage is available. Currently the terrain model generated automatically, though still usable for navigation, is rough and looks far from the real world. The slope generation techniques would need to be improved in combination with the ENC reading algorithm to allow for automatic terrain improvement without the need for manual modifications.
- *Terrain display optimisation.* Currently the whole model of terrain and bathymetry for a given ENC range is read and displayed in the system. This limits the level of displayed detail, because the number of triangles that a machine is capable of displaying at a satisfactory FPS rate is limited. However the level of detail only matches the density of ENC data. Use of intelligent paging and Level of Detail (LOD) algorithms would allow for much better looking representation of terrain that is close to the observer.
- *Automatic loading of relevant ENC files on the move. Combination of several ENC files of different types for better quality.* Currently only one ENC chart is read, and thus the navigation is limited to the chosen region. Future versions of the system could read adjacent ENC maps as

needed when the ship moves towards the edge of the chart area. The system could also combine several charts for one bigger area, as very often there are additional ENC charts of better quality for more important sub-regions of the chart, such as ports, etc.

- *Real-time chart updates.* One of the important properties of the data structure used in the system is that it is capable of being updated in real-time. This could be used for real-time updates of the model, based on chart updates distributed with any technology, for example as Notice to Mariners.
- *Extend physics (new motion types, inertia).* The current physics of the system, though suitable for present use, is limited. It could be extended for better simulation of ships by introduction of curvilinear motion, inertia, etc.
- *Automatic simulation of tides.* At present the level of water surface can be set to any value in order to simulate tides. This could be done automatically within the animation mechanism to match the changes of tides over time, based on tides tables for a given region.
- *More UI customization: predefined and custom user profiles.* The user interface could be further improved with introduction of predefined and customizable user profiles. Such profiles could allow the user to choose his preferred colours, types of data for display, parameters for collision alerts, safe depth, etc.
- *Weather profiles.* Several typical predefined weather profiles, with different settings of light, clouds, rain, and fog could be provided.
- *Dedicated hardware.* As it is the case for ECDIS the system could be deployed on dedicated hardware, rather than a PC laptop. Such hardware would be optimised for the purpose, moisture sealed, more robust and compact, prepared to be integrated with on-board devices and safely mounted, and fitted with manipulation devices.
- *Real 3D.* This last point isn’t a part of our production-stage system “wish list”. It is a huge area of research on its own. But with usage of real 3D data the system could be used for simulation of submerged vessels, such as submarines, ROVs, AUVs, shoals of fish, etc.

6 Conclusions

We believe that the ‘Marine GIS’ will be a significant improvement to the problem of maritime safety. However it is currently in the development phase and its usability and value have yet to be verified. Comprehensive tests onboard real marine vessels of different classes are planned, after the expected completion date in 2008.

Although this paper presents advancements in development of our specific marine navigation safety system, we believe that the experience gathered goes far beyond this specific area and applies to other possible 3D GIS systems. As such we believe that this report can be of benefit to other researchers who elect to work with extension of traditional 2D GIS towards incorporation of the 3D visualization.

7 References

Busset J, Fournier P (2008) Design of 3D models for the ‘Marine GIS’ 3D navigational aid. (Final project report, Ecole navale, Brest)

Dakowicz M, Gold CM (2003) Extracting Meaningful Slopes from Terrain Contours. International Journal of Computational Geometry and Applications 13:339-357

Fournier S, Brocarel D, Devogele T, Claramunt C (2003) TRANS: A Tractable Role-based Agent Prototype for Concurrent Navigation Systems. European Workshop on Multi-Agent System (EUMAS), Oxford

Gold CM (1999) An Algorithmic Approach to Marine GIS. In: Wright DJ, Bartlett D (eds) Marine and Coastal Geographic Information Systems. Taylor & Francis, London, pp 37-52

Gold CM, Chau M, Dzieszko M, Goralski R (2004) 3D geographic visualization: the Marine GIS. In: Fisher P (ed) Developments in Spatial Data Handling. Springer, Berlin, pp 17-28

Goulielmos A, Tzannatos E (1997) Management information system for the promotion of safety in shipping. Journal of Disaster Prevention and Management 6(4):252-262

International Hydrographic Bureau (2000) IHO transfer standard for digital hydrographic data edition 3.0. Special publication No. 57

International Maritime Organization (2004) SOLAS: International Convention of the Safety of Life at Sea, 1974 - Consolidated Edition. International Maritime Organization Publishing

Talley WK (2006) Determinants of the Severity of Passenger Vessel Accidents, Journal of Maritime Policy & Management 33:173-186

Ward R, Roberts C, Furness R (1999) Electronic Chart Display and Information Systems (ECDIS): State-of-the-Art in Nautical Charting. In: Wright DJ, Bartlett D (eds) Marine and Coastal Geographical Information Systems. Taylor & Francis, London, pp 149-161

The IGN-E Case: Integrating Through a Hidden Ontology

A Gómez-Pérez [1], JA Ramos [1], A Rodríguez-Pascual [2],
LM Vilches-Blázquez [2]

[1] Ontology Engineering Group – UPM, Spain
email: {asun, jarg}@fi.upm .es

[2] National Geographic Institute of Spain (IGN-E), Spain
email: {afrodriguez, lmvilches}@fomento.es

Abstract

National Geographic Institute of Spain (IGN-E) wanted to integrate its main information sources for building a common vocabulary reference and thus to manage the huge amount of information it held. The main problem of this integration is the great heterogeneity of data sources. The Ontology Engineering Group (OEG) is working with IGN-E to attain this objective in two phases: first, by creating automatically an ontology using the semantics of catalogues sections, and second, by discovering mappings automatically that can relate ontology concepts to database instances. So, these mappings are the instruments to break the syntactic, semantic and granularity heterogeneity gap. We have developed software for building a first ontology version and for discovering automatically mappings using techniques that take into account all types of heterogeneity. The ontology contains a set of extra-attributes which are identified in the building process. The ontology, called PhenomenOntology, will be reviewed by domain experts of IGN-E. The automatic mapping discovery will be also used for discovering new knowledge that will be added to the ontology. For increasing the usability and giving independence to different parts, the processes of each phase will be designed automatically and as upgradeable as possible.

Keywords: ontology creation, geographic information, feature catalogues, mapping discovery, ontology-database mapping, heterogeneity

1. Introduction

National Geographic Institute of Spain (IGN-E) wanted to integrate its information sources for building a common vocabulary reference and thus to manage the huge amount of information it held. The main reason was to offer a unified national vocabulary to different Geographical Information (GI) producers, which have different interest, necessities and work scale (national – regional – local).

IGN-E has four main databases that correspond to different scales: Conciso Gazetteer (NC) (1:1,000,000), National Geographic Gazetteer (NGN) (1:50,000), Numerical Cartographic Database (BCN200) (1:200,000) and Numerical Cartographic Database (BCN25) (1:25,000). Each database has a different constant table to store represented features and their attributes. These databases are maintained separately and present great heterogeneity in different issues as we will show below.

The active collaboration between IGN-E and OEG (Ontology Engineering Group) of UPM (Universidad Politécnica de Madrid) aims to create an integration framework for maintaining the current databases. This framework will be designed and, in the future, added to databases, and it should be built following the most automatic processes in order to solve as best as possible the heterogeneity problems that may arise.

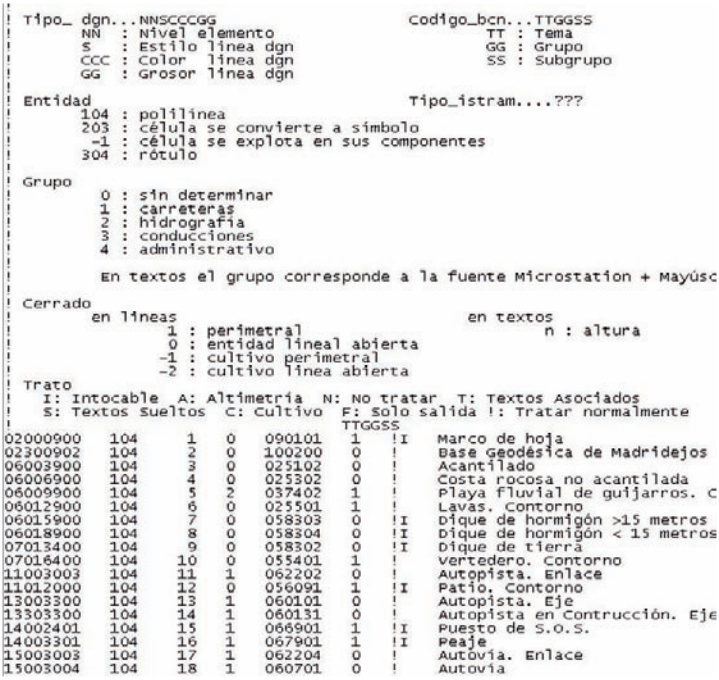
This paper is organized as follows. Section 2 presents the history and characteristics of the current catalogues involved in this integration work. Section 3 describes the problems found and the approach proposed to solve them. Section 4 focuses on heterogeneity types and levels. Section 5 covers the steps followed to integrate the different feature catalogues. Section 6 shows in detail the automatic ontology creation. Section 7 deals with the automatic mapping discovery. Finally, section 8 provides some brief conclusions and discusses some future lines of work.

2. Existing catalogues

The IGN-E has various databases and feature catalogues, but this work focuses on four main data sources: two Numerical Cartographic Database (BCN25 and BCN200) and two gazetteers (Concise Gazetteer and National Geographic Gazetteer).

With regard to the two Numerical Cartographic Databases, we can point out that they are considered as feature catalogues. This type of catalogue presents the abstraction of reality, represented in one or more sets of geographic data, as a defined classification of phenomena. It defines the feature type, its operations, attributes, and associations represented in geographic data. This type of catalogue is indispensable to turning data into usable information (ISO 19110). Next we provide some details of these data sources.

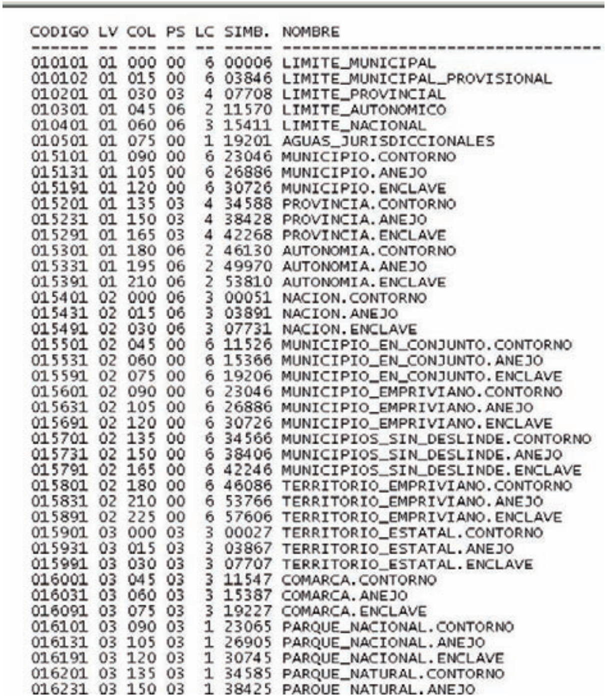
BCN25 was designed as a derived product from the National Topographic Map, which was created at a 1:25,000 scale (MTN25) in 1997, whereas MTN was created at a 1:50,000 scale (MTN50) in 1870. This long and hard project culminated at the end of the 1960s leaving behind great many changes due mainly to the continuous evolution that affected cartographic techniques during those years. From 1975 onwards, the updating of maps was carried out simultaneously with the production of a new series of maps at a1:25,000 scale (MTN25) with the aim of complementing MTN50 with some areas of special interest. However, in the 1980s these new series became a national coverage project. This Numerical Cartographic Database (BCN25) was built to obtain the 1:25,000 cartographic information that complies with the required data specifications exploited inside Geographic Information Systems (GIS) environments. Therefore, BCN25 contains essentially the same information than MTN25, though it has some additional geometric and topological properties, following a specific database oriented model and feature catalogue (Rodriguez 2005). The figure below shows a small part of the BCN25 feature catalogue.



```
!
! Tipo_ dgn...NNSCCCGG                      Codigo_bcn...TTGGSS
!       NN  : Nivel elemento                        TT : Tema
!       S   : Estilo linea dgn                      GG : Grupo
!       CCC : Color  linea dgn                      SS : Subgrupo
!       GG  : Grosor linea dgn
!
! Entidad                                   Tipo_istram....???
!       104 : polilinea
!       203 : célula se convierte a símbolo
!        -1 : célula se explota en sus componentes
!       304 : rótulo
!
! Grupo
!         0 : sin determinar
!         1 : carreteras
!         2 : hidrografía
!         3 : conducciones
!         4 : administrativo
!
!         En textos el grupo corresponde a la fuente Microstation + Mayúsc
!
! Cerrado
!        en lineas                                en textos
!                1 : perimetral                           n : altura
!                0 : entidad lineal abierta
!               -1 : cultivo perimetral
!               -2 : cultivo linea abierta
! Trato
!   I: Intocable  A: Altimetría  N: No tratar  T: Textos Asociados
!   S: Textos Sueltos  C: Cultivo  F: Solo salida !: Tratar normalmente
!                                  TTGGSS
02000900   104      1   0   090101   1   !I   Marco de hoja
02300902   104      2   0   100200   0   !    Base Geodésica de Madridejos
06003900   104      3   0   025102   0   !    Acantilado
06006900   104      4   0   025302   0   !    Costa rocosa no acantilada
06009900   104      5   2   037402   1   !    Playa fluvial de guijarros. C
06012900   104      6   0   025501   1   !    Lavas. Contorno
06015900   104      7   0   058303   0   !I   Dique de hormigón >15 metros
06018900   104      8   0   058304   0   !I   Dique de hormigón < 15 metros
07013400   104      9   0   058302   0   !I   Dique de tierra
07016400   104     10   0   055401   1   !    Vertedero. Contorno
11003003   104     11   1   062202   0   !    Autopista. Enlace
11012000   104     12   0   056091   1   !I   Patio. Contorno
13003300   104     13   1   060101   0   !    Autopista. Eje
13303300   104     14   1   060131   0   !    Autopista en Contrucción. Eje
14002401   104     15   1   066901   1   !I   Puesto de S.O.S.
14003301   104     16   1   067901   1   !I   Peaje
15003003   104     17   1   062204   0   !    Autovía. Enlace
15003004   104     18   1   060701   0   !    Autovía
```

Fig 1. Source: BCN25

On the other hand, the first version of the Numerical Cartographic Database (BCN200) at a1:200,000 scale was started in 1985. This work was developed through analogical map digitalisation of provincial maps at this scale (Sevilla 2006). Below, a part of the BCN200 feature catalogue is shown as an example of the layout that the catalogue presents.



```
CODIGO LV COL PS LC SIMB. NOMBRE
------ -- --- -- -- ----- --------------------------------
010101 01 000 00  6 00006 LIMITE_MUNICIPAL
010102 01 015 00  6 03846 LIMITE_MUNICIPAL_PROVISIONAL
010201 01 030 03  4 07708 LIMITE_PROVINCIAL
010301 01 045 06  2 11570 LIMITE_AUTONOMICO
010401 01 060 06  3 15411 LIMITE_NACIONAL
010501 01 075 00  1 19201 AGUAS_JURISDICCIONALES
015101 01 090 00  6 23046 MUNICIPIO.CONTORNO
015131 01 105 00  6 26886 MUNICIPIO.ANEJO
015191 01 120 00  6 30726 MUNICIPIO.ENCLAVE
015201 01 135 03  4 34588 PROVINCIA.CONTORNO
015231 01 150 03  4 38428 PROVINCIA.ANEJO
015291 01 165 03  4 42268 PROVINCIA.ENCLAVE
015301 01 180 06  2 46130 AUTONOMIA.CONTORNO
015331 01 195 06  2 49970 AUTONOMIA.ANEJO
015391 01 210 06  2 53810 AUTONOMIA.ENCLAVE
015401 02 000 06  3 00051 NACION.CONTORNO
015431 02 015 06  3 03891 NACION.ANEJO
015491 02 030 06  3 07731 NACION.ENCLAVE
015501 02 045 00  6 11526 MUNICIPIO_EN_CONJUNTO.CONTORNO
015531 02 060 00  6 15366 MUNICIPIO_EN_CONJUNTO.ANEJO
015591 02 075 00  6 19206 MUNICIPIO_EN_CONJUNTO.ENCLAVE
015601 02 090 00  6 23046 MUNICIPIO_EMPRIVIANO.CONTORNO
015631 02 105 00  6 26886 MUNICIPIO_EMPRIVIANO.ANEJO
015691 02 120 00  6 30726 MUNICIPIO_EMPRIVIANO.ENCLAVE
015701 02 135 00  6 34566 MUNICIPIOS_SIN_DESLINDE.CONTORNO
015731 02 150 00  6 38406 MUNICIPIOS_SIN_DESLINDE.ANEJO
015791 02 165 00  6 42246 MUNICIPIOS_SIN_DESLINDE.ENCLAVE
015801 02 180 00  6 46086 TERRITORIO_EMPRIVIANO.CONTORNO
015831 02 210 00  6 53766 TERRITORIO_EMPRIVIANO.ANEJO
015891 02 225 00  6 57606 TERRITORIO_EMPRIVIANO.ENCLAVE
015901 03 000 03  3 00027 TERRITORIO_ESTATAL.CONTORNO
015931 03 015 03  3 03867 TERRITORIO_ESTATAL.ANEJO
015991 03 030 03  3 07707 TERRITORIO_ESTATAL.ENCLAVE
016001 03 045 03  3 11547 COMARCA.CONTORNO
016031 03 060 03  3 15387 COMARCA.ANEJO
016091 03 075 03  3 19227 COMARCA.ENCLAVE
016101 03 090 03  1 23065 PARQUE_NACIONAL.CONTORNO
016131 03 105 03  1 26905 PARQUE_NACIONAL.ANEJO
016191 03 120 03  1 30745 PARQUE_NACIONAL.ENCLAVE
016201 03 135 03  1 34585 PARQUE_NATURAL.CONTORNO
016231 03 150 03  1 38425 PAROUE NATURAL.ANEJO
```

Fig. 2. Source: BCN200

The information contained in the two Numerical Cartographic Databases is structured in eight different topics (Administrative boundaries, Relief, Hydrography, Vegetation, Buildings, Communications, Piping lines and Toponymy). Each topic is coded with three pairs of digits: two for its topic, two for its group (part of homogeneous information structured in topics) and two for its subgroup (a stretch of geographic feature belongs to a group). These numbers describe and classify different features regardless of its location and spatial dimension. The following text box shows an example of how this information has been structured.

Topic 03: Hydrography
Group: 01 constant watercourse
Subgroup: 01 Watercourse symbolized with one line

Figure 1 and 2 have other codes to symbolize graphical characteristics, which are associated to Computer Aided Design (CAD) Systems. These digits belong to the following attributes: level (LV), colour (COL), weight (PS) and style (LC).

The BCN25 feature catalogue has a peculiarity. This peculiarity appears as attributes named DGN_Type (“Tipo_dgn”), Entity (“Entidad”) or Group (“Grupo”), which represent an alternative way for structuring these geographical features. These attributes are shown in figure 1 and are subdivided into different categories. An odd case is the classification proposed by the “Group” attribute, since it represents a brief, but similar classification within this catalogue. It is subdivided into five topics (No specified, Roads, Hydrography, Piping lines and Administrative).

As regards gazetteers, a widespread definition of this concept comes from (ISO 19112). This standard defines a gazetteer as a directory of instances of a class or classes of features that contain some information regarding position. Next, some of the main characteristics of the IGN-E gazetteers are described.

The National Geographic Gazetteer (see figure 3), also called Georeferenced DataBase or NOMGEO, has 460,000 entries, which belong to different features in Spanish, Galician, Catalan, Basque and Aranes (official languages of Spain). This gazetteer has 14 items with Universal Transverse Mercator (UTM) and geographic coordinates. Moreover, the gazetteer is the information source of the Web Service of the Spanish_Spatial Data Infrastructure (IDEE)[1].

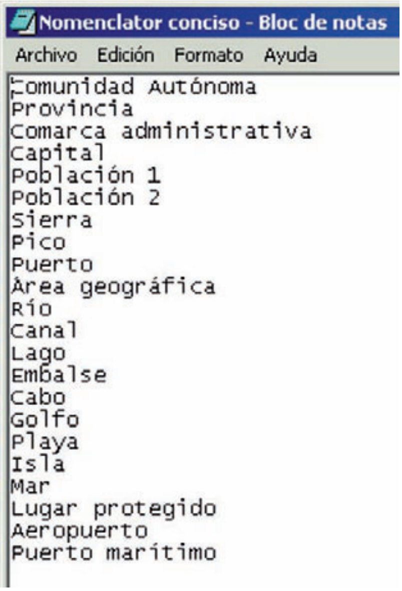
The Conciso Gazetteer (see Fig. 3) is a basic corpus of standardized toponyms created by the Spanish Geographical Names Commission. The first version has 3667 toponyms. This gazetteer agrees with the United Nations Conferences Recommendations on Geographic Names Normalization. Furthermore, the gazetteer has 17 items, of which some are mandatory: Name, Name Language, Group, Feature Type, Province, Autonomous Region, Latitude, Longitude, Map and Name Source; and others, optional: Variant, Variant Language, Before, Before Language, Municipality, Variant Source and Observations. These items are in accordance with the Spanish Gazetteer Model[2]. The Conciso Gazetteer has been created by the Spanish Geographical Names Commission. For further details, refer to (Nomenclátor Geográfico Conciso 2006).

Regarding previous data sources, there is, in some cases, a mix between geographical and cartographic concepts. On the one hand, in IGN-E gazetteers, only geographic concepts such as Reservoir (“Embalse”), Province (“Provincia”) Plain (“Llanura”), etc. appear, whereas in the BCN feature catalogues we can only find concepts specific to the GI domain (as province, river or dam) and some of their geometrical characteristics (as outline,

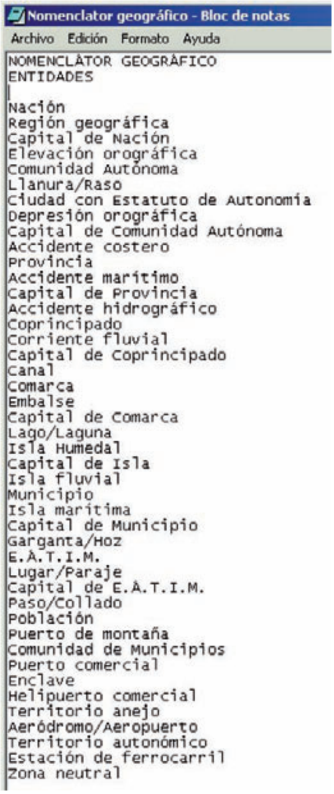
[1] http://www.idee.es

[2] http://www.idee.es/resources/recomendacionesCSG/MNEv1_2.pdf

axis, symbolized by one line and so on). This peculiarity will not have any influence on the development of our work.



```
Comunidad Autónoma
Provincia
Comarca administrativa
Capital
Población 1
Población 2
Sierra
Pico
Puerto
Área geográfica
Río
Canal
Lago
Embalse
Cabo
Golfo
Playa
Isla
Mar
Lugar protegido
Aeropuerto
Puerto marítimo
```



```
NOMENCLÀTOR GEOGRÀFICO
ENTIDADES

Nación
Región geográfica
Capital de Nación
Elevación orográfica
Comunidad Autónoma
Llanura/Raso
Ciudad con Estatuto de Autonomía
Depresión orográfica
Capital de Comunidad Autónoma
Accidente costero
Provincia
Accidente marítimo
Capital de Provincia
Accidente hidrográfico
Coprincipado
Corriente fluvial
Capital de Coprincipado
Canal
Comarca
Embalse
Capital de Comarca
Lago/Laguna
Isla Humedal
Capital de Isla
Isla fluvial
Municipio
Isla marítima
Capital de Municipio
Garganta/Hoz
E.À.T.I.M.
Lugar/Paraje
Capital de E.À.T.I.M.
Paso/Collado
Población
Puerto de montaña
Comunidad de Municipios
Puerto comercial
Enclave
Helipuerto comercial
Territorio anejo
Aeródromo/Aeropuerto
Territorio autonómico
Estación de ferrocarril
Zona neutral
```

Fig. 3. Feature Types of IGN-E Gazetteers Conciso Gazetteer (left) National Geographic (right)

3. Problems and the proposed approach

From a general viewpoint, GI is increasingly captured, managed and updated with variable levels of granularity, quality and structure by different cartographic agencies. In practice, this approach causes the building up of multiple sets of spatial databases with a great heterogeneity of feature catalogues and data models. That means a coexistence of a great variety of sources with different information, structure and semantics without a general harmonization framework. On the other hand, this heterogeneity combined with the sharing needs of miscellaneous users and information overlaps from different sources, causes several and important problems when linking similar features, to search, retrieve and exploit GI data (Vilches et al. 2007a).

From a narrow viewpoint, the most important concept for GI is the *feature* since the Open GeoSpatial Consortium (OGC 2003) has declared that a geographic feature is the starting point for modelling geospatial information. For that reason, the basic unit of GI within most models is the "feature", an abstraction of a real world phenomenon associated with a location relative to the Earth, about which data are collected, maintained and disseminated (ISO 19110). Features can include representations of a wide range of phenomena that can be located in time and space such as buildings, towns and villages or a geometric network, geo-referenced image, pixel or thematic layer. This means that, traditionally, a feature encapsulates in one entity all that a given domain considers about a single geographic phenomenon (Greenwood et al. 2003). From this point of view, we can observe that the heterogeneity associated to the feature term grows more because of the interests and necessities of different GI producers.

From an ontological perspective, no ontology has compiled the characteristics and peculiarities of Spain's geographic features. Up to now, there is only an hydrographical feature ontology of these characteristics, called *hydrOntology* (Vilches et al. 2007b). On the other hand, the use of standardized vocabularies, such as CORINE Land Cover[3], EuroGlobalMap[4] or EuroRegionalMap[5] involves an oversimplification of the existing complex reality because each GI producer (both national and local) has different feature catalogues (following their self-interests), which provokes that the overlaps between features are, quite often, not totally evident.

Taking into account these reasons, we decided to design an integration framework without having to reuse the information technology (standards, ontologies, feature catalogues, etc.) available. Fig. 4 shows the approach proposed for this integration, which implies using an ontology of features while keeping the current databases.

[3] http://dataservice.eea.europa.eu/dataservice/metadetails.asp?id=950

[4] http://www.eurogeographics.org/eng/03_projects_EGM_overview.asp

[5] http://www.eurogeographics.org/eng/03_projects_euroregionalmap.asp

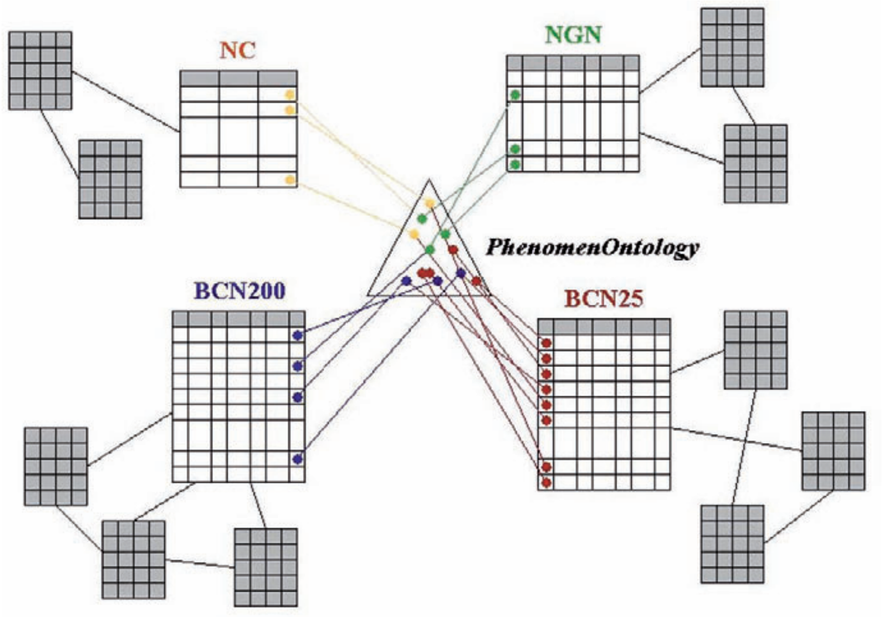


Fig. 4. Proposed approach

As can be seen, an ontology (called PhenomenOntology) conceptualises features, and their concepts will be mapped to their corresponding phenomenon row in each database.

For integrating GI by means of ontologies, some authors propose a multi-ontology system (Stuckenschmidt et al. 1999) (Hakimpour et al. 2001) (Fonseca et al. 2002). In such approach each organization integrates its information sources using a local ontology; other ontology integrates all the organization ontologies of the system. According to this approach, our integration framework will be an organization inside a multi-ontology system.

Following the classification of ontologies for the geographic world provided by (Fonseca et al. 2002), in our framework we will build a Phenomenological Domain Ontology (PDO) that we will name PhenomenOntology.

4. Heterogeneity

In the geographical information domain, any differences in data sources, disciplines, tools and repositories can cause heterogeneity (Alonso et al.

1994). Next, we describe the different approaches that tackle heterogeneity problems.

In (Bishr 1998), three different heterogeneity types (semantic, schematic and syntactic) are distinguished. First, semantic heterogeneity is usually the source of most of data sharing problems. This occurs because of the variation of models of the different disciplines and necessities, though geographical features are likely to share a common interest. Heterogeneity is subdivided into cognitive heterogeneity and naming heterogeneity. Cognitive heterogeneity is frequent when there is not a common base of definitions for the common features of different catalogues or databases, whereas naming heterogeneity is due to semantically alike features that might be named differently. For instance, *watercourse* and *river* are two names describing the same thing. On the other hand, in schematic heterogeneity the classification and hierarchical structure of the geographical feature could vary within or across disciplines. Finally, syntactic heterogeneity is divided in two types, one is related to the logical data model and its underlying DBMS (DataBase Management System), e.g., *relational* and *object oriented*, while the other is related to the representation of the spatial objects in the database.

Another approach to classify heterogeneity problems has been developed by (Hakimpour 2003). This proposal puts forwards other classification, which presents similar aspects of heterogeneity to the previous one

- Heterogeneity in the conceptual modelling: A geographical feature can be represented in one system as an object class, and in other, as a relation.
- Heterogeneity in the spatial modelling: This feature type could be represented by polygons (or a segment of pixels) in one system, while being represented by lines in the second system.
- Structure or schema heterogeneity: In this heterogeneity type it is possible that different systems hold the same name for a same feature, but different attributes or formats. Therefore, the information of each system is different.
- Semantic heterogeneity: One system may adopt a viewpoint about a feature, while the other may adopt a different one. Moreover, it is usual to find different definitions of a same feature.

In these classifications, big differences of granularity in a same domain do not appear as a heterogeneity problem; however, heterogeneity problems are presented in our catalogues because the conceptual overlap is complete, so the number of concepts that describe the geographical domain range from 22 to over 400. This gap represents other type of heterogeneity which will be taken in account.

For solving the different problems caused by heterogeneity we will have to use mapping techniques. In our integrated system, mappings are the components which relate heterogeneous elements. So, mappings have to solve the heterogenity gap between ontology and catalogues.

5. From proposal to reality

We have carried out the integration task in two phases: first, the building of PhenomenOntology, and then the mapping of the catalogues with PhenomenOntology.

To build the ontology, the members agreed to generate automatically an ontology that could cover as best as possible different feature domains (Administrative boundaries, Vegetations, Buildings, etc). For that, the domain experts decided to use the BCN25 catalogue, which is the most detailed one, for extracting the information and then with that information creating the ontology automatically. Once the ontology is created, the domain experts will have to review and modify it to cover all features presented in other catalogues.

To build mappings, OEG is creating a framework which permits adding new techniques for mapping discovery between our ontology and a database table. The discovery has to be automatic and the resultant alignments will be reviewed by domain experts. Here, we present how to include within the mapping discovery the automatic recognition of new knowledge for learning.

The process above described is carried out in two phases and to enhance usability such phases are designed as automatic sub-processes:1) For creating ontologies automatically, the configurable application, already built, permits experts to generate ontologies quickly and to evaluate the most appropriate taxonomic building criteria. 2) For mapping discovery, the automatic feature will permit to enlarge the system with other catalogues in the near future.

6. Automatic ontology creation

6.1 Scales and coverage

As we have mentioned above, each of the IGN-E catalogues corresponds to a different scale. Therefore, the number of features is inversely proportional to the scale, because the detail permits distinguishing more specialized geographical features. Then, we use a 1:25,000 scale catalogue to generate automatically an ontology using an ad hoc application. This application extracts from instances of a feature catalogue the concepts of PhenomenOntology.

Theoretically, the list of the smaller scale catalogue includes all the features of the larger scale, but this is not truth since a small number of features do not appear in the list of the smallest scale features. Examples of these features are: “cordillera” (mountain chain) or “península” (peninsula), these features cannot be drawn in a map following the smallest scale and these names only appear in a 1:1,000,000 scale catalogues. These special features or terms will be considered during the debugging phase carried out by experts of the domain, though we pretend to identify these terms in the mapping phase.

6.2 Criteria for taxonomy creation

The software developed for the automatic creation of an ontology permits selecting the criteria for taxonomy creation and its order. These criteria are based on the information contained in each row of the BCN25 feature catalogue table. First, the column *codigo* (code) stores codified information about a three-level taxonomy. Therefore, there are three separate criteria from which to extract a superclass in the taxonomy creation process: the first pair, the second pair, and the third pair of code digits. Then, the application permits extracting a super class for each different value of the chosen pair of digits. For increasing the versatility, the application permits selecting one of these pairs of digits, making possible to extract the taxonomy in different fashions: the first level of extraction attending the first pair of digits, the second level of extraction attending the third pair; or attending firstly the second pair and secondly the first pair; or any combination of one, two or three levels.

Three levels can result insufficient for a taxonomy with more than eight hundred leaves. So, we have added another criterion for creating automatically a taxonomy level: common lexical parts. At the beginning of the features, the application can identify common substrings of feature names and then create a common superclass of those concepts whose names begin with an identified substring. See an example of these criteria: we start with concepts "Autovía", "Autopista 2 carriles", "Autopista 3 carriles puente" and "Autopista 3 carriles tunnel" that are siblings; when this criterion is applied, it produces a superclass called "Autopista" (sibling of "Autovía") which has as subclasses "Autopista 2 carriles", "Autopista 3 carriles puente" and "Autopista 3 carriles túnel". Applying this criterion twice, the software produces a taxonomy that has "Autovía" and "Autopista" as siblings; then "Autopista 2 carriles" and "Autopista 3 carriles" are siblings and sons of "Autopista"; and "Autopista 3 carriles puente" and "Autopista 3 carriles túnel" are siblings and sons of "Autopista 3 carriles".

While testing the software, IGN-E and OEG noticed that the substring criterion is not useful because there is lexical heterogeneity in phenomenon names. Therefore, we created another substring criterion to solve the heterogeneity problems mentioned above, and as a result the substrings "Autovía", "AUTOVIAS", "Autovia." and "autovía-" are the same when we create a superclass.

According to the atributes of the BCN25 feature catalogue table, we created another criterion to extract superclasses (a new taxonomy level) for different values of *grupo* (group) that represents a top level classification of feature (as the first pair of code digits).

There are a total of six criteria which can be ordered and combined as we wish, while a same criterion can, sometimes, be included in the criteria list. Fig. 5 shows a schematic example of three criteria.

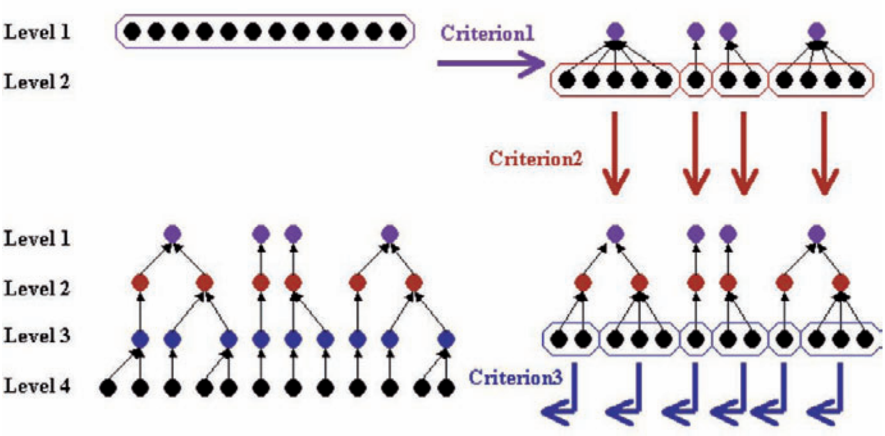


Fig. 5. Example of taxonomic level creation

6.3 Attributes by values

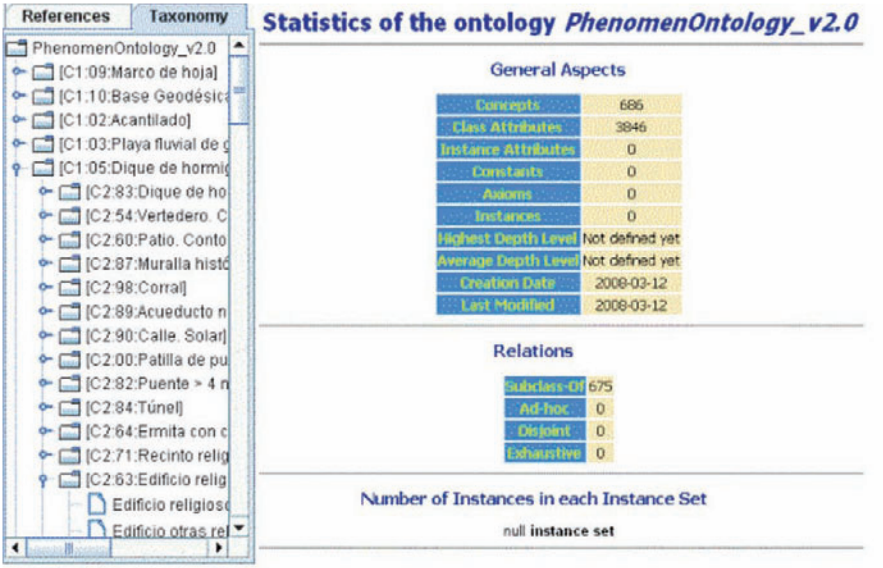
We can find other type of information in feature names that can be quite interesting for ontology conceptualisation. Indeed, feature names in databases contain the values of conceptual attributes (called data properties in the description logics paradigm) that are not explicit in database models. For example, the state of a railway, which can be “en uso” (in use), “en construcción” (under construction), “abandonado” (abandoned) or “desmantelado” (dismantled). There is an opportunity for enrichment the conceptualization attending to these attribute values held in labels. We can upgrade the scanning values by inserting concept attributes with the values found during the reviewing of the concepts.

We had the opportunity of including within the ontology creation software an analyser of feature names. To do that, the software uses a file containing the attribute names and their values; the concepts of the ontology created can have more attributes than databases columns, taking these extra-attributes the values appearing in the name. For example, the concept “FFCC doble desmantelado” (double railway dismantled) will have two extra-attributes: “número de vías” (number of tracks) with value “doble” (double) and “estado” (state) with value “desmantelado” (dismantled). The number of extra-attributes or their values is open for this application.

6.4 PhenomenOntology

IGN-E used the application developed for generating criteria combination tests and chose an automatic generated ontology with three levels (two criteria), 686 concepts and 3,846 attributes, as can view in Fig. 6. This ontology is stored in the WebODE platform (Arpírez et al. 2003).

This first version of the ontology is being refined by IGN-E experts using the WebODE Editor.



Statistics of the ontology *PhenomenOntology_v2.0*

General Aspects

Concepts	686
Class Attributes	3846
Instance Attributes	0
Constants	0
Axioms	0
Instances	0
Highest Depth Level	Not defined yet
Average Depth Level	Not defined yet
Creation Date	2008-03-12
Last Modified	2008-03-12

Relations

Subclass-Of	675
Ad-hoc	0
Disjoint	0
Exhaustive	0

Number of Instances in each Instance Set

null **instance set**

Fig. 6. PhenomenOntology v2.0 and its statistics

7. Automatic mapping discovery

In the approach here presented there are elements that relate feature stored in databases to concepts of PhenomenOntology. These elements are *mappings*.

The sets of mappings are classified into intensional and extensional; they are intensional when mappings relate elements of different set of instances, and they are extensional, when mappings relate elements of different conceptualizations. In our case, mappings relate instances of a model (rows of a table of a Relational Model) to concepts (elements of an ontology conceptualisation). We have not found similar cases in the literature and only a definition of mapping covers this type of relations. This definition is: “*A mapping is a formal explicitation of a relation between elements, or set of elements, of different conceptualizations and/or instantiations.*” (Gómez-Pérez et al. 2008).

Automatic mapping discovery is a traditional work area in which many algorithms and tools are developed; but due to our specific scenario, described above, we need to create new tools and algorithms, studying the existing reuse techniques (distance measures, matching terms, etc.).

We have identified several techniques to discover mappings following the analysis of feature catalogue. Below we present different analyses to review coverage of different heterogeneity types:

- Syntax analysis: String comparison. Before making string comparison, it would be necessary to regularize the string format (capitals, blank spaces, plurals, etc.). This kind of analysis solves, partially, semantic heterogeneity. Example: “autovia”, “Autovía.”
- Syntax analysis: Regular expression. The inclusion of names into other names might represent hiperonymy/hiponymy information. This kind of analysis solves, partially, granularity heterogeneity. Example: “autovía”, “Autovía en construcción”.
- Semantic analysis: Hiperonymy. If it is possible to access linguistics resources where looking for hiperonymy information between names or part of names. This kind of analysis solves, partially, granularity heterogeneity. Example: “muro”, “recinto amurallado”.
- Semantic analysis: Synonymy. If it is possible to access to linguistics resources to look for synonymy information between names, acronyms or abbreviations. This kind of analysis solves, partially, semantic heterogeneity. Example: “muro”, “pared exterior”.
- Semantic analysis: Root. The comparison of roots of lemmas can provide information about synonymy. This kind of analysis solves, partially, semantic heterogeneity. Example: “muro”, “muralla”.
- Semantic analysis: Definitions. IGN-E provides mapping discovery with a set of definitions about features. With these definitions we will discover new synonymy and hyponymy relations. This kind of analysis solves, partially, semantic and granularity heterogeneity. For instance: “muro”, “muralla: muro que rodea un recinto fortificado”.
- Code analysis. Codes are identifiers and we use this code information (see code description in section 2) to identify synonymy relations. This kind of analysis solves, partially, semantic heterogeneity. For instance: “064401 Vías de estación de FFCC. Vía de servicio” (BCN25), “064401 FFCC.VIA_DE_SERVICIO” (BCN200).
- Structural analysis. As mentioned above, codes are composed of three pairs of digits with taxonomic information embedded. This taxonomic information can be used in catalogues with codes for scanning an ontology. This kind of analysis solves, partially, schematic heterogeneity. Example: “064204 FFCC en construcción en puente” (BCN25) implies 06 is transports, 42 is railways (a subclass of transport), and 04 is a subclass of railways.

7.1 Knowledge discovery

To facilitate the review and depuration of the automatic creation of the ontology, during the mapping discovery phase we have studied how to identify new knowledge while discovering mappings. Therefore, in the near future we will take in account two techniques:

- Code analysis. Codes are identifiers and we use code information (see code description in section 2) to identify synonymy relations between terms; then we reuse this semantics for mapping discovery. Example: “064301 FFCC abandonado o desmantelado” (BCN25), “064301 FFCC_FUERA_DE_SERVICIO” (BCN200).
- Not identifying the relations with ontology concepts implies lacks of knowledge in the ontology added. Example: “068202 Radiofaro” (BCN200).

8. Conclusion and future work

Our application for finding lexical heterogeneity in feature names has permitted IGN-E domain experts to evaluate its main data sources.

The automatic creation of ontologies and the easy combination of criteria have permitted us to have, very quickly, different first version ontologies for evaluation; thus expert time and efforts are saved. Other advantages of the application are the identification of extra-attributes by values in the feature names and the automatic storage in an ontology management platform such as WebODE, which permits debugging the ontology by experts easily.

The mapping discovery process is now being developed and we expect to obtain the first results in a few months. The process is being conscientiously developed with the aim of incorporating it to regional and international databases. In this mapping discovery, the application has to solve the heterogeneity problem in different levels so as to identify mappings automatically. However, the results of all the automatic processes must be reviewed by domain experts to get the most successful results.

As we mentioned above, the mapping discovery phase is not concluded yet. Therefore, this work is our priority for the near future.

Once the framework is finished, the collaboration partners will develop new techniques to discover mappings, which will improve this efficiency and will cover other type of relations

When the final phase of the mapping discovery framework is finished and the new techniques are added, then the integration system will be enlarged automatically with the incorporation of other catalogues (regional and local scale), extending the integration to other GI producers.

Acknowledgements

This work has been partially supported by the National Project "GeoBuddies" (TSI2007-65677C02) and the bilateral collaboration UPM-IGN 2007-2008. We are also grateful to Rosario Plaza for her help in checking and reviewing the grammar and spelling of the paper and improving clarity and style.

References

Alonso G, Abbadi AE (1994) Cooperative modelling in applied geographic research. International Journal of Intelligent and Cooperative Information Systems, 3(1): 83-102

Arpírez JC, Corcho O, Fernández-López M, Gómez-Pérez A (2003) WebODE in a nutshell. AI Magazine

Bishr Y (1998) Overcoming the semantic and other barriers to GIS interoperability. International Journal of Geographical Information Science, 12(4): 299–314

International Standard Organization (ISO) (2003) ISO 19112:2003 Geographic Information – Spatial referencing by geographic identifiers.

Fonseca FT, Egenhofer MJ, Davis CA, Câmara G (2002) Semantic Granularity in Ontology-Driven Geographic Information Systems. Annals of Mathematics and Artificial Intelligence. Volume: 36, 2002. Issue: 1-2, pp 121-151

Gómez-Pérez A, Ramos JA (2008) Semantic mappings: out of ontology world limits. Intl. Workshop on Ontology Alignment and Visualization. March 4-7, 2008. Barcelona, Spain, pp 907-912

Greenwood J, Hart G (2003) Sharing Feature Based Geographic Information – A Data Model Perspective. 7th Int. Conference on GeoComputation. United Kingdom

Hakimpour F (2003) Using Ontologies to Resolve Semantic Heterogeneity for Integrating Spatial Database Schemata. Ph.D. thesis, Zurich University, Switzerland

Hakimpour F, Timpf S (2001) Using Ontologies for Resolution of Semantic Heterogeneity in GIS. 4th. AGILE Conference on Geographic Information Science, Brno, Czech Republic

ISO 19110 (2005) Geographic Information – Methodology for feature cataloguing.

ISO 19112 (2003) Geographic Information – Spatial referencing by geographic identifiers.

Nomenclátor Geográfico Conciso de España (2006), versión 1.0. Presentación y Especificaciones. Instituto Geográfico Nacional http://www.idee.es/ApliVisio/Nomenclator/NGCE.pdf

OGC (2003) OpenGIS Reference Model. Version 0.1.2, OGC Inc. Wayland, MA, USA

Rodríguez Pascual AF, García Asensio L (2005) A fully integrated information system to manage cartographic and geographic data at a 1:25,000 scale. XXII International Cartographic Conference. A Coruña, Spain. ISBN: 0-958-46093-0

Sevilla Sánchez C, Rodríguez Pascual AF, González Matesanz FJ, Blanco Ortega LM, Vilches-Blázquez LM (2006) Un SIG corporativo en el IGN para la gestión integrada, publicación y análisis de datos geográficos. In proceedings of XII Congreso Nacional de Tecnologías de la Información Geográfica. Camacho Olmedo, M.T.; Cañete Pérez, J.A.y Lara Valle, J.J. Ed. ISBN: 84-338-3944-6 Depósito Legal: GR-1855-2006

Stuckenschmidt H, Visser U, Schuster G, Vögele T (1999) Ontologies for geographic information integration. Proceedings of Workshop Intelligent Methods in Environmental Protection: Special Aspects of Processing in Space and Time, 13. International Symposium of Computer Science for Environmental Protection, CSEP 1999, pp 81-107

Vilches-Blázquez LM, Rodríguez Pascual AF, Mezcua Rodríguez J, Bernabé Poveda MA, Corcho O (2007a) An approach towards a harmonized framework for hydrographic features domain. In Conference Proceedings of XXIII International Cartographic Conference. 4-10 August 2007, Moscow, Russia

Vilches Blázquez LM, Bernabé Poveda MA, Suárez Figueroa MC, Gómez-Pérez A, Rodríguez Pascual A F (2007b) Towntology & hydrOntology: Relationship between Urban and Hydrographic Features in the Geographic Information Domain. In :Ontologies for Urban Development (Eds). Teller, J.; Roussey, C.; Lee, J. Springer-Verlag, 2007. ISBN: 978-3-540-71975-5

All Roads Lead to Rome – Geospatial Modeling of Hungarian Street Names with Destination Reference

Antal Guszlev[1], Lilla Lukács[2]

[1] University of West Hungary, Székesfehérvár, Hungary.
email : a.guszlev@gmail.com
[2] Institute of History of the Hungarian Academy of Sciences, Budapest, Hungary
email : llilla@map.elte.hu

Abstract

Spatial analysis of place names is a vital part of toponymic research as spatial location and relations between geographical features have a vast effect on natural and artificial name giving. Examining spatial attributes and relations between features can help to understand impacts of distance, hierarchy and other characteristics of natural and man-made objects in naming processes. This paper shows possibilities for modeling and visualizing the connections between settlements of streets which are named for their destinations in Hungary. Spatial analysis of street names can help to investigate the development of settlement and road systems.

Keywords: street names, odonyms, toponymy

1. Introduction

Spatial attributes (location, drift, height, extension, shape, spatial relations with surroundings, etc.) of objects play an important role in name giving, both the natural and artificial processes. Natural denominations became widely used place names within a community if they are “name-like” that is they suggest existing names in form and style. In addition, it is important that names should have the same denotation value for all of members of the community (Hoffmann 2007). Therefore, natural place names are usually descriptive names and the name giving is based on the characteristics of features that are able to identify the objects in human communication. Visual characteristics (e.g. spatial attributes) are easily recognizable and memorizable so they are primer identifiers in natural name giving.

For example, the map of Semaphore (Adelaide, Australia) shows different descriptive street names (Figure 1). Street names like Grand Junction Road, Port Road, and Old Port Road refer to the function of these streets. Mariners Crescent and Settlers Drive give information about inhabitants. West Street and Lower Street store the relative location of these streets.

Some streets are named after persons, e. g. Lord Hobart Way, Mary Ann Street, or after islands, e. g. Samoa, Fiji, Pit-cairn, Corfu, etc. These names are not descriptive names, they are artificial street name groups.

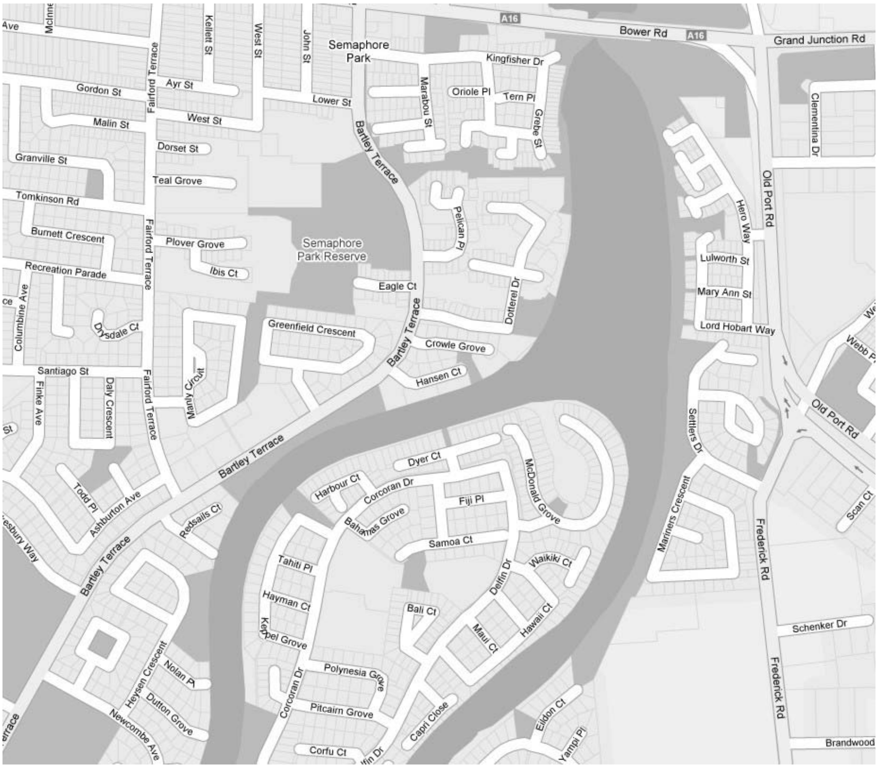


Fig. 1. Descriptive and non-descriptive street names on the map of Semaphore (Adelaide, Australia), maps.google.com

Names, because of their descriptive nature, can give information of recent or former relief, vegetation, land cover, inhabitants or history of a geographical place. Usefulness of toponymy in the reconstruction of geographical and human environment was confirmed in other studies (Conedera et al. 2006).

On the other hand, Johnson points out that street type keywords in street names can cause troubles in geocoding. This is a typical problem of foreign street names: keywords are often handled as a part of the street names in databases (Johnson 2007). This problem shows the necessity of studies on street names also in aspect of spatial databases.

Impacts of spatial attributes can be seen in artificial name giving also. In this case new names should fit into the existing naming system, in order that the new names are more likely to get used in everyday communication in a shorter time, and it is required that the toponymic corpus keeps consistency (Fülöp 1981). For this reason artificial name giving must not be high-handed (top-down driven). Creating place names is regulated by decrees and regulations. In Hungary a special expert committee (Hungarian Board on Geographical Names, Magyar Földrajzinév-bizottság) was established in the 1960s to control official name giving by authorizing new names or offering an opinion on them.

There are cultural differences in street name giving. Legal background, historical traditions, geographical environment and grammatical correctness must be considered in street name giving. In Hungary, street names which denote directions or refer to geographical places are preferred more than commemorative names (Földi 2005). Rules forbid us-ing ordinal numbers as street names, which is frequent method in the U.S.

The examination of street names (or odonymy) is a well-developed scientific field with a rich bibliography in Hungary, but is dominated by linguistic approach (Vincze 1984; Kálmán 1989; Hoffmann 2007). The novelty of our research is that we apply GI processing methods to examine Hungarian street names that contain geographical references to indicate their directions or destinations. These names can also be analyzed at local and at state level. In this research, street names of town Kaposvár were analyzed to find relationships between street names and relief, land cover, and structure of a settlement; and impacts of spatial aspects in name giving methods. (Kaposvár is the seat of Somogy County in Southwest Hungary; an average size town of about 65 000 inhabitants). In addition, street names with settlement directions in their names were modeled in GIS for the whole country area.

2. First case study: Street names of town Kaposvár (Hungary)

2. 1. Street name types based on motivation of naming

Linguists and toponymic experts analyze place names from various points of view, e.g. functional-semantic, lexical-morphological, and syntagmatic analyses (Hoffmann 2007). According to these approaches place names can be sorted by origin, denotation, morphology, etc. In this research, spatial aspects of place names have been modeled, therefore street names were classified for this purpose. Street name types are based on the motivation of naming, with a focus on spatial types. With this end in view, the following types were established for our research (this grouping is based on Fülöp's analysis on street names of (Kaposvár and Fülöp 1981)):

- Street name from place name
  - settlement name
  - watercourse or lake name
  - landscape region name
  - name of a settlement part
  - geographical generic term

  names refer to destination
  or
  names do not refer to destination
- Street name from term which refers to a characteristic of the street
- Street name from term which refers to natural environment
- Street name from term which refers to cultural environment
- Street name from personal name
- Street name from term which refers to a characteristic of human living•
- Street name from other word

Classification of street names originated from place names can be elaborated, based on types of the geographical features. Official street name giving is the authority of local governments, and there are traditions in name giving also. Therefore, percentage of name types differs in settlements. In Kaposvár there are 414 street names, 25% of which originated from geographical names.

2. 2. Street names with destination references

Street names containing place names can indicate direction or destination of the street, but there are many names without destination reference in each type. (e.g. London Road in Brighton and in Portsmouth, UK; Georges River Road in Campbelltown, Australia.)

On Figure 2, streets of Kaposvár that are named for the destination are marked with thick lines. It is clearly visible that the marked roads touch the outer boundary of the inhabited area.

Út (road) and utca (street) are the most frequent generic terms occurring in Hungarian street names. Usually, shorter roads in settlements are called utca and wider or longer roads called út. Traditionally, roads with direction reference were called út in natural name giving, but at present, this is not obligatory in (artificial) naming. For example, Figure 2 shows that the term utca is more frequent in Kaposvár than út.

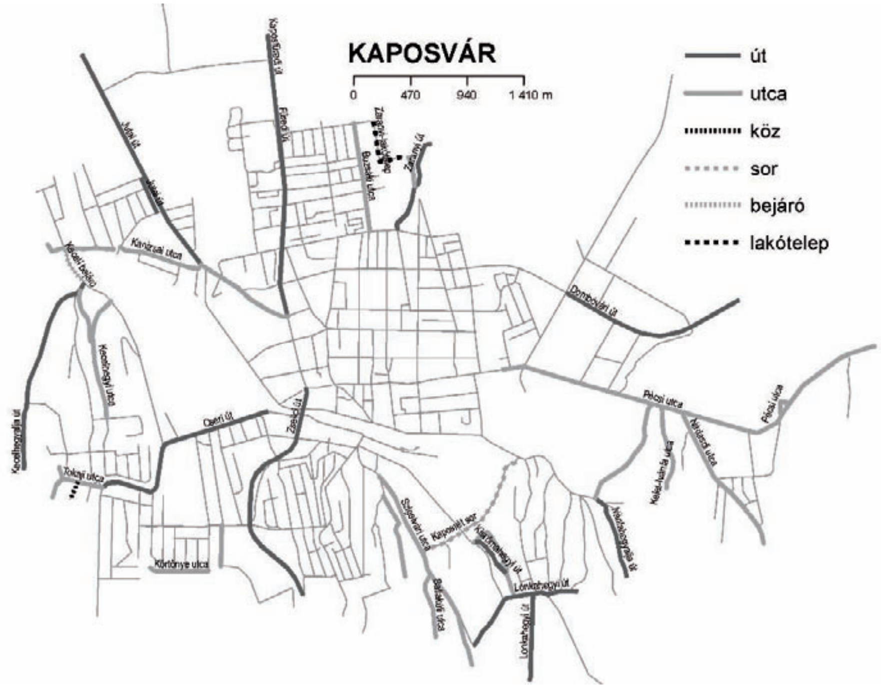


Fig. 2. Types of generic terms in street names of Kaposvár

Street names with destination or direction references can denote settlements (e. g. the neighboring settlement, the nearest town, a distant important town, or the capital of the recent or former state, etc. In other frequent cases, street names denote other geographical objects, e.g. Balaton (largest lake in Hungary) -> Balatoni út, Balatoni Street.

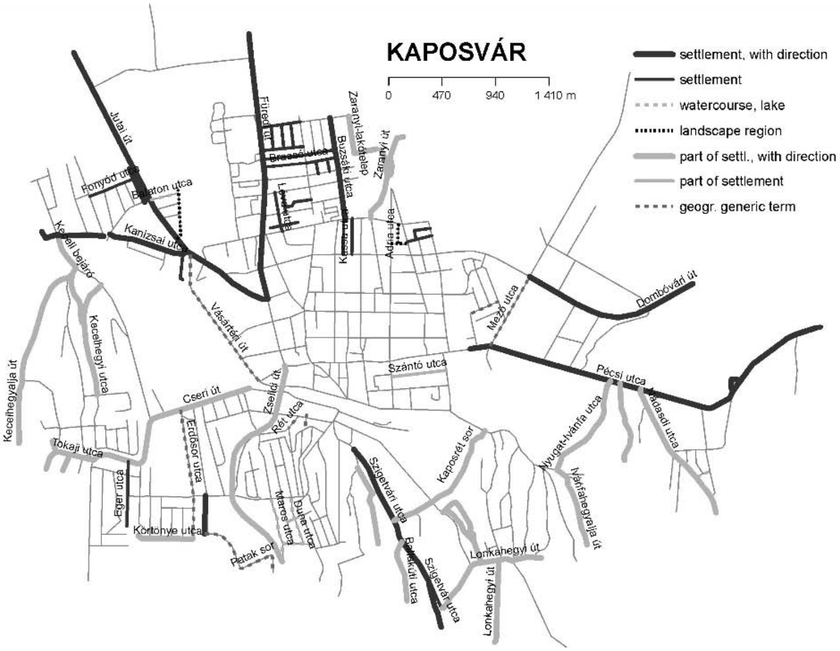


Fig. 3. Classification of street names of Kaposvár according to the type of included geographical feature

Classification and visualization of street names according to indicated destination could give information about relief, important geographical features, main transportation routes, connections (and dependences) to neighboring settlements, etc. Street names of Kaposvár have been compared to geographical characteristics of the town area (Figure 3). In case of Kaposvár, street names containing settlements and settlement parts point at different directions. Most street names with settlement destination are in the northern half of the town, and point at north. Most streets with settlement part in their names are in the southern half, and point at south. Compared to the relief and environment it can be recognized that street names reflect spatial characteristics:

- There are more hill names in the southern street names, because settlement parts are usually hills in the south.
- There are few southern destination road names including settlements, because just a few main routes leave Kaposvár to south.
- Names of northern destination roads usually include settlement names and do not include settlement parts, because there are few named hills or

beacons which could be reference, there are more destination roads, and main roads do not cross populated areas.

3. Extension to whole Hungary

3. 1. Spatial modeling and visualization of connections between settlements and street names

Names with destination reference can be modeled and analyzed by GIS tools on the whole road network of Hungary. In typical geodatabases names are only stored as attributes of geometric objects. GI systems do not support linguistic analytic methods, but by extending traditional spatial analysis functions to names we can gather information about

- relations between the hierarchy of settlements and popular name giving methods
- effects of distance on generating street names including destination
- correlations between settlements in having street names cordially pointing at each other (at similar or at different hierarchy levels)
- historical development of the road system and of main (traditional) transportation routes between settlements
- spatial differences of quantitative and qualitative attributes of toponymic phenomena (changes and simplifications of names)
- spatial distribution and texture of names with destination reference in Hungary
- etc.

3. 2. Settlement and settlement parts in street names

If we are talking about destination of a road in its name, we must remember that not just settlements are included in names, but settlement parts and settlements with shorter names. Hills, lands, estates, and other geographical and man-made features can be parts of a settlement. In other cases, former separate settlements become parts of other settlement by fusion. For historical analysis these name forms have also been identified.

Long settlement names which are usually compound words become shorter in everyday communication. Street names often are generated from these short name forms instead of the official settlement names. Approximately 700 short names were found in Hungary (there are 3152 settlements in Hungary in 2007). The list of these short name forms was not available, therefore manual collecting was necessary. After preparing this list, collected forms have been examined in street names automatically.

3. 3. Analysis of name forms

Place names are often included in street names of each Hungarian settlement (e.g. there is a Budai út in Győr, and there is a Győri út in Budapest). Similar to Kaposvár, these street names often denote the destination of the road which leaves the town or village.

A street name with destination reference differs from the place name in the form, because usually they got an -i suffix, e.g. Buda –> Budai út (Budai Street), Győr -> Győri út (Győri Street). The -i suffix generates adjectives from nouns in Hungarian language. There are some special cases with other suffixes, e.g. Ferihegy (airport in Budapest) -> Ferihegyi repülőtérre vezető út (means "the road which goes to Ferihegy Airport". Sometimes the name of a road is inverted between two settlements (e.g. the road mentioned before is starting at Buda (Budapest) with the name Győri út and ending at Győr with the name Budai út).

Occurring problems in analysis:

- Automatic identification of destination street names can be difficult, because there are other name types with -i suffixes, and often information of name origin is needed. There are street names with similar forms originated from persons (e.g. Toldi út: Toldi is a personal name here, although it means that the person came from Told village). There are some frequently used personal names, but most of personal names in street names must be recognized as unique cases.
- There are some street names with geographical names which are not destinations, just simple proper name parts of a street name referring to a former settlement, commemorating an event, or can be artificial names without semantic relations (e.g. some streets of the town Szeged commemorate European cities). Information of location or the line of bearing of streets can help to identify these streets.
- If the name root ends with letter "i", then it will not get another -i suffix (e. g. Nemti ->Nemti instead of Nemtii). In some cases, the root of the name changes when getting -i suffix. Usually a vowel is dropped for easier pronunciation (e. g. Eger-> Egri instead of Egeri). These changes are regular; therefore they can be treated automatically with defining rules for them.
- Some settlement names become shorter as a part of a street name (Székesfehérvár -> Fehérvári instead of Székesfehérvári)

- Settlement names without antecedents should also be looked for in street names
- According to a law in Hungary, every municipality has to have a unique name, so settlement names are considered unique identifiers. However, because of the shorter forms in common language, sometimes more than one settlement name can be selected (e.g Tiszafüred, Balatonfüred, -> Füredi út). In these cases distance, size and historical notability determine the name source. Increasing distance shows the decreasing order of feasibility, but historical traditions must be recognized as special cases.

The overall reliability of the analysis depends on how we deal with the above mentioned special issues and exceptions.

3. 4. Spatial Analysis

Data mining tools can be used to assist the process of exploring large amounts of data in search of recurrent patterns and relationships. Geographical data mining is a special type of data mining that seeks to perform similar generic functions as conventional data mining tools, but modified to take into account the spatial features of geoinformation, the different styles and needs of analysis, modelling and geoprocessing, and the peculiar nature of geographical explanation (Openshaw 1999).

Base data of analyses are databases of the street names and the settlements of Hungary. The database of street names contains about 126,000 records of street segments, roads, squares, and other public areas of Hungary with information of settlements and counties. The database of settlements is a basic gazetteer that contains settlements of Hungary including data of spatial location (coordinates). In order to simplify the tasks we have used static relational tables. There is a possibility to monitor and keep track of settlement and street changes, but these advanced processing functions are beyond our current aims.

Processes:

1. Filtering street names ending with letter "-i" (resulting 17,694 records) with SQL command,
2. Trimming the last character of names in the whole selected set
3. Creating relation between settlements and the trimmed street database (9916 links). The relation is simple 1:1 relation, as Hungarian settlement names are unique identifiers.

4. Connecting destination street names with coordinates of the start and end points (the destination settlement). The connection was simplified to lines as the crow flies, instead of real road routes.
5. Determining one-way or cordial two-way relations, and registering directions of lines.
6. Filtering exceptions, solving disambiguation. Finding errors could not be automated, and needed a large amount of manual checking.

3. 5. Visualization

Street names with destination reference and connection between the starting and destination settlements are worth to be implemented into geoinformation systems and visualized on thematic maps.

Some attributes have been visualized by symbolizing methods (Figure 4): Color and size of a settlement sign relate to the number of inhabitants. Arrows represent relationships between settlements of destination street names. Line width represents distance; line color denotes one- or two-way relations.

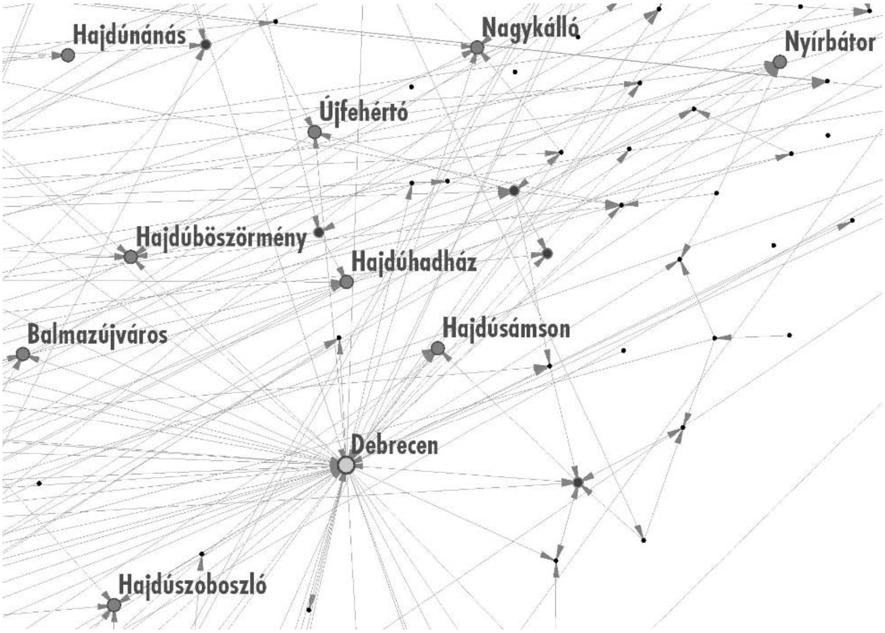


Fig. 4. Relationship between settlements based on street names

With a look at relations of the settlements, it is obvious that visual analysis of the whole destination street system is not feasible. GIS tools give opportunity to filter and analyze these relationships in many aspects. Some examples of analyses:

- Number of incoming arrows can indicate the importance of a town or village in the settlement system of the country. Settlements with many incoming arrows are usually the local (economical, cultural, etc.) centers of regions. Naturally, Budapest, the capital has the larg-est number of incoming arrows (83 street, including Budai, Pesti, and Budapesti streets).
- Length of arrows (distance between connected settlements) can refer to historical traditions. For example, names of former comitat seats or market-towns can occur in street names far from these towns.
- Most of relationships are recorded in street names of neighboring villages, because of the natural name giving of smaller communities. Therefore, it is worth to examine historical effects in street name giving by leaving neighboring villages out of consideration. (Neighboring settlements can be filtered by creating Thiessen polygons on settlement points.)
- Former settlements, settlement fusions and name changes can be explored by filtering former names and settlement name parts in street names. Building the geoinformation system of these relationships can help to reconstruct the development of settlement network in Hungary.

4. Conclusions

Spatial analysis of toponyms can help to explore basic motivations and spatial differences in name giving. Toponyms record information about relief, settlement and transportation system of an area, and relationships between human populations and geographical objects. A model of the environment can be built by analyzing spatial relations recorded in place names. A method of analyzing and visualizing street names with destination reference was elaborated in our research.

Jumping back to the first sentence of our paper: according to our street names database 31 roads lead to Rome in Hungary (Figure 5).

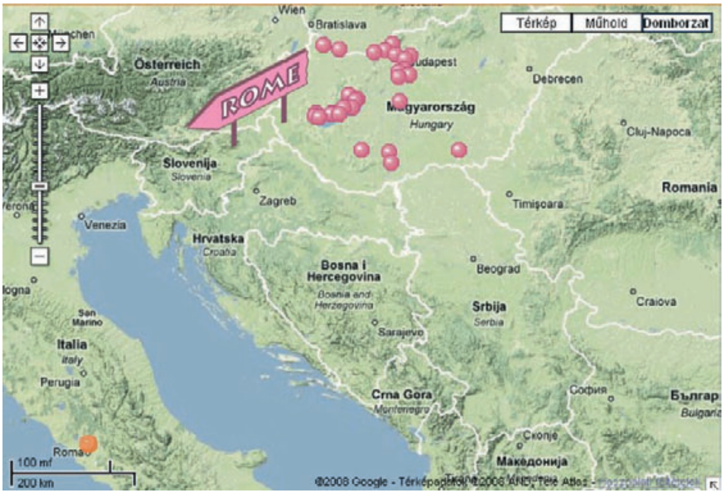


Fig. 5. Street names which contain word 'Római' (Rome) in Hungary

References

Hoffmann I (2007) Helynevek nyelvi elemzése. Tinta Könyvkiadó, Budapest

Fülöp L (1981) Kaposvár utcaneveinek névtani vizsgálata, Magyar névtani dolgozatok, Budapest

Földi E (2005) Térképi névírás (course book, draft)

Johnson JL (2007) Improving Geocoding for Unusual Road Names, ArcUser 10 (4): 20-21

Kálmán B (1989) A nevek világa. Debrecen

Merhavia (2004) Budapest City Map 1: 20 000, Budapest

Openshaw S (1999) Geographical data mining: key design issues

Conedara M et al. (2006). Using toponymy to reconstruct past land use: a case study of 'brüsáda' (burn) in southern Switzerland

Vincze L (1984) Új módszer az utcanevek vizsgálatára. Névtani Értesítő 9., Budapest

Where is the Terraced House? On the Use of Ontologies for Recognition of Urban Concepts in Cartographic Databases

Patrick Lüscher[1], Robert Weibel[1], William A. Mackaness[2]

[1] Department of Geography, University of Zurich
Winterthurerstrasse 190, 8057 Zurich, Switzerland
email: patrick.luescher@geo.uzh.ch

[2] Institute of Geography, School of GeoSciences, University of Edinburgh, Drummond St, Edinburgh EH8 9XP, Scotland, UK

Abstract

In GIS datasets, it is rare that building objects are richly attributed. Yet having semantic information (such as tenement, terraced, semi-detached) has real practical application (in visualisation and in analysis). It is often the case that we can infer semantic information simply by visual inspection – based on metric and topological properties for example. This paper explores the application of pattern recognition techniques as a way of automatically extracting information from vector databases and attaching this information to the attributes of a building. Our methodology builds upon the idea of an ontology-driven pattern recognition approach. These ideas are explored through the automatic detection of terraced houses (based on Ordnance Survey MasterMap® vector data). The results appear to demonstrate the feasibility of the approach. In conclusion we discuss the benefits and difficulties encountered, suggest ways to deal with these challenges, and propose short and long term directions for future research.

Keywords: cartographic databases, ontologies, ontology-driven pattern recognition, building types, geographical characterisation

1 Introduction

Spatial databases currently in use typically have been originally designed and produced in the 1990s. They are rich in geometry, most often include topological structuring, yet they are usually poor in semantics. Those exceptional databases that are semantically rich are restricted to rather narrow purposes – vehicle navigation being a prominent example, where rich additional information on the logics of traffic flow (e.g. one-way streets, pedestrian zones etc.), average speed and speed limits are coded onto the geometry. However, the majority of GIS applications make use of general purpose topographic databases produced either by national mapping agencies (NMAs) or by private companies (e.g. Tele Atlas, NAVTEQ). These general purpose databases are poor in semantics in particular with regards to the representation of higher order semantic concepts that extend beyond the semantics of individual, discrete objects.

This under-representation of semantics limits the utility of the database. The research community has called for methods to automatically 'enrich' such databases. What is required are methods that make *explicit* the spatial relationships and semantic concepts *implicitly* contained in spatial databases. Probably the first research community to call for 'data enrichment' was the map generalisation community (Ruas and Plazanet 1996; Heinzle and Anders 2007). In map generalisation, the special semantics embedded in spatial relations, hierarchical relations, and spatial patterns and structures are critical to modelling the context in which cartographic decisions are made. The map generalisation process utilises information linked to pattern and structure recognition (Brassel and Weibel 1988; Mackaness and Ruas 2007). For example, the decision as to whether to visualise a building on a map will partially depend on contextual information. If it is small yet isolated in a rural area, then the building may be retained and slightly enlarged; if it is in an urban area, it may be eliminated; and if it happens to be a special type of building such as a hospital, it may be replaced by a special symbol (Steiniger 2007).

Generalisation is not the only area where enriched semantics and hence cartographic pattern recognition are crucial. Building types such as tenements or terraced, semi-detached, and detached houses are rarely coded into existing spatial databases, yet, they would provide important semantic information in many practical applications: They give essential clues to prospective house buyers as to what to expect when reading through real estate advertisements (King 1994); information concerning house type is important in planning when trying to develop the right balance between different residential forms in a particular neighbourhood, in quantity surveying or in the recycling of building materials (Müller 2006; Bergsdal et al. 2007). Additionally, enriched semantics can be used to associate urban patterns with urban evolution processes and urban morphology (Camacho-Hübner and Golay 2007); or they may assist adaptation in pedestrian navigation services by considering spatial contexts specified in the database (Winter 2002).

In this paper, we present a novel approach to cartographic pattern recognition. In addition to the more 'traditional' approaches that directly rely on statistical methods and/or geometric algorithms, our approach utilises ontologies to better inform the pattern recognition process and to 'glue' such algorithms together. The paper begins by explaining why ontology-driven pattern recognition has the potential to overcome some of the limitations of traditional approaches and describes the proposed methodology (§ 2). We demonstrate how this approach affords automatic identification of terraced houses from among urban buildings represented in vector form. After presenting an ontology of terraces (§ 3), we explain how the concepts of this ontology can be transformed into an automatic recognition procedure, and we present results of this procedure using Ordnance Survey MasterMap data (§ 4). The paper goes on to identify the benefits and limitations of this technique and suggests ways of overcoming these limitations (§ 5). The conclusion reflects on future research, short and long-term.

2 Ontology-driven Cartographic Pattern Recognition

2.1 Why ontologies are useful in cartographic pattern recognition

Many specialised pattern recognition algorithms have been developed for the detection of structures and patterns specifically in an urban context (e.g. Regnauld 1996; Barnsley and Barr 1997; Anders et al. 1999; Boffet 2001; Christophe and Ruas 2002; Heinzle and Anders 2007; Steiniger et al. 2008). These techniques focus on rather specific patterns that are linked to particular generalisation operations, for instance where we wish to group buildings or to detect alignments in support of aggregation or typification operations (Regnauld 1996; Christophe and Ruas 2002). As there is often an element of fuzziness involved in pattern definitions, these algorithms are often coupled with statistical methods. It remains doubtful whether such algorithms, or a collection thereof, will be sufficient to extract more general, higher order semantic concepts such that we could comprehensively describe the semantics of the morphology of a city. There has to be something additional that enables broader synoptic description of the city form. It has been pointed out by Mackaness (2006) that abstraction from large-scale databases to highly generalised ones requires that the roles of individual features and patterns be understood and modelled explicitly. Dutton and Edwardes (2006), Kulik (2005) and Redbrake and Raubal (2004) show the importance of semantic modelling of geographic features in maps to guide user adaptation during generalisation.

In our research, therefore, we pursued a 'top-down' approach to cartographic pattern recognition of urban structures. The individual steps of this ontology-driven approach are illustrated in Figure 1: Based on textual descriptions of urban spaces extracted from the literature, we identify specific urban patterns (step 1); we then formalise these patterns, their context and hierarchical composition based on ontological descriptions (step 2). The ontological definitions of patterns are then used to deductively trigger appropriate 'low level' pattern recognition algorithms (step 3) in order to detect them in spatial databases (step 4).

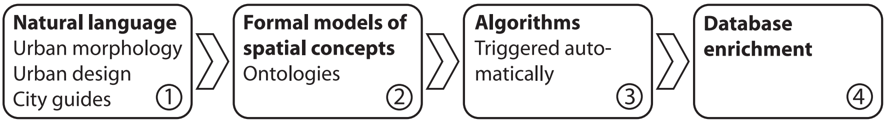


Fig. 1. Steps in the processing chain of ontology-driven pattern recognition

In this way, we can overcome some important drawbacks of methods used today:

- Current pattern recognition methods have often been developed and parameterised for specific data models and databases. For instance, if they have been developed with German ATKIS data in mind, they might assume that roads are represented by centre lines. It is anticipated that ontologies will provide meta-knowledge that improves the 'interoperability' and applicability of pattern recognition methods across different databases.
- It is often the case that existing pattern recognition algorithms cannot be adapted to take into account additional information in the detection procedure, such as topography, which may be important in describing the genesis of certain urban patterns. Ontological descriptions help make explicit all the criteria that enable us to identify a particular composition of buildings (Klien and Lutz 2005).
- The nature of geographic form means that many spatial patterns cannot be crisply defined and delineated. Therefore pattern recognition additionally depends upon the use of statistical techniques (e.g. Steiniger et al. 2008). The result of typical statistical methods may be difficult to interpret, however, as the relations that are inferred between pattern variables

are purely statistical rather than revealing causes and consequences. Ontologies, on the other hand, represent the concepts that are modelled, as well as the relations between them in an explicit way. Thus, they are inherently more transparent than statistical methods and have potentially more explanatory power.

2.2 Ontologies for cartographic pattern recognition

The term 'ontology' is defined from an engineering science perspective and is defined as an explicit specification of a shared conceptualisation (Gruber 1993). It is thus an attempt to capture the knowledge in a certain domain in a systematic way by breaking it down into the types of entities (*concepts*) that exist and the *relations* that hold between them. Ontologies can be classified according to the degree of formalisation into informal (written in natural language), semi-formal (restricted language), and formal (artificial language) ontologies (Agarwal 2005). An alternate classification is one that conforms to the degree of specialisation and is divided into top-level, domain, and task ontologies, the last being the most specific one (Guarino 1998). While a key application of ontologies is to improve the interoperability between information systems (Fonseca et al. 2002), ontologies are also employed as a method of eliciting knowledge that exists in a domain (Agarwal 2005).

In this research we seek to explain complex urban phenomenon in terms of other, possibly simpler phenomena, such that the meaning of the concept is derived from the meaning of the related concepts. We refer to the first kind as a 'higher order concept', and to the second kind as a 'lower order concept'. The lower order concepts may themselves be composite concepts, in which case they have to be broken down further into still lower order concepts. Alternatively they might be simple in the sense that they can be directly related to cartographic measures or a cartographic structure recognition algorithm.

2.3 Data enrichment using ontologies

The concept above constitutes an *ideal prototype* (a template). Real occurrences of a concept will normally comply only to a certain degree with the template. Hence, a value which expresses the degree of congruence between reality and the ideal prototype of the concept has to be calculated: where $con(C_i, R_j) = 0$ when a realisation R_j differs completely from a template C_i, and $con(C_i, R_j) = 1$ when they match perfectly.

For low order concepts *con(C_i, R_j)* is extracted by a cartographic pattern recognition algorithm. For composite concepts, which are defined by their relations to lower order concepts, *con(C_i, R_j)* has to be inferred from the congruence values of their constituting concepts. Here we distinguish between two types of relationships:

- Some relationships, such as the subclass relationship, translate to strict exclusions:

$$con(C_i, R_k) = 0 \rightarrow con(C_j, R_k) = 0 \tag{1}$$

  If C_j is a subclass of C_i. For example, if a spatial object is not a building then it cannot be a terraced house, regardless of the congruence values of the other constituting concepts, since terraced houses are a subclass of buildings.
- For other relationships, congruence values of the constituting values have to be intersected. One possibility for combining single similarity values to an overall value is by calculating a weighted linear average:

$$con(C_i, R_k) = \left(\sum w_j con(C_j, R_k)\right) \Big/ \sum w_j \tag{2}$$

  Where *con(C_j, R_k)* is the congruence value of a constituent concept of C_i and the weight w_j is an influence value of the subconcept. For reasons of simplicity, all weights were equated to 1 for this study.

Thus, the calculation of congruence values starts with the patterns at the bottom and then propagates iteratively to higher order concepts. This is similar to forward reasoning in description logics. At the end of this process, spatial objects can be annotated with the congruence value for the concepts defined in the ontology.

2.4 Related work

Our review of related work will be brief and will focus exclusively on approaches that use explicit semantic models for the recognition of spatial patterns in *vector databases*, ignoring the literature related to image interpretation and computer vision.

Sester (2000) and Anders and Sester (1997) built semantic models for the automatic interpretation of large-scale vector databases. They extracted different types of houses, streets, parcels and built-up areas from polygon data. The inductive machine learning algorithm ID3 is used to discover relevant spatial properties and relations in manually tagged data. An approach for combining spatial reasoning with description logics to formalise spatial arrangements is presented by Haarslev et al. (1994).

Many spatial concepts are inherently vague. Santos et al. (2005) used supervaluation semantics to integrate vagueness into logical reasoning. They show a prototype implementation in Prolog that classifies water bodies according to an ontology of inland water features.

Ontologies are a means to achieve semantic interoperability in a distributed environment. In this context, Klien and Lutz (2005) discuss the automatic annotation of existing datasets with concepts defined in an ontology. Their approach emphasises spatial relations between features rather than individual feature properties. Thomson (2006) sought to build land use maps from OS MasterMap data. Her intention was to use ontologies to model land use categories according to the specific spatial configurations, compositions, and relations. This is somewhat similar to a project at the Ordnance Survey which sought to identify fields such as farming land or pasture in OS MasterMap data, using ontologies (Kovacs and Zhou 2007).

We conclude our review with a few observations. First, the amount of work using semantic models for pattern recognition in cartographic vector databases is much smaller than the literature on purely algorithmic approaches. Second, much of the research reviewed in this subsection is restricted to a selected set of spatial patterns; the extensibility and the potential generality of these approaches is rarely discussed. And finally, few references have actually gone into details of instantiating the proposed ontology definitions and of implementing a prototype to prove the validity of the approach; many stay at the more theoretical level.

3 An ontology of terraced houses

«Beyond the mills … were the rows of terraces – mean little houses, with low ceilings and dark cramped rooms.» — Jane Rogers, *Her Living Image*.

In this section we want to show how textual descriptions of urban concepts can be formalised and thus serve as a basis for their detection. The concepts in this study were collected from texts on urban morphology, which is “the study of the physical (or built) fabric of urban form, and the people and processes shaping it” (Jones and Larkham 1991). The hypothesis of urban morphology is that economic and social significance of a town finds its expression in the physiognomy, which is a combination of town plan, pattern of building forms, and pattern of urban land use (Conzen 1969). Concept descriptions were complemented using dictionaries such as the Oxford English Dictionary (Simpson and Weiner 1989). By way of example, Figure 2 shows residential house types identified in the urban morphology literature.

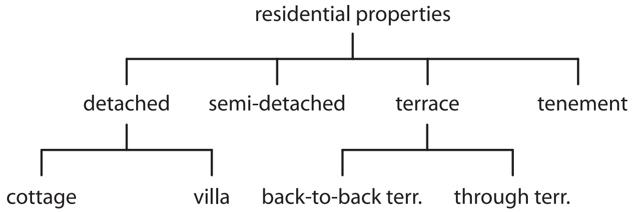


Fig. 2. Urban residential house types extracted from the glossary of urban form (Jones and Larkham 1991)

While ‘terraced house’ is generally a synonym for ‘row house’ and may therefore have different features depending on culture and construction period, the *prototype* for our formalisation is the characteristic terrace house settlement in the UK of the late Victorian and Edwardian period. It is linked to the Public Health Act of 1875, established to improve urban living conditions and resulted in re-housing of population from slum clearance areas (Conzen 1969). The demand for cheap mass housing was met by creating rows of unified buildings sharing sidewalls. Because of the low social status of the dwellers, lot sizes and room footprints were small.

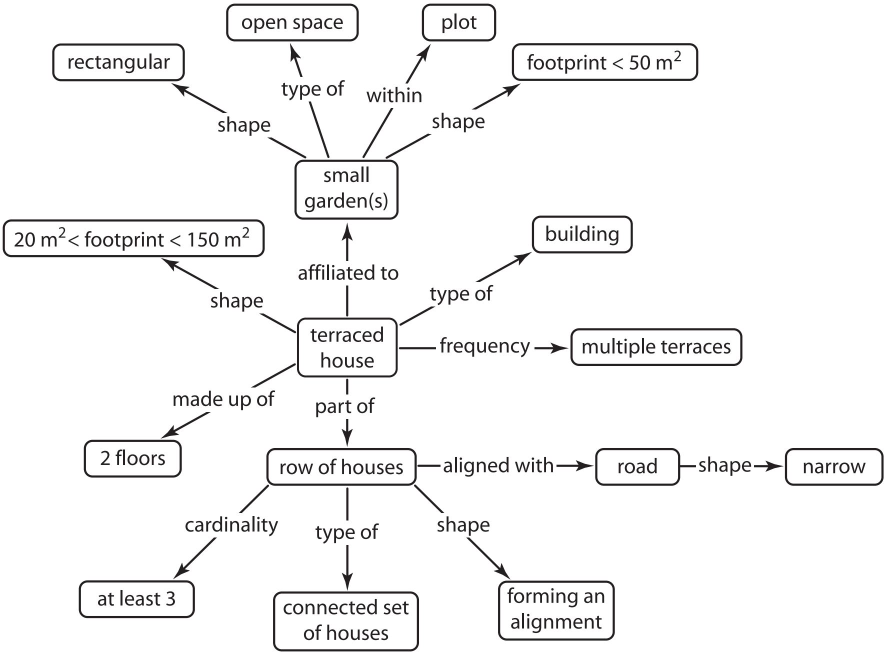


Fig. 3. An ontology of terraced houses

Terraced houses usually have small front-gardens and possibly attached sculleries and a yard at the rear. Often, multiple rows of houses form an area of a highly regular plot pattern. The ontology extracted from these descriptions is shown in Figure 3.

4 Experiment

In order to assess the data enrichment performance of the ontology-driven approach in general and the terraced house ontology in particular an experiment was carried out using OS MasterMap data for Edinburgh, Scotland, UK. OS MasterMap provides a planar topology, that is, space subdivided into polygons such that no polygons overlap, and every location is covered by exactly one polygon. The ontology was realised in a prototype for ontology-driven pattern recognition programmed in Java, tough the current prototype does not yet implement the concepts ‘small garden(s)’ and ‘narrow roads’.

4.1 Extraction and composition of low order concepts

As described in § 2.3, low order concepts can be mapped to cartographic measures. For the terraced house ontology, the following low order concepts have been implemented:

- The concept ‘building’ can be trivially extracted from OS MasterMap; an attribute encodes whether a polygon represents open land, transportation or a building.
- ’20 m^2 < footprint < 150 m^2’ was obtained using a crisp threshold for building areas.
- Since OS MasterMap does not contain any information on the height of buildings, the concept ‘made up of two floors’ had to be omitted.
- For the concept ‘row of houses’, groups of buildings were created. There are several methods that calculate alignments of buildings (see Burghardt and Steiniger 2005 for an overview). We derived the degree of alignment by grouping buildings sharing a common wall and then connecting the centroids of the buildings for groups containing at least three buildings, so that a path representing the general form of the group was formed (Figure 4a). The form of the path was assessed using the compactness of the area covered by the path. We also rated homogeneity of buildings within groups by means of the standard deviation of the building areas. Finally, the form of the path and the

homogeneity of buildings were averaged to obtain the congruence value of building groups to alignments. Figure 4b shows the congruence values for an extract of our study area: Linearly arranged, homogeneous blocks in the northwest of the extract achieve high congruence values, whereas 'perimeter-block development'-like blocks receive low congruence values.

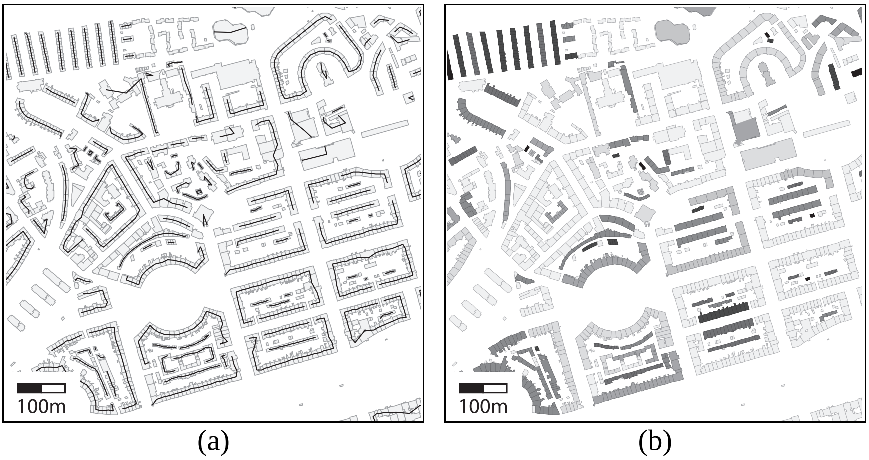


Fig. 4. (a) Paths to qualify the general form of building groups (b) Congruence of buildings to the concept 'row of houses'. Light values denote low, dark values denote high similarity. OS MasterMap data Ordnance Survey © Crown Copyright. All rights reserved.

- The concept 'multiple terraces' was derived by identifying the main axes of building groups and clustering these groups using the direction of the axes. The clusters were then qualified by means of the homogeneity of axes directions, length of axes, and homogeneity of buildings within the clusters. To this end, standard deviations were calculated and averaged as previously discussed. Figure 5 shows an example of the clusters found. Note that in the right hand part of the figure, there are two areas – marked (1) and (2) – with regular rows of buildings that have not been classified as 'multiple terrace'. This is because the footprints of the building areas are too large and hence they correspond rather to tenements than to terraced houses. The two rows marked as (3) have not been detected as being 'regular' because we defined that there must be at least three approximately parallel rows of houses for this condition to be met.

Finally, the congruence value of ‘terraced house’ was calculated by intersecting ‘building’, ‘20 m^2 < footprint < 150 m^2’, ‘row of houses’, and ‘multiple terraces’ as explained in § 2.3.

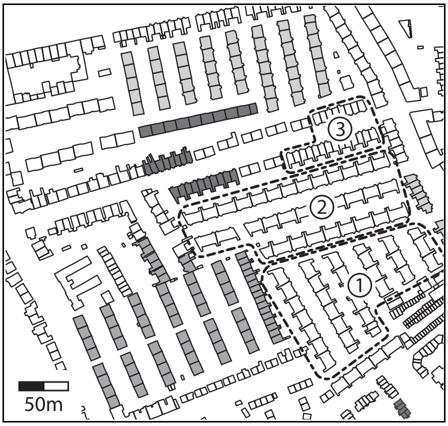


Fig. 5. Areas of multiple terraces. OS MasterMap data Ordnance Survey © Crown Copyright. All rights reserved.

4.2 Results

The classification has been carried out for an area covering a part of the City of Edinburgh, 4.6 km x 3.6 km size. The congruence values obtained were deliberately classified into the three categories in order to simplify the validation process:

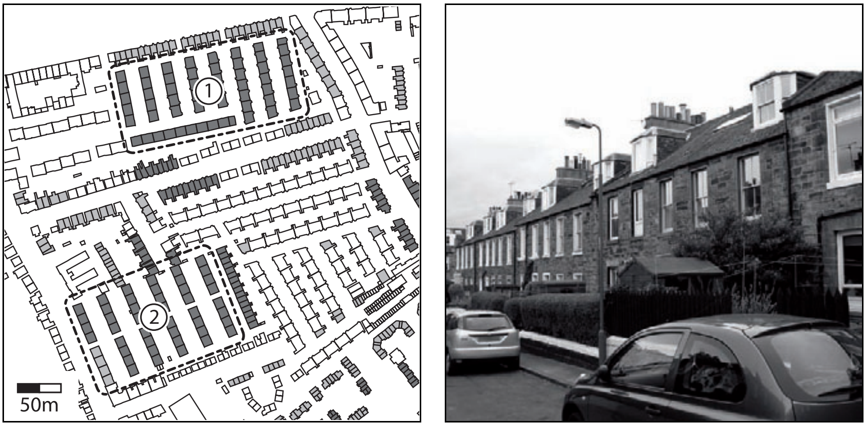
- ‘high’ congruence: *con(‘terraced house’, R_i)* > 0.8
- ‘medium’ congruence: 0.6 < *con(‘terraced house’, R_i)* ≤ 0.75
- ‘low’ congruence: *con(‘terraced house’, R_i)* ≤ 0.6

Of the 20 990 houses in the study area, 1 557 were classified as having high congruence, 5 064 as having medium congruence, and 14 369 as having low congruence with the concept ‘terraced house’. We did some ground truthing to measure the occurrence of terraced houses, but not for all of Edinburgh. The results were compared to ground truth where available, and visually compared to aerial photographs elsewhere.

The algorithm identified six larger areas of terraced houses. Five of those areas correspond to settlements known as the ‘Edinburgh Colonies’ that fit pretty nicely to our conceptualisation of terraced houses (Figures 6 and 7). There was one settlement of the ‘Colonies’ that was not classified fully as having a high congruence value, namely the North Forth Street

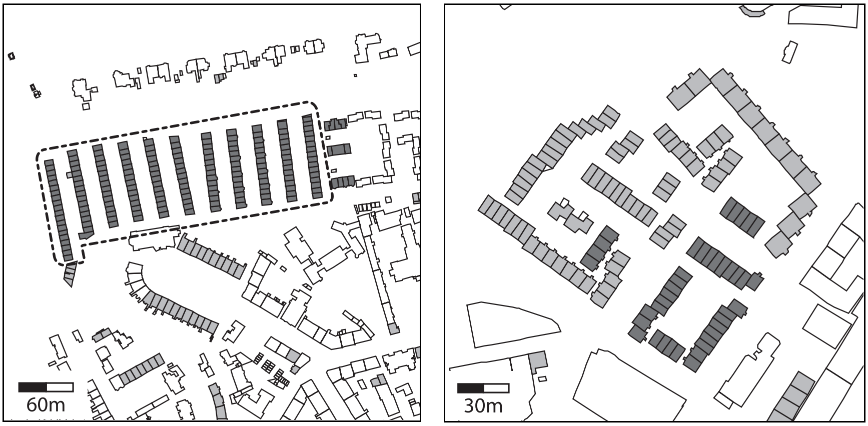
Colony (Figure 7b). The reason for this is that our algorithm for ‘multiple terraces’ extracts parallel rows of houses rather than orthogonally arranged rows such as in the North Forth Street Colony.

Finally, 775 of the 1 271 buildings classified as having high congruence could be definitively confirmed as terraced houses. This does not imply that the remaining 496 buildings with high congruence values are in fact not terraces (equivalent to an error of commission), but simply that in these cases a ground survey will be needed to confirm the result.



(a) (b)

Fig. 6. (a) Leith Links (1) and Lochend Road (2) Colonies. (b) Picture of terraced houses in the Leith Links Colony. High congruence with ‘terraced house’ concept in dark grey, medium congruence in light grey, for low congruence just building boundaries are shown. OS MasterMap data Ordnance Survey © Crown Copyright. All rights reserved.



(a) (b)

Fig. 7. (a) Stockbridge Colony. (b) North Forth Street Colony. Contrast levels as in Figure 6. OS MasterMap data Ordnance Survey © Crown Copyright. All rights reserved.

5 Discussion

5.1 Benefits

In general, the results generated are plausible. This research has shown how textual descriptions of urban patterns can be used to define an ontology that in turn can be used to inform the detection of these patterns, thus enabling enrichment of existing vector cartographic databases. Since the ontology makes the concepts and relations defining a spatial pattern explicit, it can also be used to generate graphical representations such as the one seen in Figure 3 as well as textual descriptions (or metadata) about the extracted patterns. And finally, it follows trivially from Figure 3 that it would be easy to modify concepts in the ontology of the higher order concept ‘terraced house’, or add further low order concepts to it. For instance, it would be possible to accommodate cultural differences between prototypical terraces in different regions or countries. Our ultimate aim is to extend this framework such that a domain expert can define his/her conceptualisation of *any* urban pattern as an ontology and has a useful set of low order patterns at hand that can be used to perform the detection process.

5.2 Difficulties

Operationalisation of concepts: The operationalisation of lower order patterns is not necessarily easy. One example is the concept ‘multiple terraces’, which means that a larger number of rows of terraces are arranged regularly. Regularity itself is a loose term, and there are several ways of measuring it. We defined a regular arrangement of terraces as a group of at least three approximately parallel rows of houses. The generation of such groups involves creating a buffer to both sides of each main axis and intersecting this buffer with other main axes. This works well for typical terraced houses (Figure 5), but more general definitions may be needed when different concepts are to be detected.

Another example is the derivation of alignments of houses. There exist various methods for grouping houses into alignments (Burghardt and Steiniger 2005; Christophe and Ruas 2002; Boffet 2001). They assume different conceptualisations of the constitution of alignments and hence produce different results. Therefore, the influence of the choice of implementation of the low order concepts to the inference workflow and to the recognition performance has to be investigated in detail.

Thresholds: Some of the concepts involved setting a threshold (e.g. the area of the footprint of a building). Such crisp thresholds are rather undesirable and could be improved using fuzzy membership functions (Ladner et al. 2003).

Defining a processing order: For complex concepts like terraced houses, a processing hierarchy has to be identified. The hierarchy defines the order of the inference of lower level concepts and their composition into higher level concepts. This is made difficult by the fact that lower level concepts in different sub-branches sometimes depend on each other. For example, the detection of areas of multiple terraces assumes that terraces have already been detected, but in turn also inform the detection process of terraced houses. Since we turned our ontology manually into a detection process, these interdependencies could be accounted for. With respect to a more automated operationalisation process (which is desirable because domain experts are usually not experts in programming), we need more research on how we can formally model such interdependencies.

Alternative ways of concept inference: The method to calculate congruence values of composite concepts was given in § 2.3. The strengths are its simplicity, the fact that the output is a similarity (congruence) value instead of a hard classification, and the high level of transparency of the results. Fuzzy logic would offer a similar but more complex approach.

Supervised classification methods (Steiniger et al. 2008) use training data to define characteristic properties of different classes, and hence there is no need to set thresholds. On the other hand, the performance of supervised classification depends largely on the quality of the training samples used. Furthermore, it is our opinion that using ontologies can better integrate structural knowledge about concepts into the reasoning process and hence is better adapted to detecting complex concepts.

6 Conclusions

In this paper, we have advocated the use of ontologies to better inform the recognition of spatial patterns and structures in the urban environment from cartographic vector databases. We have explained how we envisage ontology-driven cartographic pattern recognition as a novel complement to traditional algorithmic and statistical pattern recognition. For the example of terraced houses, we have developed an ontology, implemented the corresponding recognition procedure in Java, and validated it using OS MasterMap data.

There are several insights that can be gained from this work. Ontologies definitely render the recognition process more flexible (and extensible), enable greater self-documentation, and make us better equipped to compose complex concepts from simple concepts as opposed to traditional algorithmic techniques. Despite the great potential of ontology-driven approaches, they still represent a relatively unfamiliar approach in this application domain and hence pose a series of challenges for future research. Among the difficulties encountered in our study (§ 5) are the operationalisation of concepts; the proper way of dealing with thresholds and fuzziness; dealing with concept interdependencies when integrating simple to complex concepts; and alternative ways of concept inference.

In the short term we plan the following extensions to this study: Complete ground truthing to completely validate our results; application of the procedure to other study areas; modification and/or extension of the ontology of terraced houses (e.g. to accommodate cultural differences); experiments using people to study where and how they visually detect terraces; and development and implementation of ontologies of other house types (semi-detached, detached, tenement). In the mid term we envisage first integrating the different building ontologies to a ‘house’ ontology, and later to an ontology of even higher order concepts such as ‘residential area’. And in the long term we hope to develop methods for the automated ‘deployment’ of ontologies, which will facilitate the application of ontology-driven pattern recognition for domain experts.

Acknowledgements

Funding of the first author by the Swiss State Secretariat for Education and Research (SER) through COST Action C21 (project ORUS, grant no. C05.0081) is gratefully acknowledged. The OS MasterMap data used in this study were made available to the University of Edinburgh through Digimap. We are also grateful to the Institute of Geography, University of Edinburgh for hosting an extended stay of the Zurich researchers during November 2007.

References

Agarwal P (2005) Ontological Considerations in GIScience. International Journal of Geographical Information Science 19(5):501–536

Anders K-H, Sester M (1997) Methods of Data Base Interpretation – Applied to Model Generalization from Large to Medium Scale. In: Förstner W, Plümer L (eds) Semantic Modeling for the Acquisition of Topographic Information from Images and Maps: SMATI 97. Birkhäuser, Basel, pp 89–103

Anders K-H, Sester M, Fritsch D (1999) Analysis of Settlement Structures by Graph-Based Clustering. SMATI'99 – Semantic Modelling for the Acquisition of Topographic Information from Images and Maps, 7th September, Munich, Germany

Barnsley MJ, Barr SL (1997) Distinguishing Urban Land-use Categories in Fine Spatial Resolution Land-cover Data using a Graph-based, Structural Pattern Recognition System. Computers, Environment and Urban Systems 21(3–4):209–225

Bergsdal H, Brattebø H, Bohne RA, Müller DB (2007) Dynamic Material Flow Analysis of Norway's Dwelling Stock. Building Research & Information 35(5):557–570

Boffet A (2001) Méthode de création d'informations multi-niveaux pour la généralisation cartographique de l'urbain. Ph.D. thesis, Université de Marne-la-Vallée

Brassel KE, Weibel R (1988) A Review and Conceptual Framework of Automated Map Generalization. International Journal of Geographical Information Systems 2(3):229–244

Burghardt D, Steiniger S (2005) Usage of Principal Component Analysis in the Process of Automated Generalisation. Proceedings of the 22nd International Cartographic Conference, 11–16 July, A Coruna, Spain

Camacho-Hübner E, Golay F (2007) Preliminary Insights on Continuity and Evolution of Concepts for the Development of an Urban Morphological Process Ontology. In: Teller J, Lee JR, Roussey C (eds) Ontologies for Urban Development. Studies in Computational Intelligence, Vol. 61. Springer, Berlin Heidelberg New York, pp 95–108

Christophe S, Ruas A (2002) Detecting Building Alignments for Generalisation Purposes. In: Richardson DE, van Oosterom P (eds) Advances in Spatial Data Handling (10th International Symposium on Spatial Data Handling). Springer, Berlin Heidelberg New York, pp 419–432

Conzen MRG (1969) Alnwick, Northumberland: A Study in Town-plan Analysis. Institute of British Geographers, London

Dutton G, Edwardes A (2006) Ontological Modeling of Geographical Relations for Map Generalization. Proceedings of the 9th ICA Workshop on Generalisation and Multiple Representation, 25th June, Portland, USA

Fonseca FT, Egenhofer MJ, Agouris P, Câmara G (2002) Using Ontologies for Integrated Geographic Information Systems. Transactions in GIS 6(3):231–257

Gruber TR (1993) A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition 5(2):199–220

Guarino N (1998) Formal Ontology and Information Systems. In: Guarino N (ed) Formal Ontology in Information Systems. Proceedings of FOIS'98. IOS Press, Amsterdam, pp 3–15

Haarslev V, Möller R, Schröder C (1994) Combining Spatial and Terminological Reasoning. In: Nebel B, Dreschler-Fischer LS (eds) KI-94: Advances in Artificial Intelligence: 18th German Annual Conference on Artificial Intelligence. Lecture Notes in Artificial Intelligence 861. Springer, Berlin Heidelberg New York, pp 142–153

Heinzle F, Anders K-H (2007) Characterising Space via Pattern Recognition Techniques: Identifying Patterns in Road Networks. In: Mackaness WA, Ruas A, Sarjakoski LT (eds) Generalisation of Geographic Information: Cartographic Modelling and Applications. Elsevier Science, Amsterdam et al., pp 233–253

Jones AN, Larkham PJ (1991) Glossary of Urban Form. Historical Geography Research Series no.26. Institute of British Geographers, London

King AD (1994) Terminologies and Types: Making Sense of Some Types of Dwellings and Cities. In: Franck KA, Schneekloth LH (eds) Ordering Space – Types in Architecture and Design. Van Nostrand Reinhold, New York et al., pp 127–144

Klien E, Lutz M (2005) The Role of Spatial Relations in Automating the Semantic Annotation of Geodata. In: Cohn AG, Mark DM (eds) Spatial Information Theory, International Conference, COSIT 2005. Lecture Notes in Computer Science 3693. Springer, Berlin Heidelberg New York, pp 133–148

Kovacs K, Zhou S (2007) Key Challenges in Expressing and Utilising Geospatial Semantics at Ordnance Survey. Presentation held at the European Geoinformatics Workshop, 7–9 March, Edinburgh, UK.
http://www.nesc.ac.uk/action/esi/download.cfm?index=3411
Accessed 22.01.2008

Kulik L, Duckham M, Egenhofer MJ (2005) Ontology-Driven Map Generalization. Journal of Visual Languages and Computing 16(3):245–267

Ladner R, Petry FE, Cobb MA (2003) Fuzzy Set Approaches to Spatial Data Mining of Association Rules. Transactions in GIS 7(1):123–138

Mackaness WA (2006) Automated Cartography in a Bush of Ghosts. Cartography and Geographic Information Science 33(4):245–256

Mackaness WA, Ruas A (2007) Evaluation in the Map Generalisation Process. In: Mackaness WA, Ruas A, Sarjakoski LT (eds) Generalisation of Geographic Information: Cartographic Modelling and Applications. Elsevier Science, Amsterdam, pp 89–111

Müller DB (2006) Stock Dynamics for Forecasting Material Flows – Case Study for Housing in The Netherlands. Ecological Economics 59(1):142–156

Redbrake D, Raubal M (2004) Ontology-Driven Wrappers for Navigation Services. In: Toppen F, Prastacos P (eds) AGILE 2004, 7th Conference on Geographic Information Science. Crete University Press, Heraklion, pp. 195–205

Regnauld N (1996) Recognition of Building Clusters for Generalization. In: Kraak MJ, Molenaar M (eds) Advances in GIS Research II: Proceedings of the Seventh International Symposium on Spatial Data Handling. Taylor & Francis, London, pp 4B.1–4B.14

Ruas A, Plazanet C (1996) Strategies for Automated Generalization. In: Kraak MJ, Molenaar M (eds) Advances in GIS Research II: Proceedings of the Seventh International Symposium on Spatial Data Handling. Taylor & Francis, London, pp 6.1–6.17

Santos P, Bennett B, Sakellariou G (2005) Supervaluation Semantics for an Inland Water Feature Ontology. Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, July 30 – August 5, Edinburgh, Scotland

Sester M (2000) Knowledge Acquisition for the Automatic Interpretation of Spatial Data. International Journal of Geographical Information Science 14(1):1–24

Simpson J, Weiner E (1989) The Oxford English Dictionary. Oxford University Press, Oxford

Steiniger S (2007) Enabling Pattern-aware Automated Map Generalization. Ph.D. thesis, University of Zurich

Steiniger S, Lange T, Burghardt D, Weibel R (2008) An Approach for the Classification of Urban Building Structures Based on Discriminant Analysis Techniques. Transactions in GIS 12(1):31–59

Thomson T (2006) Cartographic Data Analysis and Enhancement. Objective 1: Analysis of Functional Information – Needs & Opportunity Analysis, User Requirements, and Use Case. Technical Report, Dept. of Geomatics Eng., University College London

Winter S (2002) Ontologisches Modellieren von Routen für mobile Navigationsdienste. In: Kelnhofer F, Lechthaler M (eds) TeleKartographie und Location-Based Services. Schriftenreihe der Studienrichtung Vermessungswesen und Geoinformation, Technische Universität Wien, pp 111–124

Information Processes Produce Imperfections in Data—The Information Infrastructure Compensates for Them

Andrew U. Frank

Department of Geoinformation and Cartography
Technical University Vienna
Gußhausstraße 27-29/E127
A-1040 Vienna, Austria

Abstract

Data quality describes the imperfections in geographic data. Imperfections are caused by imperfect realizations of the processes that are used to collect, translate, and classify the data. The tiered ontology gives a clarifying framework to analyze the data processes and the imperfections they introduce. The information infrastructure that processes the data and the decision methods using the data are adapted to compensate for some of the imperfections.

1 Introduction

Geographic information is used by many for different applications and these new users need the descriptions of the quality of the data to assess the fitness for their use. The general discussion assumes that low quality is negative and focuses on methods to improve the quality of the data, but data quality should not be expressed on an absolute scale—I will use the concept of absence of imperfection for this—but relative as fitness for a particular use. For example, a less detailed dataset, which is less perfect, is better fit for a use when the detail is irrelevant. The less perfect dataset may have higher quality for a particular use than a more detailed, more perfect one.

Human decision making is based on heuristics; a model of rationality would amount to perfect and complete knowledge—something humans cannot achieve. Bounded rationality (Simon 1956) considers the cost of acquiring information and suggests that decisions are taken with incomplete information. An ecological model of human decision making takes into account that humans have limited computational resources and must make decisions in limited time (Gigerenzer et al. 1999). The focus of this paper is on producing a realistic ontology in the sense of ecological rationality.

In this article I build on previous publications on ontology and data quality (Frank to appear 2008) and explore how imperfections in data processing are compensated. The tiered ontology for geographic data (Frank 2001; Frank 2003) is used and extended to include the processes that are used to transform information between the tiers. A systematic review of the information processes reveals the imperfections these processes introduce and indicates how compensation methods can be used to reduce negative influences on decisions.

The novel contribution of the article is, firstly, the generalization of information processes to focus on the imperfections they produce and, secondly, the compensatory methods that are related to each kind of imperfection based on principles of ecological reasoning.

To escape the terminological confusion I use

- "data quality" as a one-dimensional assessment of the fitness for use of data for a particular decision;
- "data imperfection" as a multi-dimensional description of how the description of the world in the data differs from reality;
- "information process" to include any process that produces, transforms or uses data, starting from observations and measurements to decision making;
- "compensation methods" to describe all information processes that reduce the negative effects of imperfections in the data on the decisions.

This paper starts in section 2 with a short review of a simplified tiered ontology. Section 3 shows how imperfections are introduced by data processing. Section 4 looks at decisions and how strategies to compensate for imperfections in geographic data is used to reduce the negative effects.

2 Ontology

An ontology describes the conceptualization of the world used in a particular context (Guarino et al. 2000): different applications may use different conceptualizations. A car navigation system determines the optimal path using the conceptualization of the street network as a graph of edges and nodes, whereas an urban planning application conceptualizes the same space as regions with properties. The ontology clarifies these concepts and communicates the semantics intended by data collectors and data managers to persons making decisions with the data.

If an ontology for an information system contributes to the assessment of the usability of the data, it must not only conceptualize the objects and processes in reality but must also describe the information processes that link reality to the different conceptualizations. Ontologies that divide conceptualization of reality in tiers, (Frank 2001; Smith et al. 2004), must describe the processes that transform data between tiers.

2.1 Tier O: Physical Reality

Tier O of the ontology 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; Sartre 1943; translated reprint 1993); sometimes Ontology in this sense is capitalized and it is never used in a plural form (in contrast, the ontologies for information systems are written with a lower case o). The observed interactions between humans is only possible if we assume that there is only one, shared physical reality.

2.2 Tier 1: Observations

Reality is observable by humans and other cognitive agents (robots, animals). Physical observation mechanisms produce data values from the properties found at a point in space and time: $v=p(x, t)$. A value v is the result of an observation process p of physical reality found at point x 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’). In GIS such observations are, for example, realized as raster data resulting from remote sensing (Tomlin 1983); similarly our retina performs many such point-observations in parallel.

2.3 Tier 2: Objects

The second tier of the ontology contains the description of the world in terms of physical objects. An object representation is more compact, especially 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 parcel of land, such as size, or form, remain the same for years and need not be observed and processed repeatedly, only ownership, land use, and land cover change and must be observed regularly.

The formation of objects—what Zadeh calls granulation (Zadeh 2002)—first determines the boundaries of objects and then summarizes some properties for the delimited regions before a mental classification is performed. For objects on a table top (Figure 1) a single method of object formation dominates: we form spatially cohesive solids, which move as a single piece: a cup, a saucer, and a spoon.

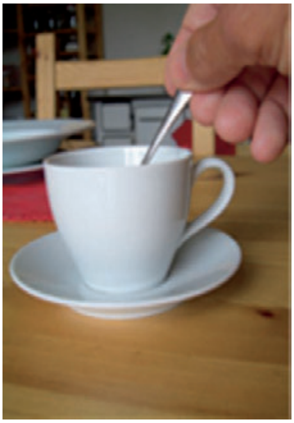


Fig. 1. Simple physical objects on a table top: cup, saucer, spoon

Geographic space does not lead to such a single, dominant, subdivision. Watersheds, but also areas above some height above sea level or regions of uniform soil, uniform land management, etc. can be identified (Couclelis 1992). However, they are delimited by different properties and can overlap (Figure 2).

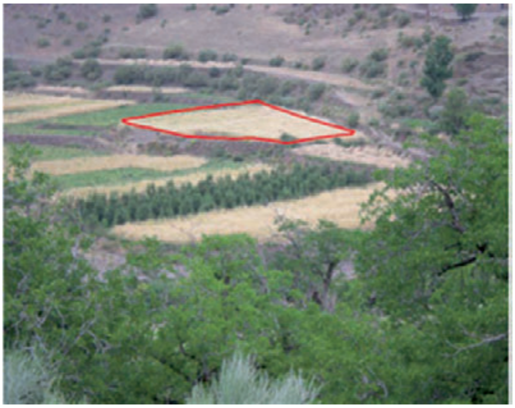


Fig. 2. Fields in a valley: multiple overlap subdivisions in objects are possible

2.4 Tier 3: Constructions

Tier 3 consists of constructs combining and relating physical objects. These constructs are generally socially coordinated. A physical object X is used to mean the socially constructed object Y in the context Z: “X counts as Y in context Z” (Searle 1995, 28) For example, a special kind of stone in the ground counts as boundary maker in the legal system of Switzerland.

Social constructions relate physical objects or processes to constructed objects or processes. Constructed objects can be constructed from other constructed objects, but all constructed objects are eventually grounded in physical objects.

3 Information Processes Transform between Tiers

Information processes transform information obtained at a lower tier to a higher tier (Figure 3).

All human knowledge is directly or indirectly the result of observations, transformed in sometimes long and complex chains of information processes. All imperfections in data must be the result of some aspect of an information process (Figure 3). As a consequence, all theory of data quality and error modeling has to be related to empirically justified properties of the information processes. The production of complex theory for managing error in data without empirical grounding in properties of information processes seems to be a futile academic exercise.

The information processes will be analyzed in the following sections to understand their effects on data, specifically how they contribute to imperfections in the data.

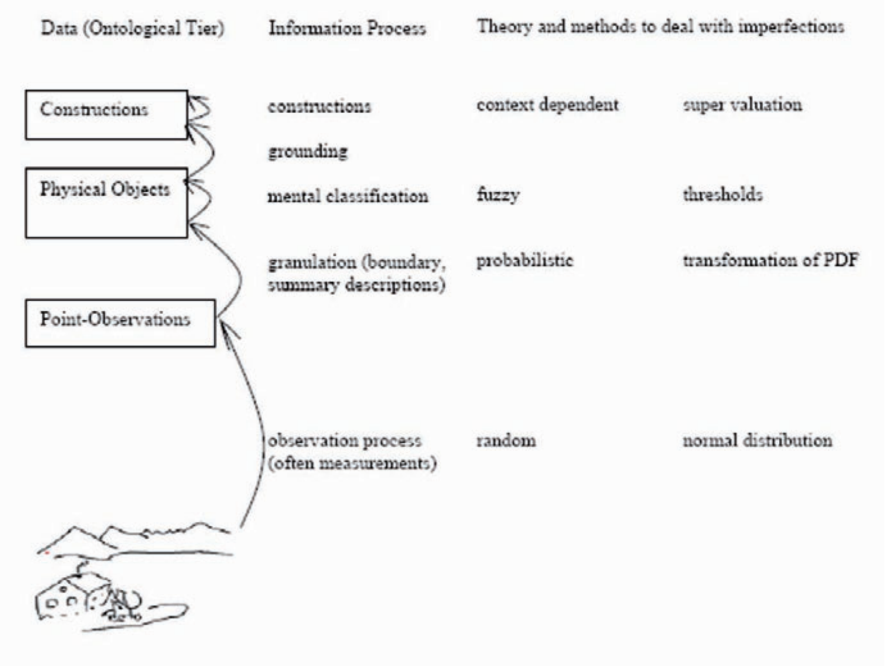


Fig. 3. Tiers of ontology and information processes transforming data between them

3.1 Observations of Physical Properties at Points

The observations of physical properties at a specific point is a physical process that links tier O to tier 1; the realization of which is imperfect in 3 ways

- systematic bias in the transformation of intensity of a property into a quantitative (numerical) value,
- unpredictable disturbance in the value produced, and
- observations focus not at a point but over an extended area.

The systematic bias can be included in the model of the sensor and be corrected by a function. The unpredictable disturbance is typically modeled by a probability distribution. For most sensor a normal (Gaussian) probability distribution function (PDF) is an appropriate choice.

A sensor cannot realize a perfect observation at a perfect point in space or time. Any finite physical observation integrates a physical process over a finite region during a finite time. The time and region over which the integration is performed can be made very small (e.g., 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. Note that the size of the area and the duration influences the result—the infamous modifiable area unit problem appears already here (Openshaw et al. 1991). The necessary finiteness of the sensor introduces a scale in the observations. The sensor can be modeled as a convolution with a Gaussian kernel. Scale effects 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), convolution seems to be a promising approach.

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 eyes seem to report about size, color, and location of objects around us. The properties of extended objects are immediately available, converted from point observations to object data without the person being conscious about the processes involved. Processes of mental formation of objects are found not only in humans, higher animals form mental representations of objects as well.

Object formation increases the imperfection of data—instead of having detailed knowledge about each individual pixel in Figure 2 only a summary description of, for example, the wheat field is retained. The reduction in size by a factor in the order of 105 of the data is achieved with an increase in imperfection.

Object formation consists of two information processes, namely, boundary identification (subsection 3.2.1) and computation of summary descriptions; (3.2.2) objects formed are then mentally classified (3.2.3).

3.2.1 Boundary identification

Objects are—generally speaking—regions in 2D or 3D that are uniform in some aspect. The field in Figure 2 is uniform in its color, tabletop objects in Figure 1 are uniform in their movement: each point of the rigid object moves with a corresponding movement vector.

An object boundary is determined by first selecting a property and a property value that should be uniform across the object. It produces a region of uniform values and boundaries for these regions. Assuming a PDF for the determination of the property of interest one can describe the PDF for the boundary line. The associated transformation function transforms the PDF of the point observation in a PDF for the boundary line (Figure 4).

3.2.2 Determination of descriptive summary data

Descriptive values summarize the properties of the object determined by a boundary. The value 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).

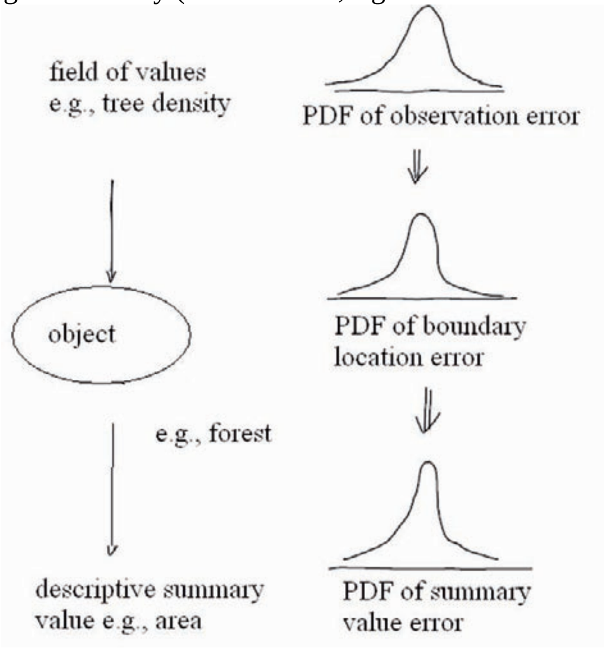

field of values
e.g., tree density
PDF of observation error
object
PDF of boundary location error
e.g., forest
descriptive summary value e.g., area
PDF of summary value error

Fig. 4. Transformation of probability distribution functions from observations to boundary and summary value

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. Given the probability distribution function (PDF) for the value of interest of the summary and the PDF for the boundary, a PDF for the summary values is obtained by transformation of the input PDF (Figure 4).

3.2.3 Mental classification

Objects once identified are mentally classified. On the tabletop, we see glasses, forks, and plates; in a landscape forest, fields, and lakes are identified. Mental classification is an information process internal to tier 2 related to the potential use of an object, the "affordances" of the object (Gibson 1986; Raubal 2002). 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 plant wheat on the field in Figure 2 some of properties of the objects involved must obtain: soil properties, climate, 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., "wheat field", "field" is a subtaxon), distinctions form a taxonomic lattice. The mental taxonomy adapts in the level of detail to the situation and can be much finer than the one implied in the vocabulary if the situation requires it (Ganter et al. 2005). Affordances are in this view are often used as bundles of distinctions.

Distinctions reflect the limits in the property values of an object, where the object can or cannot be used for a specific interaction. The decision whether the values for an object are inside the limits or not is more or less sharp and the cutoff gradual. The distinctions and classifications are therefore fuzzy values, i.e., membership functions as originally defined by Zadeh (1974).

Humans classify unconsciously and immediately the objects we encounter and retain only the mental classification. For example: the parcels suitable for planting wheat are classified as "wheat fields". The classification in the mental taxonomic lattice is an abstraction reducing the detailed information initially perceived in preparation for a probable decision. Instead of retaining detailed values for the decisive properties till the time of decision making only the classification is retained.

3.3 Constructions

Constructions are concepts that are (1) mental units, which (2) have external representations (signs, e.g., words), (3) can be communicated between cognitive agents, and (4) are, within a context, without imperfection.

3.3.1 Grounding

The agent's direct sensory experience of the world is reflected in the agent's experience of the world, an externally representable information image of reality is created duplicating the sensory "reality" in the brain (Figure 5).

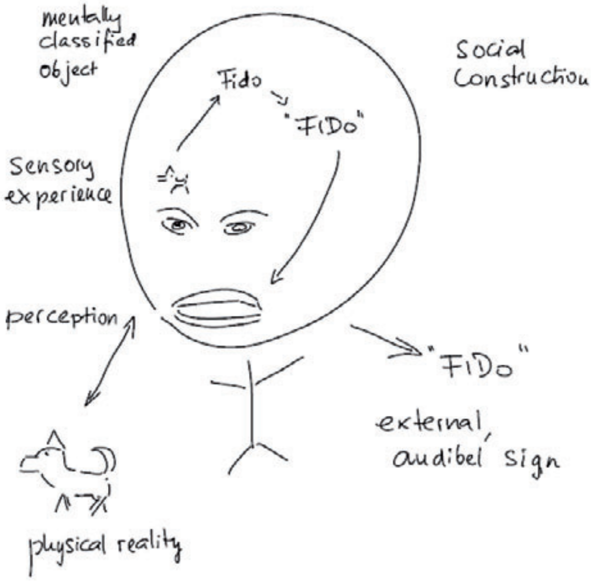


Fig. 5. The grounding of constructs in experiential concepts

Using Searle's formula for a semi-formal treatment, I posit that mental experiential concepts have corresponding representable concepts; the formal "X counts as Y in context Z" is generalized from physical objects to mental concepts: an experiential concept X counts as a representable concept Y in a particular situation and context. Note that the experiential concept—an experience of a thing in reality—can be caused by an ordinary physical object or a physical object that is intended as a sign (Eco 1976). The formula provides grounding for all constructions in mental concepts, which are all directly or indirectly grounded. Ungrounded ("freestanding Y terms" (Zaibert et al. 2004)) would be meaningless because no relation to the physical world is fixed.

3.3.2 Context

The meaning of constructions are determined in a web of concepts that are bound by the relations between the constructs. The full set of concepts that are interrelated are called the context of the construct; the semantics of the construct is determined only through the relations in this context and within this context. Notice the terminology: a person is in a real world situation, the meaning of a sign (construct) is given by context.

Considering these structures as algebraic structures of which the semantics is determined only up to a structure preserving isomorphism. This is the precondition for communication to be possible: it must be possible to translate between different representations (mental, verbal, written) while maintaining the meaning; the translations must be structure preserving mappings (Eco 2003). Such models can be described as algebras and are—in a fuzzy way—homomorphic to reality (Lawvere et al. 2005; Kuhn 2007). The "fuzzy homomorphism" between experience and mental models which must be reflected in the verbal communication seems to be sufficient to converge into a common encoding over repeated experiences.

3.3.3 Communication

Despite the fact that we do not know exactly how humans learn their mother tongue (Eco 1976; Pinker 1995) it is an empirical observation that humans establish a consensus on the meaning of external signs. Human communication is possible, even though it is not perfect! Acquiring a language means to establish a correspondence between experiential concepts and constructions.

The representable signs are constructed as models of reality. These signs may be verbal descriptions, oral or written, computational models, sketches, etc. They are strongly inter-connected by operations and relations. The fact that initial language acquisition occurs in a simplified reality and within a supportive affective environment may significantly influence how the mechanism of language acquisition works.

3.3.4 Imperfections in communication

The meaning of a sign is defined in its context and this context can vary between sender and recipient of a sign. If a sign is unique to a context, no confusion within a context is possible, but for homonyms (same sign but different meaning) and polysemy, where the same word means different things in different contexts, a potential for imperfect communication exists. WorldNet (Fellbaum 1998) documents polysemy in natural language by separating different meanings of a word in synsets.

In normal communication circumstances, multiple contexts are combined. For example, participants in a meeting each have a subjective component in their context as well as a role influenced context; much discussion in meetings serves to align the contexts of the participants (Rottenbacher 2006).

3.3.5 Constructions are without imperfections

Constructions are, unlike observations (in tier1) descriptions or classification of objects (in tier2), without imperfection and error, as long as they are used in a fixed and shared context. As an everyday example, consider a description of the paper bill in (Figure 6).

Fig. 6. Example of socially constructed objects

The length of the paper is a physical (tier 2) observation: 134 mm, with a standard deviation of 3/100 mm. Once established that this is constructed as a Czech bill of 50 Czech Crowns; there is no uncertainty, the value in the Czech context is not 49.90 or 50.05 with any probability! If we leave the Czech context, then the value expressed in Euro may be uncertain, today the exchange rate is 33.2050 Crowns per Euro, which gives an approximate value of 1.50 Euro for the bill.

The value 50 is here—unlike the measured physical length—without error, directly contradicting the often heard statement that “all data contains some error”. It is correct only for tier 1 and 2, but not for the constructions of tier 3. The imperfections of tier 3 are introduced by

- establishing the connection between experiential concepts and constructions—the ‘subsumption’ of the law where one establishes whether a concrete act was ‘murder’ or ‘manslaughter’, and
- translating between contexts.

4 Compensation Improves Decisions with Imperfect GIS Data

When the data in a GIS are used to make decisions then the decision will be affected by the imperfections in the data. Rules of error propagation, fuzzy set, and Bayesian networks, have been developed to assess and to reduce the effects of imperfections on the decision (Morgan et al. 1998). These computational approaches are complex and require large amount of data. They reflect the hope that with complete and perfect information our decisions would be perfect and assume that more work will lead to better decisions (Gigerenzer et al. 1999). This assumption is not always true—complex models tend to "overfit", use elements presented in the past to predict the future incorrectly and decision makers may rely more on the data than warranted.

Real world decision making has always only limited resources of data and time available. Gigerenzer and his group have shown that despite such limitations, decisions are often correct. Human behavior is adapted to the world and exploits the structure of the world—physical and social—to make good decisions. Efforts that blindly try to improve the data independent of how it is used to make decisions related to the highly structured world, are misguided and a simple loss of resources.

In this section I show a few compensatory effects that are built into data collection and decision strategies that help to explain why simple strategies using imperfect data are still effective.

4.1 Correlation

The aspects of the world relevant for human life are highly spatially and temporally autocorrelated. The first law of geography by Waldo Tobler says: "Everything is related to everything else, but near things are more related than distant things." (Tobler 1970, 236). This has several compensatory effects:

- Measurement errors can be reduced by averaging.
- Desirable data that is not available can be replaced by correlated data.
- Only data of nearby objects and of similar spatial and temporal frequency as the decision is affecting is relevant. Processes that are much faster or much slower, or which are much larger or much smaller in space can be ignored.

Factors that influence a decision in a minor way can be neglected; especially the influence of the factor is small compared to the imperfections of other factors.

Surveyors observe more and more often than necessary to obtain a desired value. The redundancy created by repeated measures is used to, first, reduce the (statistical) imperfection in the value; second, to assess the level of statistical imperfections (i.e., the standard derivation of the PDF) and, third, to check against blunders and therewith to improve the reliability of the value.

4.2 Granulation and Classification

The mental classification of objects by humans is a complex (subjective) and multifaceted process. It is driven by the expected decision in and focuses observations to relevant aspects. The three processes, (1) to identify the boundary of an object, (2) to determine summary values, and (3) to classify mentally, are closely related and interact. Empirical evidence shows that mirror neurons (Rizzolati et al. 2002) found in humans and (at least) apes classify not only operations the cognizant agent visually perceives but also classify the objects with respect to having the right properties to be involved in an operation. Potential interactions between the agent and objects or interactions of interest between objects stipulate conditions these objects must fulfill, expressed as a property and a range for the value of the property. In this way operations of interest indicate what properties are important and these important properties are then used to determine boundaries of objects (Frank 2006; Kuhn 2007).

4.3 Constructions in Legal Decisions

Legal or administrative decisions do not tolerate errors. They are typically subject to review by others; how are arguments about uncertainty in the data avoided?

In the context of the legal or administrative decision, the data are well defined and without imperfections; the imperfections of the observations, granulation and classification are reduced by procedural description how the relevant properties must be observed and classified, often concluded by an authoritative statement about the subsumption of a real world fact under a construction(the rules of “due process” (Black 1996).

4.4 Absorption

Life is risky and many events cannot be predicted with accuracy at reasonable cost, but it is sometime possible to find somebody else to shoulder the risk for us. Bédard has called this 'uncertainty absorption' (Bédard 1986). It comes in many forms:

Insurance: I pay another party to cover the cost of errors in my decision (e.g., fire or flooding of my home).

Guarantee: Another party guarantees the subsumption that led to a construction. Certain data in some registers, for example, the German or Swiss cadastre (Grundbuch) is by law correct and the state guarantees it.

Liability: Another party is paid to make a subsumption for me and is liable for the risks involved. Many types of certificates of professionals, e.g., building inspections certificates, include liability of the professional making the judgment. The professional often carries insurance to cover the rare events of error.

4.5 Linguistic Classification and Communication

We talk about buildings, dogs, and trees, implicitly classifying the objects we see as belonging to the class of Buildings, Dogs, or Trees. Such classifications are different from mental classifications; they are called radial categories and show prototype effects (Rosch 1978). The same word in multiple contexts has overlapping applicabilities, which share a core meaning (Figure 7).

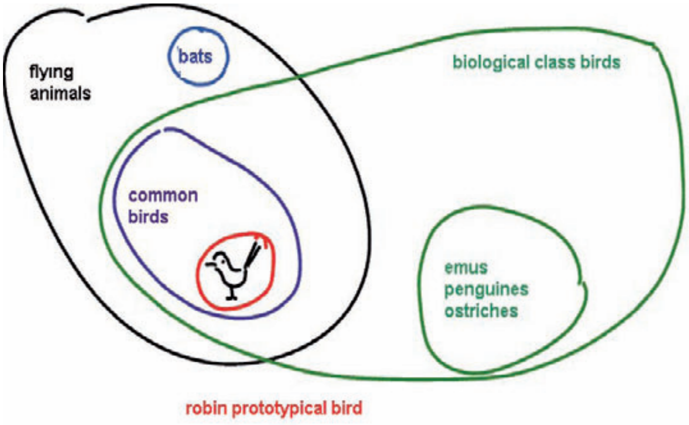


Fig. 7. Different meaning of 'Bird' in different contexts

Exemplars in the area of overlap are coded in each context with the same code. Other exemplars are coded differently, depending on the context.

Boundaries between natural language concepts (not legal or scientific constructs) tend to be at natural breaks: the distinction between cats and dogs is unproblematic, because no intermediate individuals exist (horses and donkeys provide one of the rare counterexamples!). The rich structure of reality as we experience it is approximately shared by all humans (Lakoff 1987), because they share a large part of daily experiences (eating, drinking, sleeping, etc.). This approximation of the experiential grounding is usually sufficient to establish a mapping between structure encountered in a text received and our own structure.

The theory of supervaluation gives guidelines how to deal with the integration of multiple context and reasoning in an integrated data collection. If the semantics of the context are available as formally described ontologies (e.g., in OWL (Dieckmann 2003)) then formal conversions between codes from different contexts can be attempted. The translation between contexts must always relate the signs in the context back to the grounding items and then forward to the sign of interest in the other context, but general solutions are not yet known.

4.6 Safety Margins for Decisions

If you have 2 liter of gas in the tank of your car, which uses 10 l/100km, you will not pass a filling station when you know that the next is 17 km later. We should be able to reach the next station, but the safety margin seems too low to risk to remain with no gas on the road. Similarly, most engineering decisions include safety margins—in our example a mental rule that says for example 'you should always have 5 l gas in your car'. Safety margins in engineering decisions are factors that increase the load and decreases the resistance of the materials assumed, and can be translated to a probability of failure (often 1 or 2% failure probability is acceptable) (Schneider 1995; Frank to appear 2008).

5 Conclusion

A systematic study of how imperfections are introduced in the data can be used to separate three large groups of types of imperfections

- probability related to observations of physical reality,

- fuzzy set descriptions of mental classification and subsumption, and
- constructions in contexts.

Considering that humans have adapted their methods to make decisions to their environment, what Gigerenzer calls ecological rationality, even with the imperfect data in a GIS useful decisions can be made. Several methods are used to reduce the negative effects of imperfections on decisions, some based on properties of physical reality (e.g., correlation) and others socially engineered (e.g., constructions).

Acknowledgement

I appreciated the helpful comments of three anonymous reviewers.

References

Abler R (1987) Review of the Federal Research Agenda. International Geographic Information Systems (IGIS) Symposium (IGIS'87): The Research Agenda, Arlington, VA

Black HC (1996) Black's Law Dictionary, West Publishing

Burrough PA (1996) Natural Objects with Indeterminate Boundaries. Geographic Objects with Indeterminate Boundaries. P. A. Burrough and A. U. Frank. London, Taylor and Francis: 3-28

Couclelis H (1992) People Manipulate Objects (but Cultivate Fields): Beyond the Raster-Vector Debate in GIS. Theories and Methods of Spatio-Temporal Reasoning in Geographic Space. A. U. Frank, I. Campari and U. Formentini. Berlin, Springer-Verlag. 639: 65-77

Dieckmann J (2003) DAML+OIL und OWL XML-Sprachen für Ontologien. Berlin: 21

Eco U (1976) A Theory of Semiotics. Bloomington, Indiana University Press

Egenhofer M, Frank AU (1986) Connection between Local and Regional: Additional "Intelligence" Needed. FIG XVIII International Congress of Surveyors, Toronto, Canada (June 1-11, 1986)

Fellbaum C Ed. (1998) WordNet: An Electronic Lexical Database. Language, Speech, and Communication. Cambridge, Mass., The MIT Press

Frank AU (2001) Tiers of Ontology and Consistency Constraints in Geographic Information Systems. International Journal of Geographical Information Science 75(5 (Special Issue on Ontology of Geographic Information)): 667-678

Frank AU (2003) Ontology for Spatio-Temporal Databases. Spatiotemporal Databases: The Chorochronos Approach. M. Koubarakis, T. Sellis and e. al. Berlin, Springer-Verlag. 2520: 9-78

Frank AU (2006) Distinctions Produce a Taxonomic Lattice: Are These the Units of Mentalese? International Conference on Formal Ontology in Information Systems (FOIS), Baltimore, Maryland, IOS Press

Frank AU (to appear 2008) Analysis of Dependence of Decision Quality on Data Quality. Journal of Geographical Systems: 18

Ganter B, Stumme G, Wille R, Eds. (2005) Formal Concept Analysis Foundations and Applications. Berlin, Heidelberg, Springer

Gibson JJ (1986) The Ecological Approach to Visual Perception. Hillsdale, NJ, Lawrence Erlbaum

Gigerenzer G, Todd PM, e. al. (1999) Simple Heuristics That Make Us Smart. New York, Oxford University Press

Goodchild MF, Egenhofer MJ, Kemp KK, Mark DM, Sheppard E (1999) Introduction to the Varenius Project. International Journal of Geographical Information Science 13(8): 731-745

Guarino N, Welty C (2000) Ontological Analysis of Taxonomic Relationships. Proceedings of ER-2000, 19th Int. Conference on Conceptual Modeling. E. Laender and V. Storey, Springer-Verlag

Heidegger M (1927; reprint 1993) Sein und Zeit. Tübingen, Niemeyer

Husserl (1900/01) Logische Untersuchungen. Halle, M. Niemeyer

Kuhn W (2007) An Image-Schematic Account of Spatial Categories. 8th International Conference, COSIT 2007, Melbourne, Australia, Springer

Lakoff G (1987) Women, Fire, and Dangerous Things: What Categories Reveal About the Mind. Chicago, IL, University of Chicago Press

Lawvere FW, Schanuel SH (2005) Conceptual Mathematics: A First Introduction to Categories. Cambridge, Cambridge University Press

McCarthy J, Hayes PJ (1969) Some Philosophical Problems from the Standpoint of Artificial Intelligence. Machine Intelligence 4. B. Meltzer and D. Michie. Edinburgh, Edinburgh University Press: 463-502

Morgan MG, Henrion M (1998) Uncertainty A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis, Cambridge University Press

NCGIA (1989a) The U.S. National Center for Geographic Information and Analysis: An Overview of the Agenda for Research and Education. 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, Carver S (1991) Error Propagation: A Monte Carlo Simulation. Handling Geographical Information. I. Masser and M. Blakemore. Essex, Longman Scientific & Technical. 1: 78-101

Pinker, S. (1995). The Language Instinct. New York, HarperPerennial

Randow, G. v. (1992). Das Ziegenproblem - Denken in Wahrscheinlichkeiten. Hamburg, Rowohlt Taschenbuchverlag

Raubal M (2002) Wayfinding in Built Environments: The Case of Airports. Münster, Solingen, Institut für Geoinformatik, Institut für Geoinformation

Rizzolati G, Craighero L, Fadiga L (2002) The Mirror System in Humans. Mirror Neurons and the Evolution of Brain and Language. M. Stamenov and V. Gallese, John Benjamins Publishing Company: 37-59

Rosch E (1978) Principles of Categorization. Cognition and Categorization. E. Rosch and B. B. Lloyd. Hillsdale, NJ, Erlbaum

Rottenbacher C (2006) Bewegter Planungsprozess. Vienna, Technical University Vienna. PhD

Sartre JP (1943; translated reprint 1993) Being And Nothingness. New York, Washington Square Press

Schneider M (1995) Spatial Data Types for Database Systems. Hagen, FernUniversitaet

Searle JR, Ed. (1995) The Construction of Social Reality. New York, The Free Press

Simon H (1956) Rational Choice and the Structure of the Environment. Psychological Review 63: 129-138

Smith B, Grenon P (2004) "SNAP and SPAN: Towards Dynamic Spatial Ontology." Spatial Cognition and Computing(4): 69-103

Tobler WR (1970) A Computer Model Simulation of Urban Growth in the Detroit Region. Economic Geography 46(2): 234-240

Tomlin CD (1983) A Map Algebra. Harvard Computer Graphics Conference, Cambridge, Mass

Wittgenstein L (1960) Tractatus logico-philosophicus. London, Routledge & Kegan Paul

Zadeh LA (1974) Fuzzy Logic and Its Application to Approximate Reasoning. Information Processing

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. Data Mining, Rough Sets and Granular Computing. T. Y. Lin, Y. Y. Yao and L. A. Zadeh. Heidelberg, Germany, Physica-Verlag GmbH: 3-20

Zaibert L, Smith B (2004) Real Estate - Foundations of the Ontology of Property. The Ontology and Modelling of Real Estate Transactions: European Jurisdictions. H. Stuckenschmidt, E. Stubkjaer and C. Schlieder, Ashgate Pub Ltd: 35-51

Moving from Pixels to Parcels: the Use of Possibility Theory to Explore the Uncertainty Associated object Oriented Remote Sensing

Alexis Comber[1], Alan Brown[2], Katie Medcalf[3], Richard Lucas[4], Daniel Clewley[3], Johanna Breyer[4], Peter Bunting[4], Steve Keyworth[3]

[1] Department of Geography, University of Leicester, Leicester, LE1 7RH, UK, E-mail: ajc36@le.ac.uk
[2] Countryside Council for Wales, Bangor, LL57 2LQ, UK.
[3] Environment Systems, 8G Cefn Llan Science Park, Aberystwyth, SY23 3AH UK
[4] Institute of Geography and Earth Sciences, Aberystwyth University, SY23 3DB, UK

Abstract

This paper explores the issues relating to uncertainty in the application of object oriented classifications of remote sensing data. Object oriented remote sensing software such as eCognition (now known as Definiens Developer) provides the user with flexibility in the way that data is classified through segmentation routines and user-specified fuzzy rules. However the aggregation of fuzzy data objects such as pixels to higher level parcels for the purpose of policy reporting is not straightforward. This paper explores the uncertainty issues relating to the aggregation from fine detailed (uncertain) objects of one classification system to coarser grain (uncertain) objects of another classification scheme. We show Possibility Theory to be an appropriate formalism for managing the non-additive uncertainty commonly associated with classified remote sensing data. Results are presented for a small area of upland Wales to illustrate the value of the approach.

Key words: Possibility Theory, Land Cover, Uncertainty

1. Introduction

Object oriented classification of remotely sensed imagery allows the user to manipulate groups of pixels (or objects) that have been segmented from image data. Definiens Developer, previously eCognition, (Definiens, 2008) is a recent software development for the analysis of remote sensing data. It segments the imagery and uses a hierarchical rule-based approach to classify the segmented objects, identifying features by using ancillary data in the segmentation procedure. Objects are identified from multi-resolution data in combination with ancillary information relating to spatial context, topology and position in the classification hierarchy which guide or constrain the segmentation process. Rules relating to the analysis of image and ancillary data are encoded in a “knowledge base” which specifies when they applied in the classification process. In this way segmented image objects may reflect more intuitively the features of interest on the ground than traditional, pixel based classifications. Lucas et al. (2007) note that Definiens Developer has a number of benefits which include being able to manipulate the sequence of rule application to better reflect manual classification procedures and is more heuristic than other approaches. In this way the segmentation procedures are able to better represent ‘reality’ as perceived by ecologists, field surveyors and air-photo interpreters than others remote sensing approaches. This is in contrast to many traditional remote sensing classification techniques which extract spectral information for all classes and use these simultaneously to train, for example, a maximum likelihood or minimum distance algorithm. However, the increased availability and uptake of object oriented classification software such as Definiens Developer presents the remote sensing community with a new set of problems. These relate to the use of measures of object uncertainty, the classification hierarchy and spatial characteristics of aggregation units used to summarise the data.

Increasingly classification schemes for land cover and habitats are determined by policy and legislation. As a result many recent mapping initiatives as commissioned by various institutions specify the need to generate a number of products from analysis of the same remotely sensed data and, due to its additional functionality, the use of object oriented classification approaches. Recent work by Comber (2006, 2008, forthcoming) has developed and applied a ‘*context-sensitive’* approach to land cover, vegetation and habitat mappings. Context sensitive mapping supports alternative realisations of the same remotely sensed data and allows different questions to be answered. These include questions relating to conservation (what is there), restoration (what could be there) and monitoring (what has happened there). The approach allows different policy objectives to be fulfilled and questions relating to different types of decision to be answered. The resulting maps are called *'context-sensitive'* because they have been produced to meet a particular need. For example, some habitats like 'Bog' are protected but the mapping of Bog is subject to uncertainty as it's constituent plant communities overlap with other habitats such as 'Heaths'. There may be patches of bog vegetation within the upper bound of the potential spatial extent of bog which are too small to be mapped independently of the surrounding heath vegetation. Quantifying the 'possible' or 'uncertain' spatial extent of Bog vegetation communities for habitat restoration objectives therefore is a different question to identifying the 'certain' spatial extent of Bog for legal reasons, such as prosecuting land owners for illegal burning on protected habitats.

The need for alternative realisations of the same data allows different landscape questions to be answered for different policy objectives: fulfilment of habitat reporting obligations, Common Agricultural Policy payments, monitoring of change, conservation, habitat restoration, etc. Policy makers also need land information at different levels and degrees of granularity: typically coarser information to be able report upwards (e.g. to fulfil EU or federal obligations) and finer information for local land management decisions. Recent habitat mapping projects commissioned by the Countryside Council for Wales have specified the need for the habitat analyses to report different information from the same remotely sensed data. Context sensitive mapping provides answers to these various policy objectives by presenting the information derived from analyses of the remotely sensed imagery at different scales of aggregation (summary) with different confidences according to the specific task.

The issue of how to manage the uncertainty associated with aggregating to different data scales and different information granularity (i.e. moving up and down) has not hitherto been addressed in the remote sensing literature. In this paper we explore the uncertainty associated with aggregating from a fuzzy classification at a fine spatial scale and fine thematic granularity (as needed by ecologists) to a coarser scale and coarser thematic grain (as needed for policy reporting).

2. Background

Traditional, pixel based techniques for classifying remotely sensed imagery typically match the characteristics of each pixel either to a set of predefined classes (supervised classification) or to clusters of similar pixels (unsupervised classification) for a set number of predefined classes. Each pixel is evaluated for membership to each class and membership (fuzzy or crisp) is allocated based on a specified distance function, usually Euclidian. In these approaches each class and each object (or pixel) is treated in the same way.

A hierarchical object oriented classification is slightly different. It imposes a classification order where finely grained classes (i.e. those at the bottom of a classification tree) can be allocated to objects only after coarser grained classes higher up in the hierarchy have been allocated. This allows inheritance as attributes are passed down the hierarchy and ‘child’ classes inherit characteristics from their ‘parents’. Inheritance contributes to the definition of each class as the semantics of the classes are necessarily defined not only by the rules that exist in the knowledge-base at their ‘level’ in the hierarchy but also by the attributes they inherit.

Information about habitats is collected by government agencies at a range of spatial scales and thematic granularities in order to fulfil national and international policy obligations. At the most detailed grain is National Vegetation Classification (NVC) or Phase II habitats. The NVC classifies vegetation into 286 communities grouped into 25 woodlands, 22 heaths, 38 mires etc. Communities were originally defined by a statistical analysis of species occurrence and abundance in ~70,000 samples. The full description of the communities is provided in the five volumes by Rodwell et al., (1991 et seq) and are based an expert botanists interpretation of the environment in which the communities are found. Phase I habitats capture information at a coarser grain which is required in order to fulfil the objectives of the UK Biodiversity Action Plan (UK Government, 1994). Guidance for the identification and recording of Phase I habitats in the field is given by the Joint Nature Conservancy Council (JNCC, 1990). The class descriptions, specification and semantics of Phase I habitats are predicated on the assumption that the data will be collected in the field rather than extracted from satellite imagery in the laboratory. The guidance on Phase I habitats (JNCC, 1990) was written for field surveyors and as such does not define any minimum mapping unit.

In this work a series of end members or data primitives have been identified in Definiens Developer. These are segments of single 5m pixels and represent the most ecologically detailed information that it has been possible to extract from the image data. For one of the final map product these will be aggregated into Phase I habitats. However there is considerable interest in the end-members themselves many of them are equivalent to NVC vegetation communities and are close to the thematic granularity of Phase II habitat classes which are politically and ecologically very important. The mapping described in this work combines end-members (or data primitives) to identify higher level ‘habitats’ such as Priority habitats (Jones et al., 2003) and Phase I habitats (JNCC, 1990). It builds on the work of Lucas et al., (2007) and is part of a project that is mapping Phase I habitats for the whole of Wales.

We explore the influence of the uncertainty associated with constituent end member values on the final mappings at a higher spatial and thematic level.

3. Problem

The end members identified using Definiens Developer are to be aggregated into Phase I habitats. However, the same end members can be components for a number of different habitats. Table 1 shows four example habitats and the spectral end members, different combinations of which make up each of the habitats. The spectral end members are the building blocks from which the higher level habitats are constructed. The Phase I habitats are defined in terms of their species composition and vegetation communities. Example spectral end members and how they relate to Phase I habitats are shown in Table 1.

The problem is further complicated by the ability of object oriented software to compute fuzzy memberships for each of the end members based on the extent to which the rules used to identify different them are satisfied. The result is, as in fuzzy classification, a membership value to each end member class for each pixel (they are treated as fine scaled objects). Figure 2 shows a small hypothetical area of 25 pixels by way to illustrate this problem. The uncertainty attached to each pixel for each of the 4 spectral end members imposes a further level of complexity and uncertainty to the problem of combining end-members to generate higher level classifications: how to combine the fuzzy memberships of the spatially and thematically more detailed (sub-) classes and how to determine the presence of any given Phase I habitat given the uncertainty of the end members.

Table 1. Example habitats and the proportions of their constituent spectral end members

	D.1.1 Dry Acid Heath		D.2 Wet Heath		E.1.6.1 Blanket Bog		E.1.6.2 Raised Bog	
Spectral End-members	AND (max)	OR (min)	AND (max)	OR (min)	AND (max)	OR (min)	AND (max)	OR (min)
Blanket Bog						>=0.1		
Bog Moss	<=0.05					>=0.1		
Bogs			<=0.1			>=0.25		>=0.1
Calluna		>=0.25						>=0.1
Calluna				>=0.25				
Cotton Grass	<=0.05					>=0.1		>=0.1
Festuca-Agrostis			<=0.01					
Heathy bog	<=0.05		>=0.25			>=0.25		>=0.1
Jsq-nardus				>=0.1				
Molina	<=0.25	>=0.1			<=0.1		<=0.25	
Mossy Fescue			<=0.01					
Vaccin-ium		>=0.25						

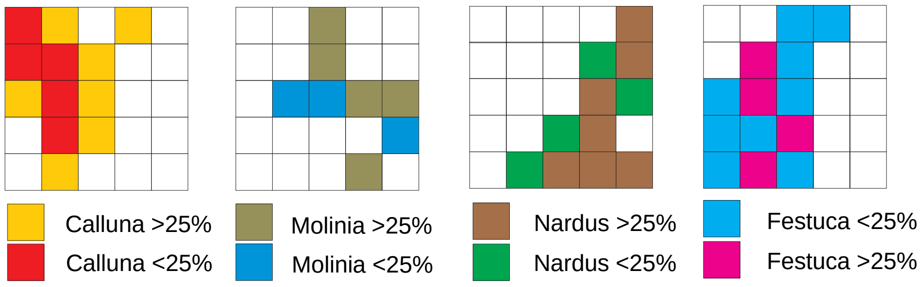


Fig. 1. An illustration of hypothetical overlapping spectral end members.

4. Method

4.1 Possibility Theory

The problem was how to combine uncertain pixel (lower level object) data with a membership values to a lower level classification in order to generate a second, higher level classification, whilst managing the uncertainty associated with each low level object. The approach taken to resolve it was to use Possibility Theory. Possibility Theory handles incomplete information using a pair of dual set-functions 'Possibility' and 'Necessity'. These are similar to the Belief and Plausibility functions in the Dempster-Shafer theory of evidence, except that whilst Dempster-Shafer is additive, Possibility is not. Instead it uses a supremum function, relating to the maximum support for any given hypothesis.

The knowledge base (a set of rules relating to image data, vector layers, spatial configuration, etc) provides the statistical model which in the object oriented classification generates a likelihood of a value between 0 and 1 for each object (in this case, pixels). The objective was to convert this into a degree of truth for a higher level object (a Phase I parcel) in terms of its membership to a higher level Phase I habitat class. For instance, the higher level object class may be regarded as 'true' if it exceeds a membership threshold. In every case, the class membership at the sub level (end member) is determined by a set of rules using remote sensing data layers and other data as input.

For each of the lower sub level classes, the classification process is a function that generates a membership for each pixel *x*:

$$F(x) \rightarrow [0, 1]$$

The higher level habitat object (parcel) ***X***, is composed of set *{w} = x1, x2, x3,..., xN* representing the spectral end member classes in this Phase I class (such as those in Table 1 for the 4 example Phase I habitats). If all the rules for the Phase I class are satisfied, the overall Possibility for that class can be assessed. The Possibility (*Poss*) function of Dubois and Prade (Dubois and Prade, 2001, who provide an excellent overview text on this topic) states that *Poss(**X**)* is the supremum of *Poss({w})*, where *w* are elements of set ***X***.

The uncertainty associated with ***X*** is given by the corresponding Necessity function, which is similar to the Belief-Plausibility pair in Dempster-Shafer, where the difference between the two functions provides a measure of the uncertainty of the inference. Necessity is defined as:

$Nec(\boldsymbol{X}) = 1 - Max(\ Poss(\neg \boldsymbol{X})\ )$ Equation 1

4.2 Application to the Data

A GIS data layer of empty parcels was generated in Definiens Developer. Each parcel was populated with Phase I habitats by assessing the intersecting pixels in the following way:

- A threshold of 0.1 was applied to the pixel membership functions;
- A zonal statistics GIS function was used to generate information about the number of pixels and their values for each spectral end member class;
- The distribution of the pixels and their memberships values in the parcel was compared with the rules for that Phase I class, as in Table 1;
- The Possibility function for each Phase I habitat was calculated from the proportion represented by the supremum of *Poss({w})*;
- The Necessity function was calculated using Equation 1.

5. Results

Results of initial Phase I classification are shown for a small upland area in Wales. Figure 2 illustrates the problem of overlapping spectral end member classes and how their distributions relate to the Phase I parcels.

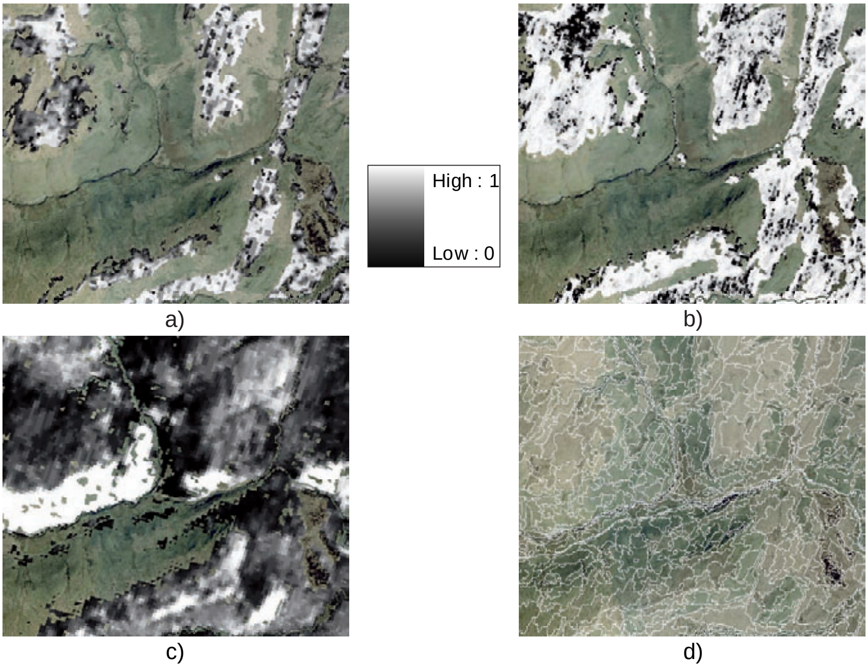


Fig. 2. Examples of the overlap for three spectral end member classes a) Blanket Bog, b) Molinia c) Festuca and d) the Phase I habitat parcels.

Figure 3 shows the distribution of the Phase I Mires (Blanket Bog, Raised Bog, Valley Mire, Basin Mire) and of the Heaths (Dry Acid Heath and Wet Heath). These examples show that by applying the rules to the spectral end member classes it is possible to generate Phase I habitat land covers.

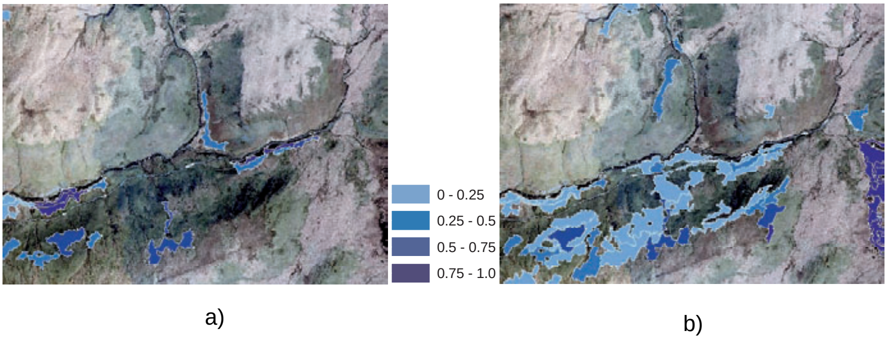


Fig. 3. The overlapping distributions and Possibility of a) Heath and b) Mire Phase I habitats

More interestingly the use of the Possibility and Necessity functions allows the uncertainty associated with those classifications to be explored, providing the basis of alternative realisations of the same data. Figure 5 shows the Necessity functions for the Heath classes and indicates the uncertainty associated with the parcels.

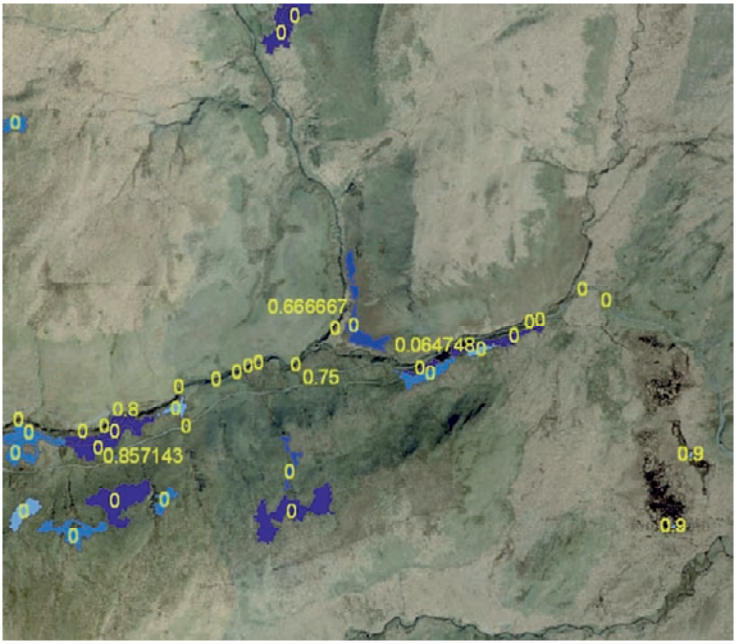


Fig. 5. The Heath parcels labelled with their Necessity values as a measure of the uncertainty associated with the classification

Finally, it is possible to map the ‘conflicts’ where more than one class has support greater than a threshold. Figure 6 shows the distribution of conflicts in the test area. The identification of conflicts indicates the locations where individual field surveyors may disagree over the class allocation, where there is considerable heterogeneity in the pixel memberships to spectral end member classes or where the rule base may need to be improved.

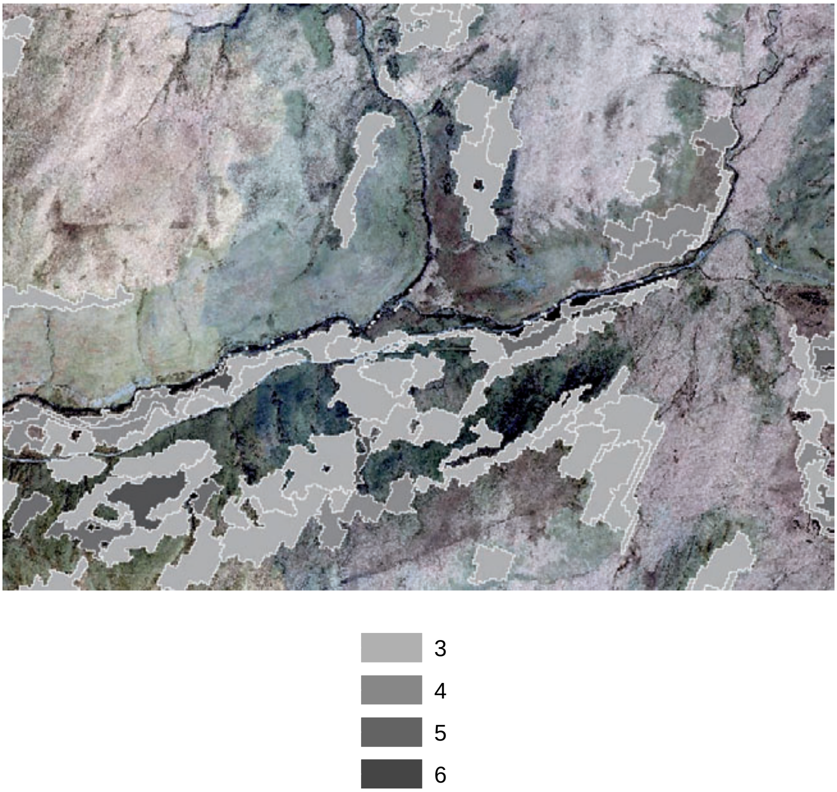


Fig. **6.** The distribution and number of ‘conflicts’ where more than one class exists

6. Discussion

Information on habitats such as Phase I habitats are needed by policy makers in order to fulfil national habitat monitoring objectives. It is important to state that the results of this work have not yet been validated – the purpose of the paper to is to introduce the problem and a possible solution. However the results and techniques reported are of relevance to a number of areas of work. First, the remote sensing community have embraced object oriented image classification software such as Definiens Developer, as reflected in the number of paper reporting its use in remote sensing journals. Hitherto, analysis and the development of techniques relating to issues of uncertainty in remote sensing have focused on the classification of the pixel. As a result fuzzy pixel classifications, once the preserve of a small group of researchers, have become mainstream techniques. The use of object oriented classifiers presents a new problem: how to move from the (lower level) pixel to the (higher level) parcel whilst managing the uncertainty associated with the lower level objects. Typically the parcel is 'crisped off' (i.e. given a single class) with some reporting of the uncertainty – see Fuller et al., 2002 as an example of some of the difficulties of trying to operationalise this for a national land cover project. However, policy makers are increasingly aware that there may be more than one way of presenting the data and are increasingly interested in additional applications and 'context sensitive' mappings – i.e. ones that are able to respond to a number of different policy questions.

Second, the ability to be able to represent and explore the uncertainty associated with allocating a parcel to any one Phase I class is important, especially in upland areas. The uplands presents a specific problem to remote sensing analyses, whether a conventional pixel based classification approach is used or an object oriented one as described here. The major issue is one of heterogeneity (Comber et al., 2004a). Any mapping of unmanaged, upland landscapes has to manage a number of inter-related issues. Principal among these are that different combinations of the same plant communities make up the same land cover classes such as the Phase I. This means that it is possible to quite correctly allocate any given area of land to a number of different classes and results in the variation commonly found in field surveys of vegetation (see Cherrill and McClean, 1999). Historically land cover was manually mapped either in the field or from aerial photography and the particular allocation depended on the individuals constructing the map: their experience, institutional context, disciplinary background etc. However, the advent of machine processing and remotely sensed imagery in the 1970s standardised these processes to some degree and land cover classification was performed on the spectral properties of the imagery. With the advent of finer resolution data and object classification the ability to mimic the human expert has increased, providing a contextual richness to the classification process.

The third aspect of this work has been the use of Dubois and Prade's Possibility Theory. In previous work, Comber has explored the use of a number of formalisms for managing the uncertainty associated with land cover classifications. These include Dempster-Shafer, Bayes, and

Endorsement theories (Comber et al., 2004b, 2004c). The advantage that Possibility Theory has over these other approaches is that where they are additive (beliefs or probabilities are combined with each other) Possibility and Necessity are not. Instead they apply a supremum function to the memberships and intuitively this makes more sense when we are seeking to combine thematic data (as represented by the spectral end member or lower level objects). Calculating the Necessity for each parcel allows an extended error matrix to be developed and the application of the techniques described in Fisher et al., (2006). Similarly, the Necessity parameterises the Possibility and gives an indication of the upper and lower bounds of the possible extent of different habitats. Ongoing work will develop these ideas.

Acknowledgements

The British National Space Centre co-funded a preparatory study under the GIFTSS programme, steered and generously supported by Dr Ian Thomas and Matt O'Donnell. The underlying classification work for this paper was conducted within the Countryside Council for Wales (CCW) contract, Mapping Habitats Across Wales Using Satellite Imagery (Contract No. FC 73-03-299). The authors would like particularly to thank Alex Turner, Stuart Mackintosh and Jane Stevens (Countryside Council for Wales) for their time and ecological discussions. Claire Horton (Welsh Assembly Government) is particularly thanked for her support and for providing access to key data.

References

Cherrill, A. and McClean, C. (1999). “Between-observer variation in the application of a standard method of habitat mapping by environmental consultants in the UK”, *Journal of Applied Ecology*, **36**, 989–1008.

Comber, A.J., Law, A.N.R., Lishman, J.R., (2004a). Application of knowledge for automated land cover change monitoring. *International Journal of Remote Sensing*, 25(16): 3177-3192.

Comber, A., Fisher, P., Wadsworth, R. (2004b). “Integrating land cover data with different ontologies: identifying change from inconsistency”, International Journal of Geographical Information Science, 18(7), 691-708.

Comber, A.J., Law, A.N.R., Lishman, J.R. (2004c). “A comparison of Bayes', Dempster-Shafter and endorsement theories for managing knowledge uncertainty in the context of land cover monitoring", Computers, Environment and Urban Systems, 28(4), 311-327

Comber A.J, Fisher, P.F. and Brown, A. (forthcoming). "Uncertainty, vagueness and indiscernibility: the impact of spatial scale in relation to the landscape elements", paper accepted for publication in ISPRS

Definiens (2008). http://www.definiens.com/ [available 10th January 2008]

Dubois, D. and Prade H. (2001). "Possibility theory, probability theory and multiple-valued logics: A clarification", Annals of Mathematics and Artificial Intelligence, 32, 35–66.

Fisher, P, Arnot, C, Wadsworth, R and Wellens, J, (2006). "Detecting change in vague interpretations of landscapes", Ecological Informatics 1(2), 163-178.

Fuller, R. M., Smith, G. M., Sanderson, J. M., Hill, R. A., and Thomson, A. G. (2002). "Land Cover Map 2000: construction of a parcel-based vector map from satellite images", Cartographic Journal, 39, 15–25.

JNCC (1990). Handbook for Phase 1 Habitat Survey: Handbook and Field Manual: Technique for Environmental Audit, Joint Nature Conservation Committee (JNCC), Peterborough

Jones, P.S., Stevens, D.P., Blackstock, T.H., Burrows, C.R. and Howe, E.A. (2003). Priority habitats of Wales: a technical guide. Countryside Council for Wales, Bangor.

Lucas, R., Rowlands, A., Brown, A., Keyworth, S. and Bunting, P. (2007). "Rule-based classification of multi-temporal satellite imagery for habitat and agricultural land cover mapping", ISPRS Journal of Photogrammetry and Remote Sensing, 62(3), 165-185.

Rodwell J.S. (editor), (1991). British Plant Communities. 5 volumes. Cambridge University Press. Cambridge.

UK Government, 1994. *Biodiversity: the UK Action Plan*, HMSO, London.

Data Matching – a Matter of Belief

Ana-Maria Olteanu Raimond, Sébastien Mustière

IGN, COGIT Laboratory, 2/4 Avenue Pasteur,
94165 Saint-Mandé cedex, France,
email: {ana-maria.olteanu; sebastien.mustiere}@ign.fr

Abstract

Nowadays, it is often that a geographic area is described by several independent geographic databases. Yet users need to fusion various information coming from these databases. In order to integrate databases, redundancy and inconsistency between data should be identified. Many steps are required to finalise the databases integration, in particular automatic data matching. In this paper, one approach of matching geographic data bearing on the belief theory is presented. This approach consists in combining criteria from knowledge such as geometry, orientation, nature of roads, names and topology. Then it is tested on heterogeneous network representing roads.

Keywords: data matching, networks, belief function, fusion, topology

1. Introduction

Generally speaking, capturing and managing efficiently geographic data was certainly the most important challenge for geographic information science in the last decades. Nowadays, the co-existence of many and heterogeneous geographical data covering the same areas raises the need to study how these data may be integrated together (Kilpelaïnen 2000; Hampe and

Sester 2002; Mustière and van Smaalen 2007). The relatively new emergence of spatial data infrastructures is a good evidence of this increasing need (Rajabifard et al. 2006). More concretely, integrating data is first necessary to efficiently combine them, which allows data users to make fruitful analyses and data producers to build rich but relatively low price products. It is also necessary to propagate updates from reference databases to other databases. It is finally useful to compare data and perform some quality analyses.

Databases integration first requires the integration of their schemas, which consists in identifying and modelling links between homologous classes and attributes in the schemas (Devogele et al. 2001; Uitermark 2001). Complete databases integration also requires to identifying homologous objects in the data or, in other words, requires *data matching* (Devogele 1997; Walter and Fritsch 1999; Beeri et al. 2004, Samal et al.2004, Mustière and Devogele 2008). Data matching is an easy process when an universal identifier is defined for the modelled features, independently of their representation in a database. This is the case for books for example, with their ISBN identifier. However, for any data where such identifiers do not exist, such the geographic data, data matching is often a complex task relying on comparison of different properties of objects.

In this paper, we present an approach for data matching based on the belief theory. We then apply this matching approach to the complex case of matching heterogeneous geographic networks with different levels of detail.

The paper is organised as follows. In the next section we briefly explain why we need to develop a new matching based on the belief theory. In Section 3, we introduce the belief theory and we present how we apply it for matching data in general and for matching networks representing roads in particular. Finally, before concluding, we show in Section 4 the results of our approach on two heterogeneous road networks.

2. Geographic data matching

Typical matching of non-geographic data relies on two types of objects properties. The first one is the comparison of pseudo-identifiers, which are most of the time represented by textual attributes like the name of a person. This is done by string comparison measures. Most of them rely on the so-called "edit-distance" or "Levenshtein distance" (Levenshtein 1965).

The second type of compared properties is the nature of objects. This is usually done by means of "semantic distances" between concepts represented in one or several ontologies (Rodriguez and Egenhofer 2003).

When a geographic feature is represented by means of different objects in distinct databases, location is the most pregnant and invariant property: objects located at similar places usually represent the same feature and are thus homologous. Consequently, geographic data matching approaches do usually focus on the comparison of locations of objects. If objects have a detailed shape represented by lines or polygons, geometrical measures can also be compared like the orientation or shape. Especially for networks, important properties of objects are also their spatial relationships with other objects, and particularly topological relationships.

In this context, most approaches for geographic data matching (Beeri et al. 2004; Samal et al 2004) and mainly networks matching heavily rely on the comparison of the geometrical and topological properties of objects (Walter and Fritsch 1999; Voltz 2007; Zhang et al, 2007; Mustière and Devogele 2008). Nevertheless, most of these processes do not holistically consider both spatial and non-spatial properties. One of the reasons of that is the difficulty to set up efficient and generic ways to combine different information while matching. The main consequence is that the matching processes are ad-hoc processes, adapted to particular cases where the heterogeneity of data to match is limited. In particular, most approaches for matching networks are adapted to networks with close levels of details (Walter and Fritsch 1999; Voltz 2007; Zhang et al, 2007).

We developed in previous works a process for matching networks with different levels of details (Devogele 1997; Mustière and Devogele 2008). This process is made of some steps roughly matching nodes together, some steps roughly matching arcs, and then some final steps combining the previous pre-matching to make the final decisions of matching. It proved to be efficient on various themes (Mustière 2006; Mustière and Devogele 2008).

However, this process has only been applied in practice to networks coming from the same producer, limiting their heterogeneity. An important insight from experiments of this process is that we sometimes needed to develop some simple but ad-hoc patches to take into account non-spatial properties, to get the best results for different networks like railways, roads, rivers or electric lines (Mustière 2006, Mustière and Devogele 2008). To go further, an efficient process should be able to simply take into account several properties (like the name, position, type or width of roads for example). Another insight from the results of our previous approach is that the efficiency of the process depends on its ability to manage imperfections encountered in the data. Indeed, values of properties may be imprecise, erroneous or even missing, and these imperfections make the process difficult.

In order to overcome these limitations, i.e. have a more generic, efficient and independent of imperfections matching process, we propose in this paper another data matching approach, based on the belief theory, also called the Dempster-Shafer model. It has been introduced by (Dempster 1968) and proved to be efficient to model imperfect knowledge by (Shafer 1976). Its major advantages in our matching context are as following:

- it allows to model imperfect knowledge, be it imprecise, uncertain or incomplete,
- it provides techniques to efficiently combine sources of knowledge in order to make decisions,
- it explicitly manages conflicts between sources of knowledge.

Our matching approach based on the belief theory has already been proved to be efficient for isolated data (Olteanu 2007). We present in the next section its main principles and how we adapt it to networks matching.

3. Matching Approach based on the Belief Theory

3.1 The general framework of the Belief Theory

The belief theory first supposes the definition of a frame of discernment $\Theta=\{H_1, H_2,\ldots, H_N\}$, which is a finite set of N hypotheses corresponding to the potential solutions of a given problem. From this frame of discernment, let us define 2^Θ, the set of all subsets of Θ defined by:

$$2^\Theta = \{\phi,\{H_1\},\{H_2\},\{H_1,H_2\}...\{H_1...H_N\},\Theta\} \quad (1)$$

where a subset $\{H_i, H_j\}$ represents the proposition that the solution of a problem is one of these hypothesis, i.e. either H_i or H_j. A key point of the belief theory is the basic belief assignment, i.e. a function that assigns to a proposition P, with $P\in 2^\Theta$, a value named the mass of belief, noted *m(P)*, that represents how much a criterion-called source of information-believes in it. For example, let us consider a process of data matching based on distances between features and a proposition stating that two given features are homologous. The closer the two features are, the stronger the criterion believes that the features are homologous, and thus the more the value of the mass of belief is important.

A basic belief assignment is a function $m : 2^{\Theta} \rightarrow [0,1]$ such that:

$$\sum_{P \subseteq \Theta} m(P) = 1 \quad (2)$$

The belief theory offers tools to combine several sources of information such as the Dempster’s rule of combination (Shafer 1976). This rule defines how to combine several belief functions in order to determine a new belief function expressing the combination of beliefs. To make a decision, i.e. to determine which proposition is the right one, different criteria have to be combined, potentially leading to a conflicting situation. The belief theory provides different operators to manage this conflict (Shafer 1976; Smets et al. 1994). It is yet important to note that Dempster’s rule does not apply if two sources are completely contradictory.

After the combination of sources a decision can be made. The belief theory also offers several decision rules developed in the literature (e.g. the maximum of plausibility, belief or the pignistic probability, see more details in (Smets et al. 1994)).

3.2 Description of the matching network process

In this part, let us consider two geographic datasets to match. They could be of any type, but for the sake of clarity suppose that they represent arcs of networks. For each feature belonging to one dataset *DataSet1*, matching generally first consists in looking for potential homologous features in the other dataset *DataSet2*. Then these candidates are analysed in order to determine actual matching links. Our matching process follows this approach and consists in four steps:

1. The first step consists in defining the frame of discernment. For each arc *arc1* in *DataSet1*, we look for close arcs $\{arc2_j\}_{j=1..m}$ in *DataSet2* (for more details concerning this rough selection of candidates see (Mustière and Devogele, 2008)). The frame of discernment is then the set of hypothesis A_j expressing that "*arc2*$_j$ *is homologous to arc1*". Due to the fact that a feature may have no homologous feature at all, another hypothesis, *NM*, standing for "*the feature arc1 is not matched at all*", is added.
   Therefore, the frame of discernment is now defined as follows:

$$\Theta_{arc1} = \{A_1, A_2, ...A_j, ...A_m, NM\} \ (3)$$

To compute the basic belief assignments, a local approach that analyses each candidate separately, is used according to a particular case of the belief theory, named the *specialised sources* (Appriou, 1991). Each source specialises on a hypothesis and assigns a mass of belief to it. In our case, a source coincides to a data matching criterion (like geometry, orientation, semantics, as explained later).

Let consider S_j, a subset of 2^{Θ}, defined as: $S_j = \{A_j, \neg A_{j,} \Theta\}$. Where:

- A_j is the hypothesis that *arc2*$_j$ is homologous to *arc1*.
- $\neg A_j = \{A_1 \ldots A_{j-1}, A_{j+1} \ldots A_m, NM\}$ is the hypothesis that *arc2*$_j$ is not homologous to *arc1*, i.e. the *arc1* is either matched to another candidate, or not matched at all.
- $\Theta = \{A_1, \ldots A_m, NM\}$ is the hypothesis expressing ignorance, i.e. the criterion does not know if *arc2*$_j$ is the right candidate or not.

2. The second step is an independent analysis of each proposition set S_j. It consists in computing the masses of belief for each criterion, and then combining criteria, i.e. for each set of propositions S_j the masses of belief are combined to provide a combined mass of belief synthesizing the knowledge from the different criteria. A similar approach was presented in (Royère 2002). Criteria are described in section 3.3.
3. In the third step, the results of the second step for each S_j are combined, in order to provide an overall view. Thus, the results for two candidates are combined, and these results are combined with the results for the third candidate and so on.
4. Finally, the fourth step is the actual decision among hypotheses.
   This is done using the maximum of "pignistic" probability. Within the context of the Transferable Belief Model, (Smets et al. 1994) defines and justifies the use of this decision rule. This rule makes the decision only among the simple hypotheses and thus there is no uncertainty at this level. However, let us precise that all propositions, which contain this hypothesis, are taken into account in the computation of pignistic probability, to choose the "best" simple hypothesis.
   Mention that the hypothesis with the highest mass of belief was chosen in our process. If this maximum is close to the second one, this raises some doubts about the validity of the results. Doubtful results can thus be highlighted and interactively checked.

3.3 Criteria of the matching network process

In order to illustrate the approach, we propose and describe here five criteria to match networks. Those criteria represent the sources in the frame of the belief theory.

The geometrical criterion

The geometrical criterion is based on the distance between arcs, which is defined as the Hausdorff semi-distance d_H (see (Mustière and Devogele 2008) for more details). Table 1 (on first column) represents the basic belief assignments for this criterion.

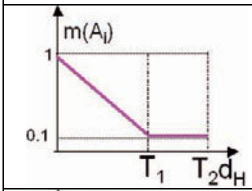
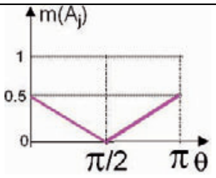
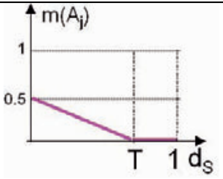
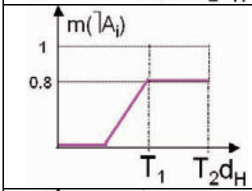
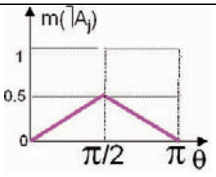
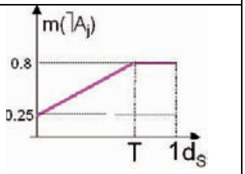
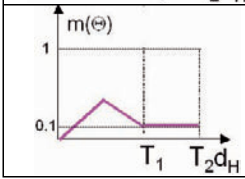
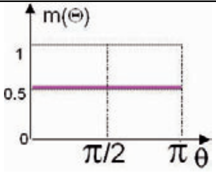
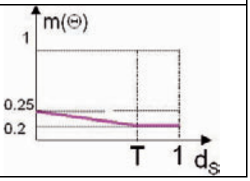
- The curve on the upper part of the figure represents the mass of belief for hypothesis A_j, "*arc2$_j$ is homologous to arc1*". It expresses that the closer the candidate *arc2$_j$* is to *arc1*, the more the criterion believes that these arcs are homologous. It also expresses that beyond a given threshold T_1, the criterion considers this hypothesis equally improbable, whatever the distance is. To avoid to definitively eliminate a candidate that is relatively far, the mass of belief allocated to this hypothesis is never 0, but ranges from 0 to 0.1.
- The curve in the middle represents the mass of belief for the proposition $\neg A_j$. A second threshold T_1 (fixed to $T_2/2$ in our tests) is defined in order to model imprecision of the spatial location. If the distance is less than T_1, the criterion considers this proposition as very improbable, but if the distance is between T_1 and T_2, the proposition begins to become more and more probable and, finally, when the distance is greater than the threshold T_2, the proposition becomes highly probable. The bigger is the distance, the higher the probability of the proposition is.
- Finally, the curve on the bottom represents the mass of belief for ignorance. For this criterion, ignorance is important when Euclidian distance is about T_1, i.e. when the candidate is neither close enough to conclude that it is it the right homologous arc, but nor far enough to conclude that is not it the right homologous arc.

The orientation criterion

Geometric properties of the features are also used to define the orientation criterion. This criterion consists in comparing local orientations between *arc1* and *arc2$_j$*. More specifically, it measures the differences between the orientations of tangents to *arc1* and to *arc2$_j$* respectively at the point of *arc1* nearest to *arc2*$_j$, and at the point of *arc2*$_j$ nearest to *arc1*. If the angle between two arcs is about 0 radians, arcs are relatively parallel and have the same orientation; if the value of the angle is closed to π, arcs are parallel but have opposite directions; if the value of angle approximates π/2, then arcs are perpendicular.

The orientation criterion is depicted on Table 1 (on second column). This criterion does not reject totally candidates that are not parallel, but assigns an important mass of belief to the ignorance, in this case. However, if arcs are relatively parallel, the criterion believes that the candidate arc could be the right homologous one, and it allocates an important mass of belief to the hypothesis "*arc2*$_j$ is homologous to *arc1*" but also to ignorance in order to express that this criterion is not sufficiently significant alone. Similarly, if arcs are perpendicular, the criterion assigns an important mass of belief to the hypothesis "*arc2*$_j$ is not homologous to *arc1*" and to ignorance.

Table 1. Basic belief assignments for geometrical (on first column), orientation (on second column) and semantic (on third column) criteria.

Geometrical criterion	Orientation criterion	Semantic criterion
		
		
		

The semantic criterion

Besides spatial knowledge, there are some properties such as functional road classification, name or number, which can be used in the matching process. In this paper, a criterion named the semantic criterion and based on nature of the features, is defined. To compare the different natures, a semantic distances d_S is required (Wu and Palmer 1994). This distance is computed using an ontology, which has been obtained by automatic extraction from textual specifications of the two datasets (Abadie et al. 2007). In Table 1 (on third column), the basic belief assumptions for the semantic criterion are presented. Thus if the semantic distance between *arc1* and the candidate $arc2_j$ is equal to 0, i.e. the arcs have the same nature, a mass of belief equal to 0.5 is assigned to the assumption "$arc2_j$ is homologous to *arc1*", so that this criterion supports this hypothesis but without allocated a strong belief to this candidate. If the semantic distance is greater than a threshold *T*, the mass of belief is shared between the hypotheses "$arc2_j$ is not homologous to *arc1*" and ignorance.

The name criterion

The name criterion is based on the comparison of names of arcs and relies on a distance, d_T, between strings using the Levenshtein distance (Levenshtein 1965).

Road name is a significant attribute that may bring useful knowledge in the matching process. However, it is difficult to use. On the one hand, the attribute value is not always fulfilled, in particular for minor roads. On the other hand, when it is fulfilled, differences between two datasets exist. For example, a road may be named according to a national nomenclature in one database and to a European nomenclature in the other one. Furthermore, names updates may have occurred and may not be reflected in both datasets.

For all these reasons, the name criterion is less reliable than other criteria. Correspondingly, where names are not available, ignorance has a significant weight and it should be explicitly modelled.

The first column of Table 2, shows the basic belief assignment for the name criterion, which is a discrete criterion, corresponding to four cases when two arcs *arc1* and $arc2_j$ are compared:

- case a): the attribute is not filled in both datasets. In this case, criterion can not make a decision, assigning a major mass of belief to ignorance. The complement of ignorance is divided into hypotheses "arc2j is homologous to arc1" and "arc2j is not homologous to arc1",
- case b): only one arc has a name value. When two arcs are compared and only one has a name, there are few chances that the two arcs are homologous, so the criterion believes that the arcs are not the right homologous assigning to this hypothesis a relevant mass of belief, with a low ignorance,

- case c): arcs have different names. When both arcs have distinct names, the criterion believes that the candidate is not the right homologous one, and assigns to this hypothesis a significant mass of belief, with ignorance equals to 0.3. To manage cases in which two arcs with different names representing the same reality, due to different nomenclature, the criterion does not completely reject the hypothesis "arc2j is homologous to arc1", and thus assigns it a low but positive mass of belief.
- case d): arcs have the same names. In this case, it is highly probable that the two arcs are homologous. Due to the fact that arcs are matched, and not roads, cases when arc1 has candidates with the same name frequently appear. Thus, the criterion assigns the same mass of belief equals to 0.5 to the hypothesis "arc2j is homologous to arc1" and to ignorance.

The neighbourhood criterion

In order to have a holistic analysis and to take into account the relations betweens geographic arcs, the neighbourhood criterion is defined. To do that, the process is actually slightly more complex than described in the previous section. In a first step, the matching process is performed using the four previous criteria as described in Section 3.2. Then, in a second step, the results of the first step are used to initialise the masses of belief of the neighbourhood criterion, and the final matching is performed with the five criteria as described in Section 3.2.

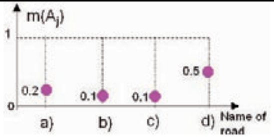
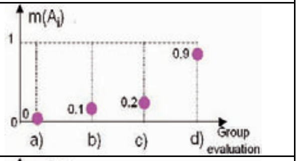
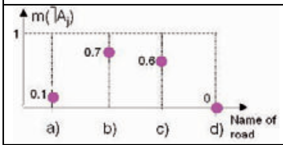
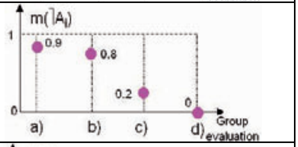
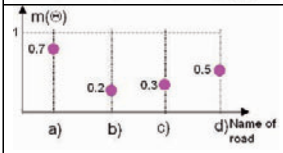
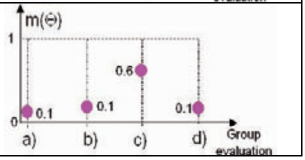
Let us define more in detail this assignment of belief. According to Section 3.2, 1:n correspondences are generated by the first step, i.e. one arc *arc2*$_j$ in the less detailed dataset is matched with n arcs in the most detailed dataset. Then, for each arc, *arc2*$_j$, belonging to the less detailed dataset, their n homologues arcs are grouped into connected groups. If only one group has been identified, the group is evaluated as being sure: case d) of the second column of Table 2. That is to say that all arcs belonging to the group are considered to be well-matched. Otherwise, if several groups are found, we analyse how neighbours of *arc2*$_j$ are matched in the first step, and especially if corresponding arcs of the neighbours are connected to the groups. Four cases are distinguished and depicted on Figure 1. Assignments of belief are shown on the second column of Table 2:

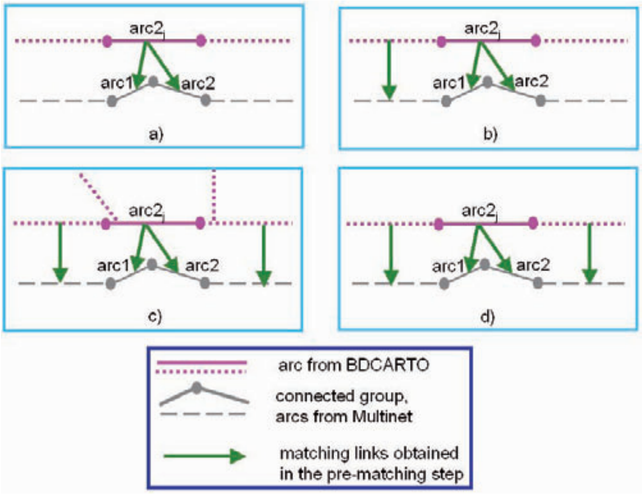
a. no connection is found. In this case, the criterion believes that "*arc2*$_j$ is not homologous to *arc1*",
b. only one connection is found. In this case, the criterion slightly believes that "*arc2*$_j$ is homologous to *arc1*",

c. some connections are found, but not for all neighbours. In this case, the criterion significantly believes that "*arc2*$_j$ is homologous to *arc1*",

d. connections are found for all neighbours. In this case, the criterion strongly believes that "*arc2*$_j$ is homologous to *arc1*".

Table 2. Basic belief assignments for name (on first column) and neighbourhood (on second column) criteria.

Name criterion	Neighbourhood criterion
	
	
	



arc2$_j$
arc1
arc2
a)
b)
c)
d)
arc from BDCARTO
connected group, arcs from Multinet
matching links obtained in the pre-matching step

Fig. 1. Different cases of matching results used to assign masses of belief for neighbourhood criterion

4. Experimentation

In this section tests data are presented and some matching results are illustrated and discussed.

4.1. Tests Data

Our experiments concern the matching of road and street networks taken from two databases: BDCARTO® and MultiNet®. These databases have different scales, producers and purposes. They are so highly heterogeneous. Our test area covers approximately 760 km^2.

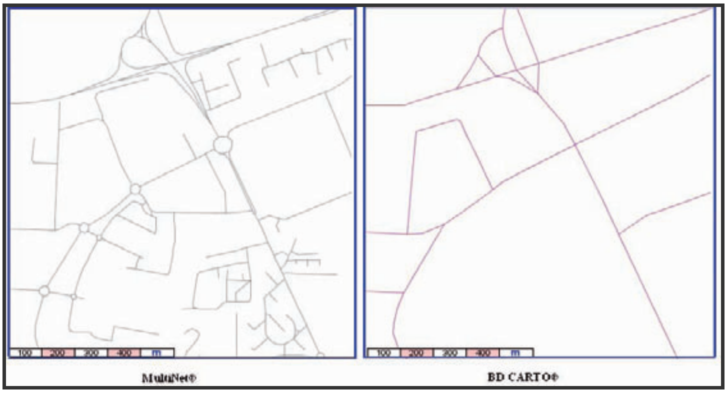


Fig. 2. Excerpt of road networks to match: MultiNet® (left) and BD CARTO® (right)

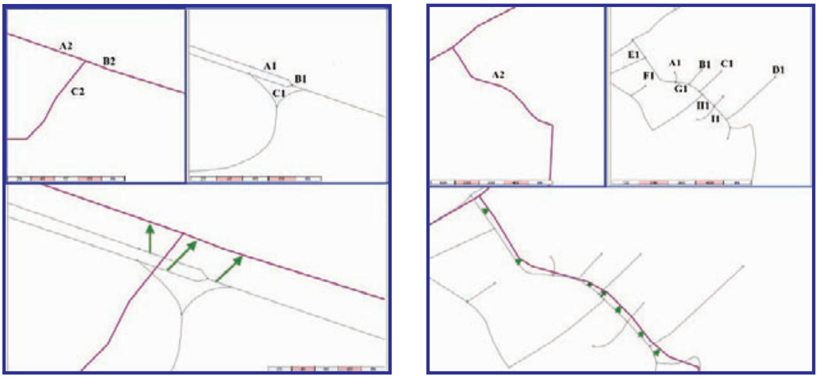
As shown on Figure 2, MultiNet is more detailed than BDCARTO. This is the most important difference between these datasets, but there are also features such as pathways and rugged ways that are represented in the less detailed database BDCARTO but not in MultiNet. Differences of modelling and representation also exist: each database has a specific representation of the real world according to its applications, perception and purposes. First, BDCARTO is build by IGN-France, a national mapping agency. It is used to make maps at 1:100,000 scale or 1:250,000 scale and to make geographical analyses at regional and departmental levels. It has an accuracy ranging from one to several decametres. Roads have attributes such as their classification, vocation, route number, name, physical state, etc. Next, MultiNet is built by TeleAtlas, a private company. Its accuracy goes from five to twelve metres in many detailed sections such as city street network. It contains thirteen thematic units of which road and street network. The letter is focusing especially on road and street network description for navigation applications including attributes describing the function, classification, name, route number, traffic of roads, but also attributes for geocoding.

To illustrate the differences between features representations in the two datasets, highways are modelled by a single line representing its axis in BDCARTO, and by two parallel lines representing the axis of each single carriageway in MultiNet. Another important difference between the datasets concerns the roundabouts that are represented by a set of arcs in MultiNet whereas there are generally represented by a single node in BDCARTO.

4.2 Matching results

In this section some results obtained by the matching process described in Section 3 are illustrated. As MultiNet (12,725 arcs) is the most detailed database, for each feature belonging to it we look for candidates in BDCARTO (2,063 arcs), before choosing among them the most probable one according to our approach based on the belief theory. In the next figures the less detailed dataset BDCARTO is displayed on the upper left side, the more detailed dataset MultiNet is displayed on the upper right side, and both datasets as well as links representing the matching results are displayed on the bottom.

As a global result, 60% of arcs belonging to MultiNet and 82% of arcs belonging to BDCARTO are matched by the process. On Figure 3, two examples of typical efficient matching results are depicted. Figure 3a) shows that our matching process does not necessary match arcs to the nearest neighbours in the other dataset. Arcs A1, B1 and C1 are matched to A2 and B2 respectively, and not to arc C2. This result is possible thanks to the orientation criterion that believes that C2 is neither homologous to A1, B1 nor C1. Figure 3b) shows unmatched arcs belonging to MultiNet (A1, B1, C1, D1) (hypothesis "Not Matched" was chosen) even if there are closed to some candidates. Moreover, on this example A2 is correctly matched to arcs E1, F1, G1, H1, I1.



a) b)

Fig. 3. Results of the data matching process

Figure 4 reveals the utility of using many criteria and in particular the neighbourhood criterion. Geographic context has a certain importance in the matching process. Thus, if the geographic context is not taken into account, over-matched matching results could occur, especially in urban areas. In some cases, due to the fact that the datasets have different scales, different features representing streets and roads are very close and have the same orientations. Therefore, the only criterion that could be decisive is the semantic criterion. Unfortunately it is usually not the case because our two datasets have heterogeneous administrative classifications in urban areas, thus leading to a too imprecise semantic criterion.

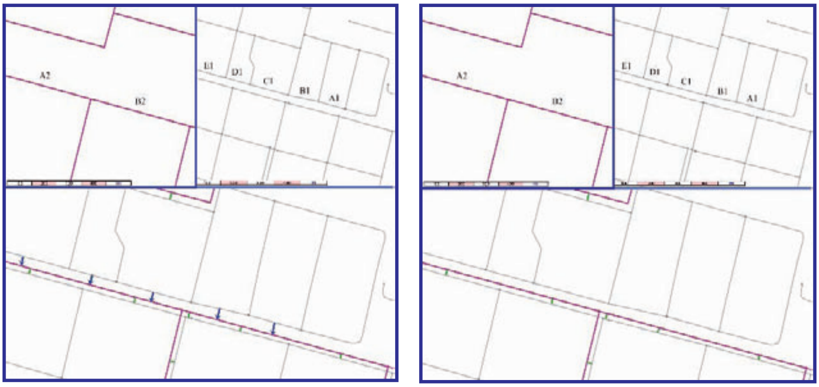


Fig. 4. Matching results, without the neighbourhood criterion on the left, and with it on the right

Consequently, the matching process without the neighbourhood criterion is not able to distinguish between right homologous objects and wrong ones: see Figure 4 on the left (arrows represent over-matched results, simple lines represent correct results).

When a more holistic analysis is made, by adding the neighbourhood criterion, over-matched results are eliminated: see Figure 4 on the right.

Similarly, Figure 5 illustrates the utility of the name criterion. On the left, the arc A2 representing the road “D81” is matched with arcs A1, B1, D1 and E1. Links (A2, D1) and (A2, E1) are wrong, because D1 and E1 are not arcs belonging to the road “D81”. This mistake is corrected when the name criterion is added to the matching: see Figure 5, on the right.

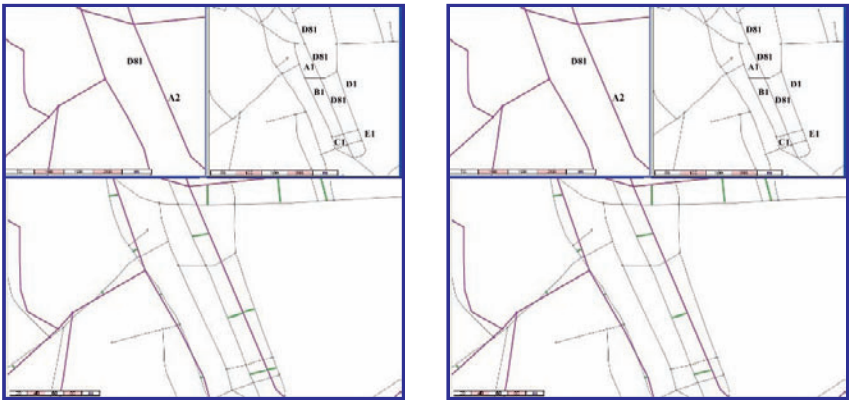


Fig. 5. Matching results, without the name criterion on the left, and with it on the right

5. DISCUSSION AND COMPARISON

Our results show that even if pieces of knowledge brought by criteria are imperfect and incomplete, the combination of them makes that, the process based on the belief theory generally succeeds to match homologous features.

Typically, although knowledge used to define the name criterion is incomplete, and filled only for 40% of the arcs, its addition to the matching process proved to be efficient. Such a criterion is less important that the others, but the belief theory offers tools to weaken a criterion and also to model the incompleteness by assigning an important mass of belief to ignorance. Thus, results not only are not deteriorated when adding imperfect knowledge, but they are also improved.

Our proposed approach has also the advantage over some existing approaches to manage cases where topological errors exist, because it combines geometric and attributes information. It also efficiently matches cases where a road is represented by two lines in a dataset and by one line in the other one.

However, some special cases are not well processed. For example, roundabouts are not well-managed. In BDCARTO, more roundabouts are represented by a single node, whereas in MultiNet they may be represented by a set of arcs. While our matching process does not match arcs to nodes, it is not efficient in these cases, and it matches arcs composing the roundabout with arcs in MultiNet that are the closest arcs and for which the differences between orientations are relatively limited. To improve this, we can imagine to previously detect roundabouts and then to add a new criterion. Doing this we would introduce some additional knowledge in the process.

Our approach is "one-way": each arc of one database (the most detailed) is matched to one arc in the other one (the less detailed). It may happen that several arcs of the detailed network are matched to the same arc of the other network: this is normal. But if network had similar levels of detail, this should not happen, or should at least be studied deeply. To improve this aspect a post-process could be imagined such as:

- analyse the pignistic probabilities for the two features that are matched with the same candidate, and then to select the one with the most important probability.
- introduce new criteria to carry out a more detailed analysis of features. For example, (Samal et al., 2004) proposes a matching approach that analyses the geographic context of features, i.e. the local systematic shifts between matched features. This idea could be introduced as a new criterion, exactly like we introduced the neighbourhood criterion.

Another important point to discuss on is the tuning of the process. As we can see in Table 1, curves defining the belief assignments are not similar: they vary from criterion to criterion. They also have different weights in the process. This is due to the fact that each criterion relies on different knowledge that is furthermore more or less perfect. First, this flexibility is a key advantage of the approach: it allows to precisely modelling different knowledge. Second, it may be thought of as a drawback, as tuning the process may become fastidious. However, we believe that this may not be such a big issue. In our experiments, we defined thresholds from the data specifications and especially from the known precision of the data. Moreover, we think that selection of precise thresholds is not so important, because we combine many criteria, and because the curves are relatively smooth, only approximate thresholds are necessary.

In our experiments, in order to define which criteria must be used and how to model their associated masses of belief, a detailed analysis of data has been made and expert knowledge has been used. Another raised question is thus how this step could become more generic and easier for end users. Two outlooks are possible. A first solution is to develop a method for optimizing parameters that is a compromise between matching results quality and number. A second solution is to determine thresholds and weights by data mining. This open issue still needs to be studied.

6. Conclusion

In this paper a matching approach bearing on the belief theory is presented. It combines criteria that are based on imperfect knowledge such as the position, nature or orientation of geographic data. This approach is tested on heterogeneous geographic networks with different scales. Despite that both data and knowledge are imperfect, our matching process matched correctly most parts of networks thanks to the combination of criteria.

The proposed process can still be improved, and other matching criteria defined. However, we believe that the explicit representation of imperfection, like imprecision, uncertainty and incompleteness or ignorance is very promising for studying geographic data.

Acknowledgements

The authors would like to thank the CERTU and Magali Di Salvo for its collaboration on this research project through giving us temporally access to the Multinet® data. The authors would also like to thank the anonymous reviewers for their very fruitful remarks.

References

Abadie N, Olteanu A-M, Mustière S (2007) Comparaison de la nature d'objets géographiques. In : Colloque Ingénierie des Connaissances, 3 juillet, 2007, Grenoble

Appriou A (1991) Probabilités et incertitudes en fusion de données multi-senseurs Revue Scientifique et Technique de la Défense 11, pp 27-40

Beeri C, Kanza Y, Safra E, Sagiv Y (2004) Object fusion in Geographic Information System. Proceedings 30th VLDB conference, 2004, Toronto, Canada, pp 816-827

Dempster A (1968) Upper and lower probabilities induced by multivalued mapping. Annals of Mathematical Statistics, (AMS-38), pp 325-339

Devogele T (1997) Processus d'intégration et d'appariement des bases de données géographiques-Application à une base de données routière multi-échelles, PhD thesis, Université de Versailles, France

Devogele T, Parent C, Spaccapietra S (2001) On spatial database integration. International Journal of Geographical Information Science, 12(4), pp 335–352

Hampe M, Sester M (2002) Real-time integration and generalization of spatial data for mobile applications. Geowissenschaftliche Mitteilungen, Maps and the Internet, Wien, Heft (60), pp 167-175.

Kilpelaïnen T (2000) Maintenance of Multiple Representation Databases of Topographic Data. The Cartographic Journal, 37 (2), pp 101-107

Levenshtein VI (1965) Binary Codes Capable of Correcting Deletions, Insertions, and Reversals, Soviet Physics - Doklady, 10(8), 707-710. Translated from Doklady Akademii Nauk SSSR, 163(4), pp845-848, 1965

Mustière S (2006) Results of experiments on automated matching of networks. Proceedings of the ISPRS Workshop on Multiple Representation and Interoperability of Spatial Data, 22-24 February 2006, Hanover, Germany, pp 92-100

Mustière S, Devogele T (2008) Matching networks with different levels of detail. GeoInformatica, on press, to be published in 2008

Mustière S, van Smaalen J (2007) Database Requirements for Generalisation and Multiple Representations. In: Mackaness W, Ruas A, Sarjakoski T (eds), The Generalisation of Geographic Information: Models and Applications, Elsevier, pp 113-136.

Olteanu A-M (2007) Matching geographical data using the Theory of Evidence. Proceedings of 20th ICC, 5-9 August 2007, Moscow, Russie

Rajabifard A, Binns A, Masser I, Williamson I (2006) The role of sub-national government and the private sector in future spatial data infrastructures. International Journal of Geographical Information Science, 20(7), pp 727-741

Rodriguez MA, Egenhofer MJ (2003) Determining semantic similarity among entity classes from different ontologies. IEEE Transactions on Knowledge and Data Engineering, 15(2), pp 442- 456

Royère C, Gruyer D, Cherfaoui V (2002) Data association with believe theory. Proceedings of International Conference of Information Fusion, Washington, pp 23-29

Samal A, Seth SC, Cueto K (2004) A feature-based approach to conflation of geospatial sources. International Journal of Geographical Information Science, 18(5), pp 459-489.

Shafer G (1976) A Mathematical Theory of Evidence. Princeton University Press.

Smets P, Kennes R (1994) The Transferable Belief Model. Artificial Intelligence, 66, pp 191-234.

Uitermark H (2001) Ontology-Based Geographic Data Set Integration. PhD thesis, Universiteit Twente, the Netherlands, 2001.

Wu Z, Palmer M (1994) Verb Semantics and Lexical Selection. Proceedings of the 32nd Annual Meetings of the Associations for Computational Linguistics, pp 133-138

Volz S (2006) An iterative approach for matching multiple representations of street data. Proceedings of the ISPRS workshop on Multiple Representation and Interoperability of Spatial Data, 22-24 February 2006, Hanover, Germany, pp 101-110

Walter V, Fritsch D (1999) Matching Spatial Data Sets: a Statistical Approach. International Journal of Geographical Information Science, 13(5), pp445-473

Zhang M, Shi W, Meng L (2005) A generic matching algorithm for line networks of different resolutions. Proceedings 8th ICA workshop on Generalisation and Multiple Representation, July, 2005, Coruña, Spain

Deriving Topological Relationships Between Simple Regions with Holes

Mark McKenney, Reasey Praing, and Markus Schneider[⋆]

Department of Computer and Information Science & Engineering, University of Florida
Gainesville, Florida, USA
{mm7,rpraing,mschneid}@cise.ufl.edu

Abstract

Topological relationships between objects in space are of great importance in many disciplines. Due to the lack of local topological information between components, i.e. faces, in the model of topological relationships between complex regions, recently, *localized topological relationships* have been defined for complex regions based on the relationships between simple regions with holes. However, unlike for simple regions, topological relationships between simple region with holes are not widely implemented. Therefore, in this paper, we propose an approach to derive topological relationships between simple regions with holes based on well known topological relationships between the simple regions as their components. This will allow localized topological predicates between complex regions to be implemented using only topological relationships between simple regions. Furthermore, localized topological predicates between complex regions can be used to implement topological relationships between complex regions. Therefore, this work allows topological relationships between complex regions to be implemented using only topological relationships between simple regions.

1 Introduction

The exploration of relationships between spatial objects is an important topic in fields such as artificial intelligence, robotics, VLSI design, linguistics, CAD, and GIS. Object relationships can be used not only to gain information about the objects involved but also for inferring new, non-explicit information as well as creating fast access and indexing structures in spatial databases. Specifically, topological relationships have been the focus of extensive research for a long time. This research includes the design of models of topological relationships between all types of spatial objects as

[⋆] This work was partially supported by the National Science Foundation (NSF) under grant number NSF-CAREER-IIS-0347574.

well as related topics like the exploration of topological relationships as a reasoning tool.

Models for topological relationships have predominantly considered *simple* spatial data types. A simple point object is defined as a single pair of coordinates, a simple line object is given as a non self-intersecting connected curve, and a simple region object is represented as an areal object topologically equivalent to a closed disc. A well-known model that defines the topological relationships between simple spatial objects is the 9-intersection model (9IM). The commonly known set of eight topological relationships originally defined by the 9IM between simple regions includes the relationships *overlap, meet, inside, contains, coveredBy, covers, equal,* and *disjoint*. This model was then extended to support topological relationships between *complex* spatial types. Roughly, a complex point is defined by a set of disjoint simple points. A complex line is composed of a set of blocks of connected simple lines. A complex region is defined as a set of one or more faces, each possibly containing holes.

The application of the 9IM to complex spatial data types has raised awareness of the global nature of the 9IM. That is, the 9IM considers the interior, exterior, and boundary point sets of the whole objects, and ignores the fact that complex spatial objects are composed of individual and separate components. As a result, local topological information regarding the relationship between individual components from each object is lost. Recently, this problem has been addressed through the introduction of *localized topological relationship* models, which are able to represent the topological relationships between components of complex regions, and *hybrid topological relationship* models, which can represent both the global and local topological relationships between complex regions (McKenney et al., 2007). However, these models are based upon topological relationships between *simple regions with holes*, a data type which is not typically implemented in spatial database systems. Furthermore, the complete set of topological relationships between complex regions is currently not implemented in any commercial spatial database system. Thus, the integration of local and hybrid topological relationship models, as well as the global topological relationships cannot be fully utilized in spatial systems at this time.

Although the complete set of topological relationships between complex regions, local topological predicates, and hybrid topological predicates are not fully implemented in commercial systems, implementations of the well known eight topological relationships between simple regions are commonly available. Therefore, the overall goal of this paper is to develop a method by which the topological predicates between complex regions can be characterized using only the eight topological predicates between simple regions. It has been shown that topological relationships between complex regions can be defined based on topological relationships between simple regions with holes using the localized topological relationship model. In this paper, we achieve this goal by developing a method by which topological relationships between simple regions with holes can be defined based on topological predicates between simple regions. Such a method will allow localized topological relationships between complex regions to be directly implemented on top of any system that provides the topological predicates between simple regions.

The remainder of this paper is structured as follows: Section 2 introduces related work. Section 3 demonstrates our concept of identifying topological relationships between simple regions with holes using topological relationships between simple regions. Finally, Section 4 gives conclusions and discusses future work.

2 Related Work

In this section, we consider previous works on spatial data models including the definition of different types of regions as well as the 9-intersection model which characterizes the topological relationships between them.

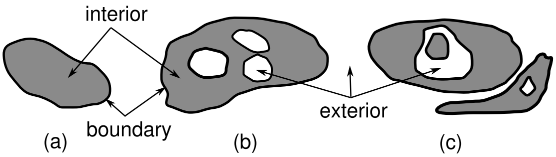


Fig. 1. A simple region (a), a simple region with holes (b), and a complex region (c).

As defined in Egenhofer and Franzosa (1991); Schneider and Behr (2006), a region consists of the interior, the boundary and the exterior point sets. Based on this definition, Figure 1 illustrates the differences between these point sets. A simple region (e.g. Egenhofer et al., 1989) is an areal object topologically equivalent to a closed disk (Figure 1a). A simple region with holes (e.g. Egenhofer et al., 1994) is made up of an outer polygon denoting its outer boundary and zero or more hole polygons representing its holes (Figure 1b). All holes must be completely contained within the outer polygon and can share a finite number of boundary points with the outer cycle and with other holes. A complex region (e.g. Schneider and Behr, 2006) is composed of faces where each face is a simple region with holes (Figure 1c).

Topological relationships between spatial objects can be defined by the 9-intersection model (e.g. Egenhofer and Franzosa, 1991; Egenhofer and Herring, 1990) by evaluating the non-emptiness of the intersection between all combinations of the interior ($^\circ$), boundary (∂) and exterior ($^-$) of the objects involved. A 3×3 matrix with Boolean value elements, as illustrated in Figure 2, describes the topological relationship between each pair of spatial objects. Table 1 shows the 8 topological relationships between simple regions.

Originally defined for simple regions, the 9IM has been extended to handle simple regions with holes (e.g. Egenhofer et al., 1994), and complex spatial objects

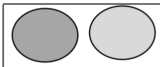
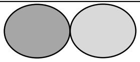
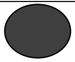
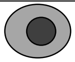
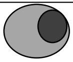
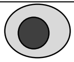
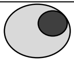
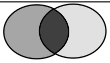
 $\begin{pmatrix}0&0&1\\0&0&1\\1&1&1\end{pmatrix}$ disjoint	 $\begin{pmatrix}0&0&1\\0&1&1\\1&1&1\end{pmatrix}$ meet	 $\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}$ equal	 $\begin{pmatrix}1&0&0\\1&0&0\\1&1&1\end{pmatrix}$ inside
 $\begin{pmatrix}1&0&0\\1&1&0\\1&1&1\end{pmatrix}$ coveredBy	 $\begin{pmatrix}1&1&1\\0&0&1\\0&0&1\end{pmatrix}$ contains	 $\begin{pmatrix}1&1&1\\0&1&1\\0&0&1\end{pmatrix}$ covers	 $\begin{pmatrix}1&1&1\\1&1&1\\1&1&1\end{pmatrix}$ overlap

Table 1. The 8 topological relationships between simple regions.

(e.g. Clementini and Di Felice, 1996; Schneider and Behr, 2006). The model in Egenhofer et al. (1994) characterizes the topological relationships between two simple regions with holes as the conjunction of topological relationships between their underlying simple regions (each of the outer cycles and the holes is considered a simple region). For two simple regions with holes A and B with n and m holes respectively, a matrix of $(n+1)(m+1)$ elements represents the topological relationship between A and B. This means that under this model, the number of topological relationships between two simple regions with holes is dependent on the number of holes in each region, resulting in an arbitrary number of relationships. A similar approach between composite regions can be found in Clementini et al. (1995). To avoid having an infinite set of valid topological relationships, the finite set of topological relationships between simple regions with holes based on the 9IM that is independent of the number of holes in each object has been identified in

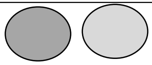
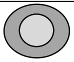
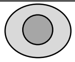
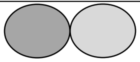
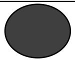
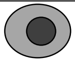
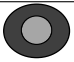
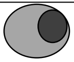
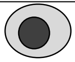
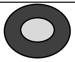
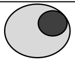
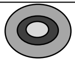
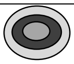
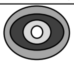
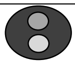
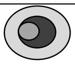
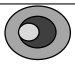
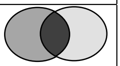
 $\begin{pmatrix}0&0&1\\0&0&1\\1&1&1\end{pmatrix}$ disjoint	 $\begin{pmatrix}0&0&1\\0&1&0\\1&1&1\end{pmatrix}$ fillingHole	 $\begin{pmatrix}0&0&1\\0&1&1\\1&0&1\end{pmatrix}$ holeFilled	 $\begin{pmatrix}0&0&1\\0&1&1\\1&1&1\end{pmatrix}$ meet	 $\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}$ equal	 $\begin{pmatrix}1&0&0\\1&0&0\\1&1&1\end{pmatrix}$ inside
 $\begin{pmatrix}1&0&0\\1&1&0\\1&0&1\end{pmatrix}$ holeCoveredBy	 $\begin{pmatrix}1&0&0\\1&1&0\\1&1&1\end{pmatrix}$ coveredBy	 $\begin{pmatrix}1&1&1\\0&0&1\\0&0&1\end{pmatrix}$ contains	 $\begin{pmatrix}1&1&1\\0&1&0\\0&0&1\end{pmatrix}$ holeCovers	 $\begin{pmatrix}1&1&1\\0&1&1\\0&0&1\end{pmatrix}$ covers	 $\begin{pmatrix}1&1&1\\1&0&0\\1&1&1\end{pmatrix}$ insideOverfill
 $\begin{pmatrix}1&1&1\\1&0&1\\1&0&1\end{pmatrix}$ containsOverfill	 $\begin{pmatrix}1&1&1\\1&0&1\\1&1&1\end{pmatrix}$ insideContains	 $\begin{pmatrix}1&1&1\\1&1&0\\1&0&1\end{pmatrix}$ coversCoveredBy	 $\begin{pmatrix}1&1&1\\1&1&0\\1&1&1\end{pmatrix}$ coversOverfill	 $\begin{pmatrix}1&1&1\\1&1&1\\1&0&1\end{pmatrix}$ coveredByOverfill	 $\begin{pmatrix}1&1&1\\1&1&1\\1&1&1\end{pmatrix}$ overlap

Table 2. The 18 topological predicates between simple regions with holes.

McKenney et al. (2007). This set consists of 18 topological relationships as shown in Table 2. These relationships are used to define a model that preserves *local topological relationships* between complex regions while still maintaining global information. However, the approach assumes that an implementation of topological relationships between simple regions with holes exists. It turns out that this is not the case in many of today's spatial database management systems. Therefore, in this paper, we introduce a method for deriving the relationships between simple regions with holes from using only the relationships between their underlying simple regions.

$$\begin{pmatrix} A^{\circ} \cap B^{\circ} \neq \varnothing & A^{\circ} \cap \partial B \neq \varnothing & A^{\circ} \cap B^{-} \neq \varnothing \\ \partial A \cap B^{\circ} \neq \varnothing & \partial A \cap \partial B \neq \varnothing & \partial A \cap B^{-} \neq \varnothing \\ A^{-} \cap B^{\circ} \neq \varnothing & A^{-} \cap \partial B \neq \varnothing & A^{-} \cap B^{-} \neq \varnothing \end{pmatrix}$$

Fig. 2. The 9-intersection matrix for topological relationships.

3 Constructing topological relationships between simple regions with holes

As stated in Section 2, a simple region with holes is constructed of an outer cycle and finitely many hole cycles. Each of these cycles defines the boundary of a simple region. Our approach to deriving topological relationships between simple regions with holes A and B is to discover the individual entries of the 9IM depicting the topological relationship between A and B by examining the interactions of the individual cycles of A with the cycles of B. We then show how such interactions can be discovered using topological predicates between simple regions. This allows us to characterize a topological relationship between simple regions with holes (Table 2) using only topological predicates between simple regions (Table 1).

As a matter of notation, we indicate the 9IM representing the topological relationship between two simple regions with holes A and B as $M_{(A,B)}$. Furthermore, we indicate the matrix entry corresponding to whether or not the interiors of the objects intersect as $M_{(A,B)}{}^{\circ\circ}$, the entry corresponding to whether or not the interior of A and the boundary of B intersect as $M_{(A,B)}{}^{\circ}\partial$, etc. We use the notations $\omega(A)$ and $\iota(A)$ to indicate the set of simple regions formed by the outer cycle and hole cycles of a simple region with holes A, respectively. Note that $\omega(A)$ is a singleton set since a simple region with holes has a single outer cycle. The *closure* of a simple region with holes A is denoted as $\overline{A}$ and is defined as the union of the boundary and interior of A.

Given two simple regions with holes A, and B, we must now show how to determine the values for the entries in $M_{(A,B)}$. Here, we only need to characterize the value for entries $M_{(A,B)}{}^{\circ\circ}$, $M_{(A,B)}{}^{\circ}\partial$, $M_{(A,B)}{}^{\circ-}$, $M_{(A,B)}\partial\partial$, and $M_{(A,B)}\partial^{-}$. This is because $M_{(A,B)}{}^{--}$ is always true and the characterization of entries $M_{(A,B)}\partial^{\circ}$, $M_{(A,B)}{}^{-\circ}$, and $M_{(A,B)}{}^{-}\partial$ is the same as that of entries $M_{(A,B)}{}^{\circ}\partial$, $M_{(A,B)}{}^{\circ-}$, and $M_{(A,B)}\partial^{-}$ respectively with A and B swapped. In order to determine the entries of $M_{(A,B)}$, we make several observations about the interactions of the components of two simple regions with holes A and B:

Observation 1 The interiors of A and B intersect if the interiors of their outer cycles intersect, unless one of the outer cycles is completely contained in a hole of the opposing region. As an example, consider Figure 3. This illustrates the case when the interiors of two simple regions with holes do not intersect, even though the interiors of the simple regions defined by their outer cycles do intersect.

Observation 2 The interior of A intersects the boundary of B if the simple region representing any hole or outer cycle from B intersects the interior of the outer cycle of A and that region is not completely contained in a hole of A.

Observation 3 The interior of A intersects the exterior of B if the interior of the outer cycle of A intersects the exterior of the outer cycle of B or the interior of the simple region representing any hole cycle of B if this hole of B is not completely contained in a hole of A (Note that the interior of the simple region representing a hole cycle in a simple region with holes represents the exterior of the simple region with holes).

Observation 4 The boundaries of two simple regions with holes intersect if the boundaries of any of their cycles intersect.

Observation 5 The boundary of A intersects the exterior of B if the boundary of the outer cycle of A intersects the exterior of the outer cycle of B, or if the boundary of any cycle from A intersects the interior of a simple region representing a hole cycle of B.

We state Observations 1 - 5 formally in Theorem 1.

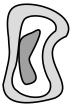


Fig. 3. Two simple regions with holes.

Theorem 1. *Let A and B be two simple regions with holes. The values of entries in the 9IM describing their topological relationship can be written as:*

$$
\begin{aligned}
&(i) && M_{(A,B)}{}^{\circ\circ} = (\exists c \in \omega(A), \exists d \in \omega(B) | c^\circ \cap d^\circ \neq \varnothing)\\
& && \quad \wedge\ \neg(\exists c \in \omega(A), \exists h \in \iota(B) | \overline{c} \subseteq \overline{h})\\
& && \quad \wedge\ \neg(\exists j \in \iota(A), \exists d \in \omega(B) | \overline{d} \subseteq \overline{j})\\
&(ii) && M_{(A,B)}{}^{\circ}\partial = (\exists c \in \omega(A), \exists d \in \omega(B) \cup \iota(B) | c^\circ \cap \partial d \neq \varnothing)\\
& && \quad \wedge\ \neg(\exists h \in \iota(A), \exists d \in \omega(B) | \overline{d} \subseteq h^\circ)\\
&(iii) && M_{(A,B)}{}^{\circ -} = (\exists c \in \omega(A), \exists d \in \omega(B) | c^\circ \cap d^- \neq \varnothing)\\
& && \quad \vee\ (\exists c \in \omega(A), \exists j \in \iota(B), \forall h \in \iota(A) | c^\circ \cap j^\circ \neq \varnothing\ \wedge\ \overline{j} \nsubseteq \overline{h})\\
&(iv) && M_{(A,B)}\partial\partial = \exists c \in \omega(A) \cup \iota(A), \exists d \in \omega(B) \cup \iota(B) | \partial c \cap \partial d \neq \varnothing\\
&(v) && M_{(A,B)}\partial^- = (\exists c \in \omega(A) \cup \iota(A), \exists d \in \omega(B) | \partial c \cap d^- \neq \varnothing)\\
& && \quad \vee\ (\exists c \in \omega(A) \cup \iota(A), \exists h \in \iota(B) | \partial c \cap h^\circ \neq \varnothing)
\end{aligned}
$$

Proof.

(i) The interiors of two simple regions with holes can only intersect if the simple regions represented by their outer cycles intersect. However, because holes are allowed in simple regions with holes and the interior of a simple region representing a hole is part of the exterior of the simple region with holes, then the interiors of two simple regions with holes cannot intersect if one outer cycle is completely contained in a hole of the opposing region. This follows from the definition of simple regions with holes stating that a simple region with holes contains only a single outer cycle. Therefore, the interior of a hole is entirely part of the exterior of the simple region with holes.

(ii) Note that the boundary of a simple region with holes consists of the boundaries of both the outer cycle and the hole cycles. This is in contrast to the interior of a simple region with holes which consists of the difference of the interior of the simple region representing the outer cycle, and the interiors of all simple regions representing hole cycles. Therefore, if the boundary of a simple region representing either an outer or hole cycle from *B* intersects the interior of the outer cycle of *A*, then the boundary of *B* intersects the interior of *A* unless one special case occurs. This special case is when *B* is a subset of the interior of a hole in *A*. In other words, if the outer cycle of *B* is completely contained in the interior of a simple region representing a hole in *A*, then the boundary of *B* cannot intersect *A*. Note that it is impossible for the intersection of the boundary of *B* and the interior of *A* to be completely contained in more than one hole of *A* since this would either require two holes to meet along a boundary, or for the holes to connect as to form a second face of *A*, both of which are prohibited by the definition of simple regions with holes.

(iii) If the interior of the simple region defined by the outer cycle of *A* intersects the exterior of the simple region defined by the outer cycle of *B*, then the boundary of *A* intersects the exterior of *B*. This follows from the definition of simple regions. Furthermore, the interior of *A* intersects the exterior of *B*. This follows from the fact that if the boundary of a simple region with holes intersects the exterior of a second simple region with holes, the interior of the first simple region with holes must also intersect the exterior of the second region with holes (McKenney et al., 2007). Alternatively, the interior of the simple region defined by the outer cycle of *A* can intersect the interior of a hole of *B*. However, if that hole is completely contained in a hole of *A*, then the interior of *A* will not intersect the exterior of *B*.

(*iv*) The boundary of a simple region with holes consists of the boundaries of all the outer cycles and the hole cycles of the region. Therefore, if the boundary of any cycle of a simple region with holes intersects the boundary of any cycle of a second simple region with holes, then the boundaries of the two regions intersect. There are no special cases for boundaries of hole cycles or outer cycles.

(*v*) The boundary of *A* intersects the exterior of *B* if the boundary of any cycle in *A* intersects the exterior of the simple region defined by the outer cycle of *B*, or the interior of the simple region defined by any hole cycle of *B*. This follows from the fact that the exterior of a simple region with holes consists of the exterior of the simple region defined by its outer cycle and the interiors of all simple regions defined by its hole cycles, and the fact that the boundary of a simple region with holes consists of the boundaries of all cycles in the region. □

At this point, we can determine the 9IM, and hence the topological relationship, between two simple regions with holes by examining the interactions of the simple regions defined by their outer and hole cycles. However, the goal of this paper is to characterize these topological relationships by using topological predicates between simple regions. We now define a method to do this. By examining Theorem 1, it is clear that we need to know all interactions between the simple regions defined by all cycles of two simple regions with holes *A* and *B* in order to determine their 9IM. Therefore, we must represent the topological predicates that hold between the simple regions that define the outer cycles of *A* and *B*, hole cycles of *A* and *B*, the outer cycle of *A* and all hole cycles of *B*, and the outer cycle of *B* and all hole cycles *A*. If this information is known, then (based on the 9IMs between simple regions) we can directly determine the values for parts (*i*), (*ii*), (*iv*), and (*v*) from Theorem 1. This follows from the fact that each of these parts is defined based on the existence or non-existence of certain predicates holding between pairs of components from two simple regions with holes. However, knowing which predicates hold between the above combinations is insufficient to determine the value for part (*iii*) of Theorem 1. This is because the second part of the conjunction requires knowledge of the existence of a certain topological configuration between three components of simple regions with holes (namely that there exists a hole in one simple region with holes that is contained in the interior of the outer cycle of a second simple region with holes, and is not contained in a hole of the second simple region with holes). Therefore, in order to determine the topological relationship between two simple regions with holes *A* and *B* based on topological predicates between simple regions, we must discover all the topological predicates that hold between the simple regions representing the outer cycles from *A* and *B*, the outer cycle from *A* and all hole cycles from *B*, the outer cycle from *B* and all hole cycles from *A*, and all hole cycles from *A* and all hole cycles from *B*. Furthermore, we must indicate whether the special situation in part (*iii*) of Theorem 1 holds. Thus, we characterize a topological relationship between two simple regions with holes as a *component based topological relationship* (CBTR), which consists of four sets of topological predicates, containing all topological predicates that exist between the outer cycle of one region and the outer cycle of the other, the outer cycle of one region and all hole cycles of the other, etc, and two boolean values indicating whether the special situation for part (iii) of Theorem 1 holds between each A and B and between B and A. We formally define the CBTR between two simple regions with holes as:

Definition 1. *Let A and B be two simple regions with holes and P_{SR} be the set of topological predicates between simple regions. The CBTR that describes their topological relationship is a six-tuple $CBTR = (OO, OH, HO, HH, B_{HHO}, B_{OHH})$ defined as:*

$$
\begin{aligned}
&OO = \{p \in P_{SR} | \exists c \in \omega(A), \exists d \in \omega(B) : p(c,d)\}\\
&OH = \{p \in P_{SR} | \exists c \in \omega(A), \exists j \in \iota(B) : p(c,j)\}\\
&HO = \{p \in P_{SR} | \exists h \in \iota(A), \exists d \in \omega(B) : p(h,d)\}\\
&HH = \{p \in P_{SR} | \exists h \in \iota(A), \exists j \in \iota(B) : p(h,j)\}\\
&B_{OHH} = (\exists j \in \iota(B), \exists h \in \iota(A), \exists c \in \omega(A) : (equal(j,c)\\
&\qquad \vee\ coveredBy(j,c)\ \vee\ inside(j,c))\\
&\qquad \wedge\ \neg(equal(j,h)\ \vee\ coveredBy(j,h), inside(j,h)))\\
&B_{HHO} = (\exists h \in \iota(A), \exists j \in \iota(B), \exists d \in \omega(B) : (equal(h,d)\\
&\qquad \vee\ coveredBy(h,d)\ \vee\ inside(h,d))\\
&\qquad \wedge\ \neg(equal(h,j)\ \vee\ coveredBy(h,j), inside(h,j)))
\end{aligned}
$$

We are now able to compute a CBTR between any two simple regions with holes by using only topological relationships between simple regions. Furthermore, because the topological relationships between simple regions are known, it is clear that given a CBTR between two simple regions with holes, we can derive the 9IM between simple regions with holes from it based on Theorem 1. To do this, we need to convert our characterizations in Theorem 1 such that each of the entries is defined based on topological relationships between simple regions. Using the 9IM of the 8 topological relationships between simple regions (Table 1) and Theorem 1, we obtain the following characterization of the 9IM entries for simple region with holes.

Corollary 1. *Let A and B be two simple regions with holes. Let p be a topological relationship between simple regions. We denote $p_{OO}(A,B)$, $p_{OH}(A,B)$, $p_{HO}(A,B)$, and $p_{HH}(A,B)$ as the topological relationship between the outer cycles of A and B, the outer cycle of A and a hole of B, a hole of A and the outer cycle of B, and a hole of A and a hole of B respectively. The values of the entries in the 9IM describing the topological relationship between A and B can now be written as:*

$$
\begin{aligned}
(i)\ \ & M_{(A,B)}{}^{\circ\circ} = (equal_{OO}(A,B) \lor inside_{OO}(A,B) \lor coveredBy_{OO}(A,B)\\
& \quad \lor contains_{OO}(A,B) \lor covers_{OO}(A,B) \lor overlap_{OO}(A,B))\\
& \quad \land \neg equal_{OH}(A,B) \land \neg inside_{OH}(A,B) \land \neg coveredBy_{OH}(A,B)\\
& \quad \land \neg equal_{HO}(A,B) \land \neg contains_{HO}(A,B) \land \neg covers_{HO}(A,B)\\
(ii)\ \ & M_{(A,B)}{}^{\circ}\partial = ((contains_{OO}(A,B) \lor covers_{OO}(A,B) \lor overlap_{OO}(A,B))\\
& \quad \land \neg equal_{HO}(A,B) \land \neg contains_{HO}(A,B) \land \neg covers_{HO}(A,B))\\
& \quad \lor ((contains_{OH}(A,B) \lor covers_{OH}(A,B) \lor overlap_{OH}(A,B))\\
& \quad \land \neg equal_{HH}(A,B) \land \neg contains_{HH}(A,B) \land \neg covers_{HH}(A,B))\\
(iii)\ & M_{(A,B)}{}^{\circ-} = disjoint_{OO}(A,B) \lor meet_{OO}(A,B) \lor contains_{OO}(A,B)\\
& \quad \lor covers_{OO}(A,B) \lor overlap_{OO}(A,B) \lor B_{OHH}\\
(iv)\ & M_{(A,B)}\partial\partial = meet_{OO}(A,B) \lor equal_{OO}(A,B) \lor coveredBy_{OO}(A,B)\\
& \quad \lor covers_{OO}(A,B) \lor overlap_{OO}(A,B) \lor meet_{OH}(A,B)\\
& \quad \lor equal_{OH}(A,B) \lor coveredBy_{OH}(A,B) \lor covers_{OH}(A,B)\\
& \quad \lor overlap_{OH}(A,B) \lor meet_{HO}(A,B) \lor equal_{HO}(A,B)\\
& \quad \lor coveredBy_{HO}(A,B) \lor covers_{HO}(A,B) \lor overlap_{HO}(A,B)\\
& \quad \lor meet_{HH}(A,B) \lor equal_{HH}(A,B) \lor coveredBy_{HH}(A,B)\\
& \quad \lor covers_{HH}(A,B) \lor overlap_{HH}(A,B)\\
(v)\ \ & M_{(A,B)}\partial^{-} = disjoint_{OO}(A,B) \lor meet_{OO}(A,B) \lor contains_{OO}(A,B)\\
& \quad \lor covers_{OO}(A,B) \lor overlap_{OO}(A,B) \lor disjoint_{HO}(A,B)\\
& \quad \lor meet_{HO}(A,B) \lor contains_{HO}(A,B) \lor covers_{HO}(A,B)\\
& \quad \lor overlap_{HO}(A,B) \lor inside_{OH}(A,B) \lor coveredBy_{OH}(A,B)\\
& \quad \lor overlap_{OH}(A,B) \lor inside_{HH}(A,B) \lor coveredBy_{HH}(A,B)\\
& \quad \lor overlap_{HH}(A,B)
\end{aligned}
$$

Given any valid configuration of two simple regions with holes A and B, we can compute their CBTR and determine the 9IM representing the relationship between them using Corollary 1. For example, consider two simple regions with holes A and B as shown in Figure 4. B has a single hole covered by A which has no hole. The CBTR for this scene can be computed by using any available implementation of topological relationships between simple regions. We obtain $CBTR(A,B) = (\{inside\},\{covers\},\varnothing,\varnothing,\ true,\ false)$. By applying this CBTR to Corollary 1, we obtain $M_{(A,B)}{}^{\circ\circ} = true$; $M_{(A,B)}{}^{\circ}\partial = true$; $M_{(A,B)}{}^{\circ-} = true$; $M_{(A,B)}\partial^{\circ} = true$; $M_{(A,B)}\partial\partial = true$; $M_{(A,B)}\partial^{-} = false$; $M_{(A,B)}{}^{-\circ} = true$; $M_{(A,B)}{}^{-}\partial = true$; and $M_{(A,B)}{}^{--} = true$. The 9IM corresponding to the above value of these entries is exactly the same as the 9IM for *coversOverfill* shown in Table 2. Therefore, the topological relationship between simple regions with holes A and B is identified as *coversOverfill*.

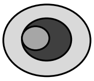


Fig. 4. A simple region with one hole B (light) and a simple region with no hole A (dark) with their shared part shaded the darkest.

4 Conclusions

Although there has been a large amount of research in modeling topological relationships between regions with holes and complex regions, the implementation of these relationships is generally not as widely available as that of topological relationships between simple regions. In this paper, we have shown how we can derive topological relationships between simple regions with holes using only topological relationships between simple regions. This opens up the possibility of using the simpler and more straight forward implementation of topological relationships between simple regions as the basis for implementing topological relationships between simple regions with holes as well as more complex topological relationships such as local and hybrid topological relationships between complex regions.

References

Clementini E and Di Felice P (1996) A Model for Representing Topological Relationships between Complex Geometric Features in Spatial Databases. Information Systems, 90:121–136.

Clementini E, Di Felice P, and Califano G (1995) Composite Regions in Topological Queries. Information Systems, 20:579–594.

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

Egenhofer MJ, Frank A, and Jackson JP (1989) A Topological Data Model for Spatial Databases. In *1st Int. Symp. on the Design and Implementation of Large Spatial Databases*, pages 271–286. Springer-Verlag.

Egenhofer MJ and Franzosa RD (1991) Point-Set Topological Spatial Relations. Int. Journal of Geographical Information Systems, 5:161–174.

Egenhofer MJ and Herring J (1990) Categorizing Binary Topological Relations Between Regions, Lines, and Points in Geographic Databases. Technical report, National Center for Geographic Information and Analysis, University of California, Santa Barbara.

McKenney M, Pauly A, Praing R, and Schneider M (2007) Local Topological Relationships for Complex Regions. In *10th Int. Symp. on Spatial and Temporal Databases*, pages 203–220.

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

Spatial Rules Generate Urban Patterns: Emergence of the Small-World Network

H. Rezayan [(a)], M. R. Delavar [(a)], A. U. Frank [(b)], A. Mansouri [(c)]

[a] Dept. of Surveying and Geomatic Eng., Eng. Faculty, University of Tehran, Tehran, Iran, (rezayan, mdelavar)@ut.ac.ir.

[b] Dept. of Geo-Information E-127, Technische University Wien, Gusshausstr. 27-29, A-1040 Vienna Austria, frank@geoinfo.tuwien.ac.at

[c] Dept. of Landscape Architecture, Fine Art Faculty, University of Tehran, Tehran, Iran, seyedamir.mansouri@gmail.com

Abstract

Objective explanation of urban patterns requires regeneration of these patterns. We defined eight simple spatial rules for locating a building in space and used these rules to simulate re-generation of the small-world network pattern, which is an archetype in structures of cities. We provided a spatial description of how these rules act generating the mentioned pattern. The description is based on using local spatial predictability of the physical reality, incorporating basic spatial global rules, and reducing the indeterminacy of the simulation model. The results show that following the spatial rules derived from the physical reality, it is difficult to avoid generating the small-world network. This clarifies problem of the urban design approaches damaging the small-world network patterns in contemporary cities. The results also propose the small-world network characteristics for cities that are not pre-planned, or more properly organic cities, settled on flat lands.

Keywords: Small-World Network, Spatial Rule, Physical Reality

1 Introduction

Although each city has unique story (Lynch 1981), archetype patterns exist in cities (Salingaros 1998; Hillier, 1989; Hillier, 2001; Carvalho and Penn, 2004; Alexander 2002a, 2002b). Human beings interaction with each other and especially their interaction with the physical reality (the sun, topography, winds, water, and food resources) to derive the maximum utility from living together generate these patterns.

Complexity of cities makes it difficult to investigate the urban patterns and their real causes. This usually leads us to subjective explanations, dominated by political and economical issues. Objective explanation of complex urban patterns requires regeneration of these patterns through effective incorporation of spatial rules.

We studied eight simple spatial rules and the patterns they generate in a simulated cellular environment. These rules are: 1) *Distance-to-Center-of-Gravity*, 2) *Distance-to-Road*, 3) *Free-Space*, 4) *Share-Free-Space*, 5) *Adjoining-Free-Space*, 6) *Access-Space*, 7) *Adjoining-Access-Space*, and 8) *Sun-Position* rule. The rules are selected based on:

1. Simplicity which enables more clear description and justification of the results and interaction of rules[1];
2. Inclusion of rules from related work. Hillier (1989) introduced *Free-Space*, *Share-Free-Space,* and *Adjoining-Free-Space* rules and showed that they can regeneration the beady ring pattern which is an archetype pattern in cities[2];
3. Similarity of the patterns they generate to the reality which is discussed and visually validated by urban designers[3]. This validation provides an overall test of wholeness of the rules[4].

[1] For example among different hypothesis of how a human define nearness and farness, we used Euclidean distance. We will select more complex hypothesis when this one failed generating the patterns we expected.

[2] The beady ring pattern depicts "…constant variation in the width of the open space so fatter areas of space are linked to other fatter areas through thinner spaces. (Hillier 1989, p. 339)"

[3] Urban designers are more familiar with the structural geometric patterns of cities than urban planners and urban architectures. Urban design acts as a mediator and sits at the interface between urban architecture and urban planning. Urban architecture and planning focus on artistic and socio-economic factors and urban design emphasizes on physical attributes that usually restrict its scale of operation to arrangements of streets, buildings, and landscapes (Batty et al. 1998).

[4] Alexander's wholeness viewpoint (Alexander 2002a and 2002b) are adopted to bridge the gap of subjective and objective world of living world.

4. Increasing determinacy of the simulation model using the local spatial predictability of the physical reality. This quality of space is missed or underestimated in previous studies, like Hillier (1989) and Lechner et al. (2006), for they incorporated the physical reality, but as a second-class entity. They firstly simplify space (for example to a horizontal and isotropic space) by removing or reducing the complexities produced by the physical reality. After defining their rules in the simplified space, artificial random processes are used to reconstruct the removed complexity of space. Providing a deterministic model using the local spatial predictability, we considered the sun and pre-existent roads (the effects of access to sun light and roads) here[5];
5. Depicting inseprability of macro and micro world. We defined two basic global rules: *Distance-to-Center-of-Gravity* and *Distance-to-Road* are global rules. It is intended that the global rules be neutral and not being selected to exert any specific pattern[6] and the patterns be generated through interaction of rules, especially local rules. Then we admit explicit global rules only and also test the expected rule-set for generation of different patterns. These will reduce the risk of defining global rules which may dominate the local rules or exert specific patterns.

Our target pattern here is the small-world network pattern. It is an archetype in cities (Salingaros 1998). In the small-world network each node can reach most of the other nodes in short steps, despite they are not neighbors.

The small-world network pattern brings out the required level of connectivity to support life in a city (Salingaros 1998). This pattern represents dynamicity, stability, scale independence, and power-law distribution in a structure, like World Wide Web, social networks, and urban geometric structures. Investigation of the small-world network will provide us with a basis to study emergence of other urban patterns like Zipf's law distribution of length of passages (Carvalho and Penn, 2004) and fractal patterns.

We defined thirteen plausible rule-sets from the mentioned eight rules. These rule-sets are numerically evaluated to define how they generate the small-world network pattern. For each simulation, the graph that represents connections of the simulated free spaces (and the pre-existent roads) are created. Then the mean-shortest path length and the clustering coefficient of the graph are calculated and compared with the trend of the small-world network. Besides, we investigated the scale-freeness quality of the patterns

[5] We will let in other characteristics of the physical reality like topography in further work.

[6] That the rules do not exert specific patterns are considered in definition of local rules, but this rarely happens for their simplicity and locality.

comparing the degree-distribution of their graphs against the power-law distribution and checking stability of its exponent while the patterns grow.

The rules and rule-sets presented in this paper provide one of the possible descriptions for the small-world network pattern of cities. It is intended to clarify the role of spatial rules and geometric structure of cities in bringing life to cities and enrich the urban designers' knowledge about what rules generate cities.

The paper includes 7 sections. The small-world network is described in section 2. Section 3 explains the simulation model. The global and local spatial rules are introduced in section 4 and also their validation process is discussed. The determinacy of the simulation model is described in section 5. Section 6 numerically evaluates emergence of the small-world network pattern in the model. In the end, conclusions are presented in section 7.

2 The Small-World Network

Small-world network is a class of random networks that its nodes are connected by both long and short links (Salingaros, 2001). Then each node in the small-world network can reach most of the other nodes by a small number of steps, although most of the nodes are not neighbors (Fig. 1).

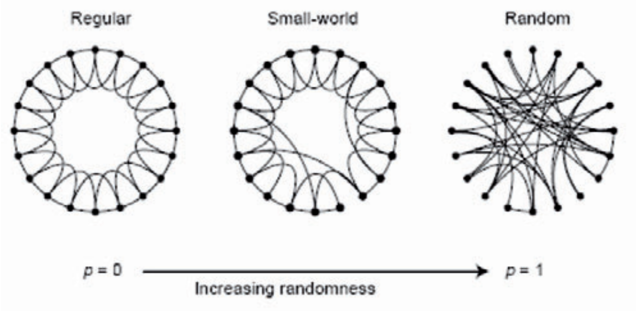


Fig. 1. The small-world network stands between a regular and a random network. It is highly clustered like a regular graph, yet with small characteristic path length, like a random graph. Here *p* represents the probability of connecting a node/vertex to far nodes/vertices, rather than immediate nearest neighbor nodes. It is defined by node's degree. In this figure, nodes have 4 degrees[7] (Watts and Strogatz, 1998).

[7] The proposed method by Watts and Strogatz (1998) for generation of a network with specific probability *p* is to rewire a fully regular graph. They introduce the rewiring process as "We choose a vertex and the edge that connects it to its nearest neighbour in a clockwise sense. With probability *p*, we reconnect this edge

Watts and Strogatz (1998) compared the mean-shortest path length and the clustering coefficient of regular, small-world, and random networks. They define the mean-shortest path length as "… the number of edges in the shortest path between two vertices, averaged over all pairs of vertices. (Watts and Strogatz, 1998)" The clustering coefficient is also defined as the average of edges exists between neighbors of a vertex to maximum number of edges between them over all vertices. The maximum number of edges that can exist between the n vertices is *n*(n-1)/2* for an undirected network.

Watts and Strogatz (1998) showed that the mean-shortest path length of a small-world network and random networks are similarly small, but the clustering coefficient of a small-world network is larger than what is expected for random networks (Fig. 2). It means that a small-world network has few high degree nodes, known as hubs, and the rest of the nodes are peripheral, low degree nodes. It brings stability against changes may happen in the peripheral nodes. It makes the small-world network pattern reliable enough to support life of networks like World Wide Web or a city.

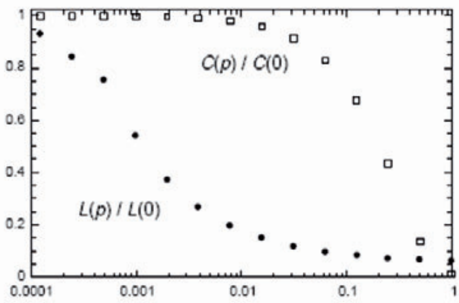


Fig. 2. The mean-shortest path length *L(p)* and the clustering coefficient *C(p)* for the family of randomly rewired graphs with 1000 nodes which have 10 degrees. The x-axis represents the probability of networks (*p*) in logarithmic scale. The values are normalized using *L(0)*, *C(0)* for a regular lattice. (Watts and Strogatz, 1998)

to a vertex chosen uniformly at random over the entire ring, with duplicate edges forbidden; otherwise we leave the edge in place. We repeat this process by moving clockwise around the ring, considering each vertex in turn until one lap is completed. Next, we consider the edges that connect vertices to their second-nearest neighbours clockwise. As before, we randomly rewire each of these edges with probability p, and continue this process, circulating around the ring and proceeding outward to more distant neighbours after each lap, until each edge in the original lattice has been considered once. (Watts and Strogatz, 1998)"

Degree-distribution of the small-world network fits the power-law distribution. It means that the small-world network is scale-free. The small-world network also encourages movement for it inherits the predictability of regular networks and accessibility of random networks.

For a network of spaces in a city, the mean-shortest path length represents how far you should go to be able to reach urban facilities like stations and shops. The clustering coefficient reflects how stable and reliable these accesses are, considering the continuous changes may happen due to human activities, like accidents or constructions, or environmental conditions (e.g. bad weathers) that can hinder or block normal flows in a city. These are the basic characteristics of an urban structure which is alive (Salingaros, 2003).

Improving the damaged small-world network pattern of our contemporary cities, Salingaros (2003) encourages expansion of short links (pedestrian) among the mostly long links (motorway roads) that have already overcome our cities. It is also considered as an alternative to improve the pedestrian passages in cities or even design pedestrian cities. Although it seems that we mastered developing long links like motorways while destroying short links, the current problem is how to do the reverse and expand short links and how to handle the interactions between these mediums of transportation. Filling the gap here, we need to develop our objective knowledge about what causes what (Salingaros, 2003). The question is how pedestrian passages are generated and interact with other mediums of transportation like motorway roads and subways.

In the simulation model presented in this paper, we considered pedestrian passages and motorway roads as our transportation mediums. This case is typical in historical cities that have many short pedestrian links and a few long motorway links usually connect two or more cities together (Lynch 1981, Salingaros 2003, Alexander 2002a). Considering this and the entities of the physical reality we are using (the sun and pre-existent roads or more properly the effects of access to sun light and roads), we expect that simulations become more similar to structure of cities that are not preplanned rather than our contemporary cities[8]. Urban designers name these cities as organic cities (Lynch, 1981).

[8] Then the results do not prescribe solutions for our contemporary cities, but describe what is missed or underestimated. They provide urban designers with more objective description of patterns of cities. It can help in developing more effective theories and applicable procedures for supporting life in our cities

3 Simulation Model

We used a cellular simulation environment. Each cell can be empty or represent a building, a free space, or a road.

The free spaces emerge during simulation. They generate the pedestrian passages between the buildings.

Roads are pre-existent entities in the environment. The width of the roads is fixed at one cell (which is equal to the width of the generated pedestrian passages) to avoid the possible influence of changing width of the roads.

Each simulation starts by putting one or more buildings as seeds in the environment. At each step of the simulation, one building is added to the environment using the rules discussed in the section 4 and considering the buildings already located. Accumulation of the buildings generates the urban patterns.

The proposed simulation approach is a space allocation. It is not an agent-based simulation that is being used for simulation of human decisions (Epstein and Axtell, 1996). Using jargon of the agent-based simulations, however, the proposed simulation would be an agent-based simulation that uses a multitude of agents act in a queue (sequentially and not simultaneously) and accumulate their decisions in the environment. No memory and learning is required here.

4 Rules Validation

The simulated patterns using the eight rules, which are introduced in this section, were presented to four urban designers[9] and they were asked whether the patterns are meaningful or not. The invalidated rules were dropped[10].

Among the rules, *Distance-to-Center-of-Gravity* and *Distance-to-Road* are global rules and the rest are local rules. The *Distance-to-Center-of-Gravity* rule and the *Free-Space* rule are the basic rules, which are used in most of the simulations by default[11]. In the figures shown hereafter, building seeds are shown in dark gray, buildings in gray, roads in light gray, free spaces in white and empty spaces in black.

[9] Two PhD and two MSc scholars.

[10] We did not discuss here the rules that are not validated.

[11] When these two rules are not used, the case is declared in the text.

4.1 Distance-to-Center-of-Gravity

This rule asserts that people prefer to live near each other (Tobler, 1979). In its simplest form, the rule formulates that the sum of the inverse distances from a location to all the buildings already exist in the environment affects the value of the location for adding a new building (Eq. 1). It is the distance to the gravity center of the buildings (Fig. 3a).

$$Value\text{-}of(aLocation) \approx \sum_{aBuilding}^{Buildings} \frac{1}{Distance(aLocation, aBuilding)} \tag{1}$$

This rule aggregates the buildings, defining the feasible locations with maximum value among the immediate neighbors of the buildings already exist. It forms a layer-by-layer pattern (Fig. 3b).

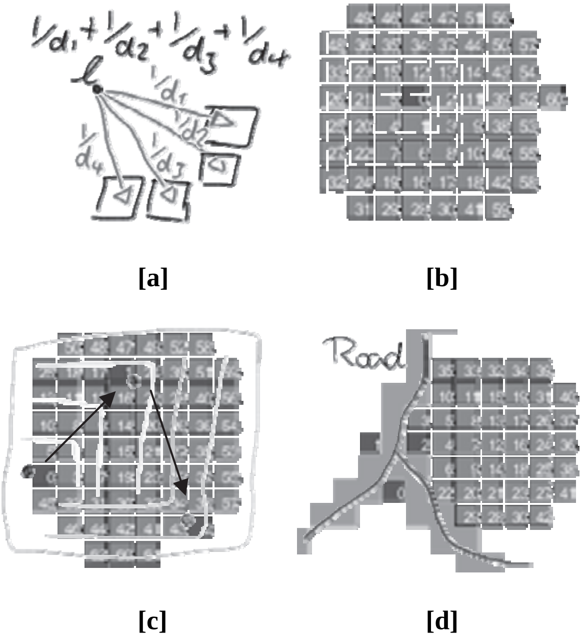


Fig. 3. [a] Schematic representation of the Distance-to-Center-of-Gravity rule. It values each location of space as sum of its inverse distances to existing buildings. **[b]** This rule creates layer-by-layer patterns around the seed placed in the environment by the user. **[c]** When detached seeds are used to initiate the simulation the generated pattern moves like a wave to cover all seeds. **[d]** This pattern can not pass entities like roads.

Starting the simulation with more than one seeds, the pattern starts from one of the seeds. It grows aggregating the buildings layer-by-layer moving wave-like towards the other closest seed and continues until include other seeds (Fig. 3c). After that, the pattern grows as described for Fig. 3.b.

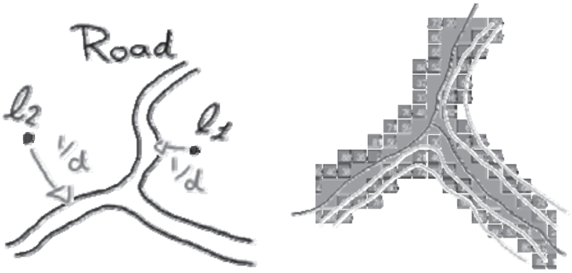
Any entities that occupy parts of the outer layer of an aggregation of buildings will block further growth of the pattern at those parts (Fig. 3d). Then a road will behave as an impassable edge against this rule, while in reality a road may hinder the crossing but does not block it.

4.2 Distance-to-Road

This rule says that distance to the roads affects the value of the location for adding a new building (Fig. 4a). It maximizes the inverse distance of a location to the road (2).

$$\mathit{Value\text{-}of(aLocation)} \approx \max\Big(\bigcup_{\mathit{aBuilding}}^{\mathit{Buildings}} \frac{1}{\mathit{Distance(aLocation,aBuilding)}}\Big) \qquad (2)$$

This rule aggregates buildings around the roads. The pattern grows layer-by-layer along the roads (Fig. 4b). Unlike the *Distance-to-Center-of-Gravity* rule, the roads behave here as passable edges.



[a] [b]

Fig. 4. [a] Schematic representation of the Distance-to-Road rule. It values each location (l) based on inverse shortest distance to roads. **[b]** This rule creates layer-by-layer patterns, defined by the parallel lines, around the roads.

The value derived using this rule just depends on the shape of roads and the location in the environment. So this value is static and does not change after adding new buildings. But the value derived from the *Distance-to-Center-of-Gravity* rule increases through the time. Then combining these two rules, their values are normalized.

4.3 Free-Space

Hillier (1989) introduced *Free-Space* rule as each building should have an exclusive free space attached to it (Fig. 5a and 5b). This free space is sort of a front yard. The free space enables going in and out of a building.

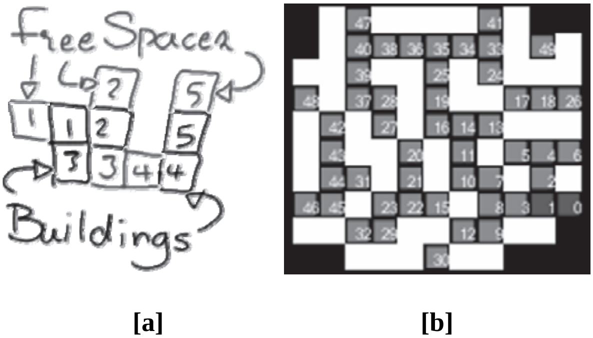


Fig. 5. [a] Schematic representation of the Free-Space rule. **[b]** This rule is the basis of other local rules defined here.

4.4 Share-Free-Space

Hillier (1989) introduced this rule. It says that free space of a building can be shared between and among two or more other buildings (Fig. 6a). The main effect of this rule is compactness of the free spaces (Fig. 6b).

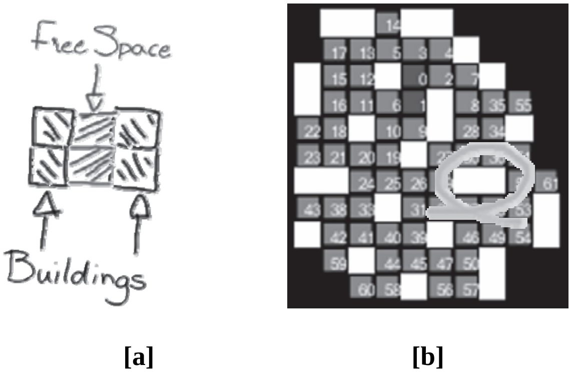
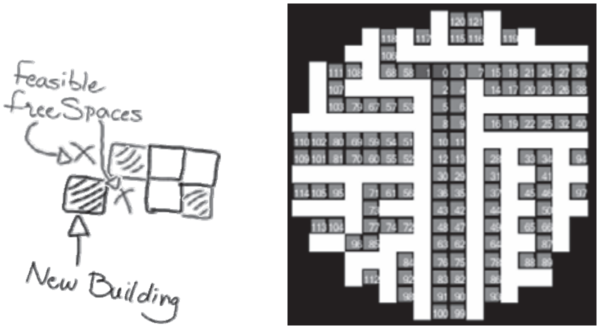


Fig. 6. [a] Schematic representation of the Share-Free-Space rule. It discards assignment of new free space to a building if any free spaces still exist in 4 neighbors of the building. **[b]** Compacting the free spaces, this rule also creates enclosed detached free spaces like what is marked in the figure.

It also generates rows of buildings formed by pair of buildings, which are located back to back. This pattern is more realistic than the patterns generated by sole usage of the *Free-Space* rule that generates rows of building from single buildings located side by side. This rule may also generate detached free spaces bounded by buildings, like an exclusive passage in a building complex (Fig. 6b).

4.5 Adjoining-Free-Space

Hillier (1989) introduced this rule that forces a free space to adjoin other free spaces, if possible (Fig. 7a). Emergence of terraced buildings is the main effect of this rule (Fig. 7b).

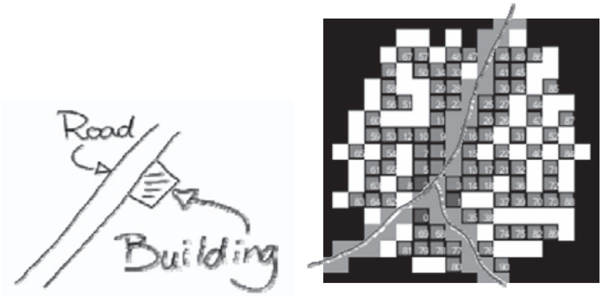


[a] **[b]**

Fig. 7. [a] Schematic representation of the Adjoining-Free-Space rule. **[b]** It creates terraced buildings.

4.6 Access-Space

This rule says that a road can provide its neighbor buildings with their required free spaces (Fig. 8). The neighbor buildings share the road as their free spaces. Their entrances face toward the road. This rule creates more real patterns around the roads than using the *Distance-to-Road* rule alone. It also compacts free spaces along the roads. The *Access-Space* rule causes roads to behave as passable edges even when rules like the *Distance-to-Road* are not used.



[a] **[b]**

Fig. 8. [a] Schematic representation of the Access-Space rule. **[b]** This rule causes compaction along the roads. It enables the pattern to pass the roads.

4.7 Adjoining-Access-Space

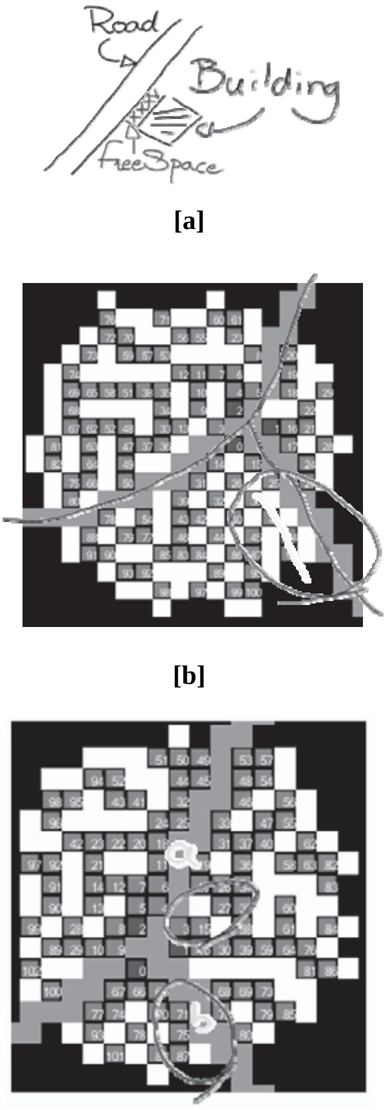
This rule states that the buildings that are near a road tend to face toward the road. These buildings use the gap between themselves and the road as their free spaces (Fig. 9a). The expected effect of this rule is near-road buildings that follow the shape of the road (Fig. 9b). The free space of these buildings somehow widens the road and looks like sidewalks.

Emergence of these near-road buildings, however, is not strict. The locus of the feasible locations to apply this rule is just the two cells buffer of the roads. No specific rule entails location of building in this locus.

It seems that the *Distance-to-Road* rule can overcome this situation, but it firstly fills the first layer of buildings adjoining the road (one cell buffer of the road). So practically, the second layer that contains the expected locations for the *Adjoining-Access-Space* rule will be blocked and it can not be applied for them.

Adding the *Distance-to-Center-of-Gravity* rule, we can avoid filling all locations in the first layer which is adjoining the roads. This provides the opportunity of locating the near-road buildings.

In this situation, adding other rules like he *Access-Space* rule will affect the pattern and location of the near-road buildings. Then retreated buildings (Fig. 9c{a}) or rows of buildings along a road (Fig. 9c{b}) may emerge.



[a]

[b]

[c]

Fig. 9. **[a]** Schematic representation of the Adjoining-Access-Space rule. **[b]** It creates buildings parallel to roads and **[c]** retreated buildings, marked as {a}, and rowed buildings, marked as {b}.

4.8 Sun-Position

This rule holds that each building tends to attain the maximum sun light in their attached free space. Then we assign the free space of a building based on the cardinal directions. The directions are prioritized as south, east, west, and north (Lynch, 1981). This rule discards any random selection of free spaces.

5 Determinacy of the Simulation Model

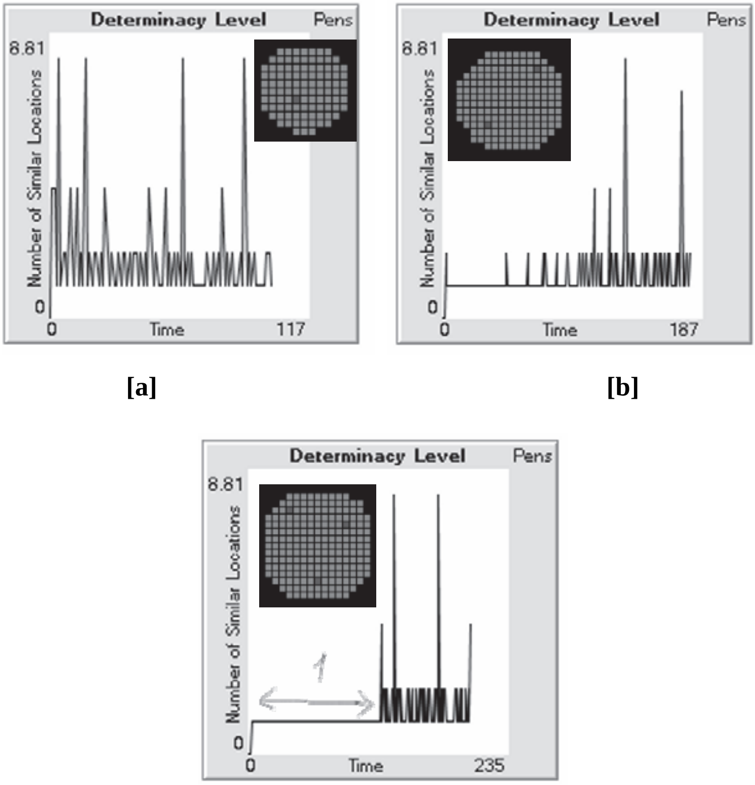
Although the free spaces can be located deterministically using the *Sun-Position* rule, still we are using a random generator to locate the buildings. It seems that deterministic location of buildings should be satisfied through combination of rules.

We just discussed location of buildings (not the free spaces) in the next two sub-sections. The point is to define how the rules distinguish locations of buildings. At each iteration, we count the ties that are the locations which have the same value equal to the maximum value (1 and 2). One of these locations should be selected for putting a new building. Then higher determinacy will be achieved if the number of ties decreases.

5.1 Global Rules

The *Distance-to-Gravity-Center* rule provides the simplest pattern. It causes the highest level of indeterminacy (Fig. 10). The number of similar ties varies from 1 to 8 (Fig. 10a). When there is more than one seed the process shows higher determinacy (Fig. 10b and Fig. 10c) until the pattern includes all the seeds. Then the determinacy decreases and become similar to the simulation with one seed (Fig. 10a). Sole usage of the *Distance-to-Road* rule causes high indeterminacy, as there are many locations with similar shortest distance to roads (Fig. 11).

Using the *Distance-to-Gravity-Center* rule and the *Distance-to-Road* rule together the number of similar locations reduces to 2-3 locations (Fig. 12). The *Distance-to-Road* rule increases the determinacy, especially when the pattern meets the roads (Fig. 13a). This is due to the irregularity exerted by the roads. More regular roads cause more indeterminacy (Fig. 13b). More complex roads make faster emergence of determinacy happen in the model (Fig. 13c).

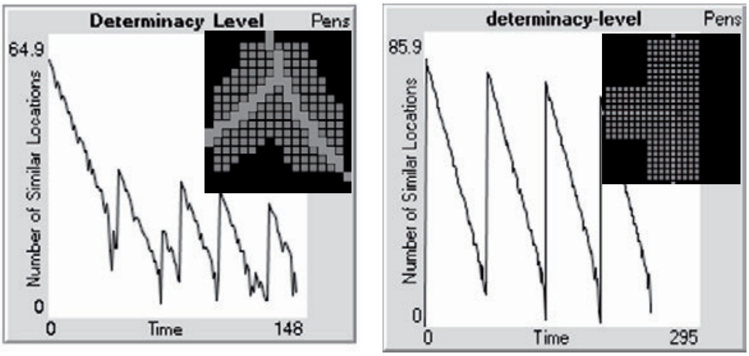


[a]

[b]

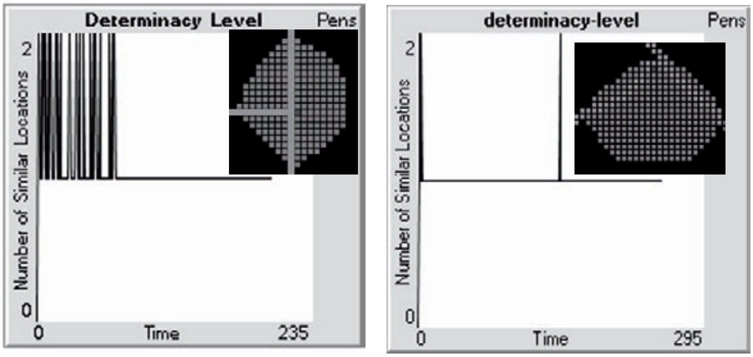
[c]

Fig. 10. Model determinacy just using the *Distance-to-Gravity-Center* rule. **[a]** Using one seed causes indeterminacy from the start and it continues. **[b]** Using two separate seeds reduces the indeterminacy until the pattern includes both the seeds. **[c]** With three separate seeds, simulation starts without indeterminacy and continues until the pattern includes all the seeds.



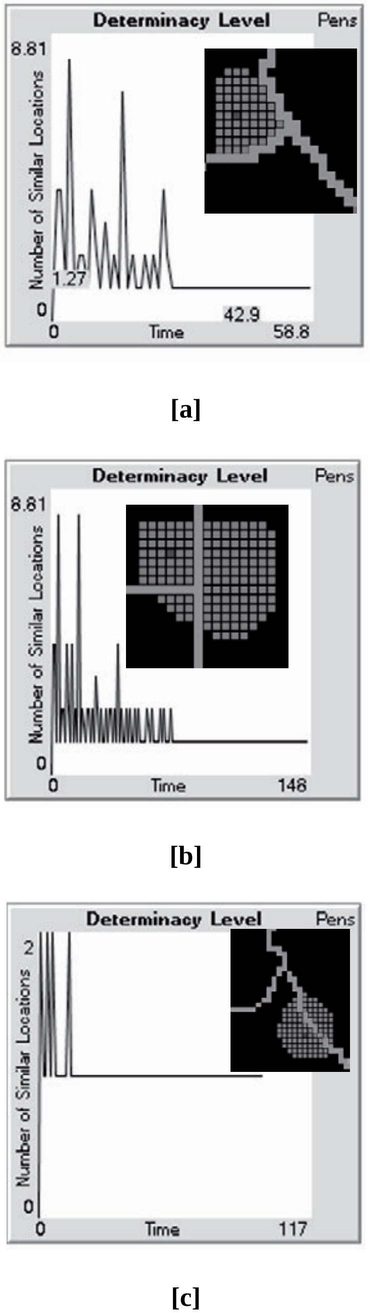
[a] **[b]**

Fig. 11. Model determinacy just using the *Distance-to-Road* rule causes **[a]** high indeterminacy which is represented as jagged rise and fall of the determinacy level **[b]** even when the roads are regular and simple.



[a] **[b]**

Fig. 12. Model determinacy rises using the *Distance-to-Road* rule and the *Distance-to-Road* rule together. **[a]** Using regular roads, determinacy varies between 2 and 3 possible locations. **[b]** Using irregular roads improves determinacy as 3 ties rarely happen.



[a]

[b]

[c]

Fig. 13. Model determinacy adding the *Distance-to-Road* rule, when having **[a]** Irregular simple roads, **[b]** Regular roads, or **[c]** Irregular complex roads.

5.2 Local Rules

Adding the free spaces increases the irregularity in the generated patterns and reduces the indeterminacy of the simulation model. The simulations carried out for different combination of rules and initial states (arrangement of seeds and geometry of roads) shows that still minor indeterminacy is possible, especially at first iterations when a few buildings still exists in the environment (Fig. 14). As the simulation goes on, the indeterminacy disappears. Combining the defined global and local rules, we did not observe any indeterminacy after 30 iterations.

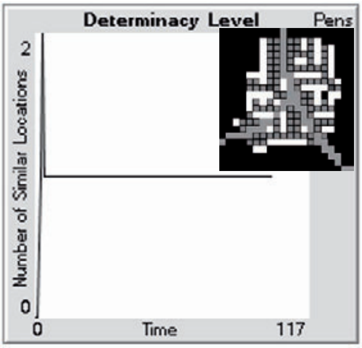


Fig. 14. Model determinacy combining the defined global and local rules. It shows some random peeks at the beginning of the simulation, when few buildings still exist in the space. Repeating the simulation with different arrangements of seeds and different geometry of roads, the number of similar locations remains under 2 locations.

6 Numerical Evaluation of the Small-World Network Pattern Emergence in the Model

Fig. 15 shows the small-world network characteristics as a semi-logarithmic graph (Fig. 2) generated, using the rewiring approach introduced by Watts and Strogatz (1998) in section 2, with 4 degrees at 0.01 intervals. We selected 4 degrees for the small-world network to enable each node to become a cross intersection that are typical in urban structures. The 0.01 interval is selected arbitrarily. It affects our comparison precision as we derive the nearest value of *p* on x-axis separated at 0.01 intervals (Table 1).

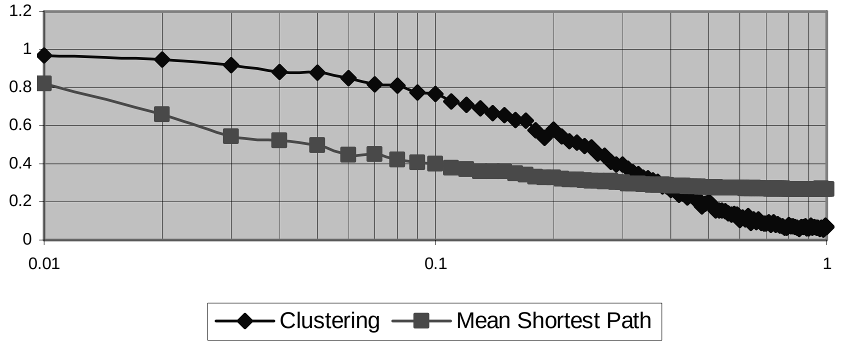


Fig. 15. Characteristics of the small-world Network for 100 nodes with 4 degrees at 0.01 intervals. The Maximum differences between the clustering coefficient and the mean-shortest path length, ranges from 0.37 to 0.40, happens between p=0.03 and p=0.09. The value for p=0 is dropped because of using logarithmic scale in x axis.

Table 1 lists different rule-sets used to generate the patterns. We executed the simulation for each rule-set, considering different (about 10) arrangement of seeds (single, multiple, attached, and detached) and different geometry of roads (regular and irregular). Then the averages of the following values are calculated (Table 1):

- the clustering coefficient,
- the mean-shortest path length,
- $P_{(clustering\ coefficient)}$ and $P_{(mean\text{-}shortest\ path\ length)}$ that are derived from the graph showed in Fig. 15, represent the corresponding probability value of the small-world networks that produce the simulated clustering coefficient and the mean-shortest path length values,
- D_p as the absolute difference between $P_{(clustering\ coefficient)}$ and $P_{(mean\text{-}shortest\ path\ length)}$, and
- PLE_{mean} and PLE_{Std} which represent mean and standard deviation of the power-law distribution exponent fitted to the degree distribution of the simulation.

Table 1. Emergence of the small-world network pattern in the simulation model.

Rules Set	Roads Exists?	Distance-to-Road	Share-Free-Space	Adjoining-Free-Space	Access-Space	Adjoining-Access-Space	Sun-Position	Clustering Coefficient	Mean-Shortest Path Length	$P_{\text{clustering coefficient}}$	$P_{\text{mean-shortest path length}}$	Dp	PLE_{mean}	PLE_{Std}
1							✓	0.485	0.638	0.260	0.030	0.230	-2.237	0.621
2			✓	✓				0.586	0.764	0.187	0.019	0.169	-3.042	0.564
3				✓			✓	0.678	0.642	0.143	0.027	0.117	-1.361	0.342
4			✓	✓			✓	0.496	0.810	0.240	0.015	0.225	-2.726	0.468
5				✓				0.777	0.523	0.100	0.045	0.055	-1.870	0.485
6			✓	✓				0.594	0.841	0.180	0.015	0.165	-3.066	0.591
7	✓			✓			✓	0.791	0.474	0.090	0.065	0.025	-2.109	0.427
8	✓			✓				0.882	0.488	0.050	0.060	0.030	-2.360	0.296
9	✓	✓		✓				0.772	0.393	0.103	0.103	0.007	-1.484	0.830
10	✓	✓		✓			✓	0.781	0.428	0.090	0.090	0.030	-1.055	0.847
11	✓	✓		✓	✓		✓	0.874	0.502	0.050	0.050	0.000	-1.497	0.538
12	✓	✓		✓	✓	✓		0.894	0.457	0.045	0.060	0.015	-1.772	0.521
13	✓	✓		✓	✓	✓	✓	0.847	0.463	0.065	0.060	0.005	-2.038	0.472

The *Distance-to-Center-of-Gravity* and *Free-Space* rules are set on by default.
The values are averages of the results of the simulations carried out for each of the rule-sets. Then D_p might not be exactly equal absolute difference of $P_{\text{(clustering coefficient)}}$ and $P_{\text{(mean-shortest path length)}}$.
Small D_p and PLE_{Std} represent emergence of a valid and stable the small-world network pattern that we expect for a city. Large D_p shows that the simulated pattern does not refer to unique small-world network.
Large PLE_{Std} represents instability of values calculated for the generated small-world network.

Then the main results are as follow:

- The rule-sets 9 to 13 assert the discussions in section 2 that existence of a few roads (long links) among a large amount of pedestrian passages (short links) have significant effect in emergence of small-world networks. This effect is visible, even when the *Distance-to-Road* rule is off (rule-sets 7 and 8). The *Distance-to-Road* rule improves the small-world network pattern more.
- The *Share-Free-Space* rule damages the small-world network pattern (rule-sets 2, 4, 6, and 14). Combination of the *Adjoining-Access-Space* rule and the *Sun-Position* rule has generative effects (rule-sets 1, 3, and 5).
- The p value of the results with low D_p and PLE_{Std}, like the rule-sets 10 and 13, ranges from 0.05 to 0.09 that approximately fits in the range of the maximum differences between the clustering coefficient and the mean-shortest path length which happens between p=0.03 and p=0.09 (Fig. 15). The rule-sets 10 and 13 also have similar PLE_{Std} equals 0.472 and quite similar PLE_{mean} around -2.0.

7 Conclusions

We propose that the values of the rule-sets 10 and 13 (Table 1) represent the characteristics of not pre-planned cities, or more properly organic cities, settled on flat areas (no topography effect). Consider that the rules and the rule-sets introduced here are just one possible alternative to describe the studied urban patterns.

Investigation of the small-world network pattern provides us a basis to study emergence of other urban patterns like Zipf's law distribution of length of the passages (Carvalho and Penn, 2004) and fractal patterns. The rule-sets 10 and 13 (Table 1) are the first candidates to derive these patterns.

We are still using a random generator to define location of the buildings. The introduction of new entities, rules, and attributes from the physical reality, like topography and access to natural resources, will reduce the indeterminacy (the number of ties) in the simulation model. We might be able to omit the random generator.

Our numerical evaluation showed that D_p remains under 0.1 when roads exists in the simulation model (Table 1). It seems that following spatial rules derived from the physical reality, it is difficult to avoid emergence of the small-world network pattern. This observation restates the criticism made by urbanists like Alexander and Salingaros against our contemporary cities which their effective small-world network patterns are damaged. They say that people have to be educated to behave against the archetypes and create such unconnected (not so alive) cities. They refer to defectiveness of our modern (20th century) urban architecture and design educations and plans made by governments. The problem is that economical and political issues and top-down design approach have dominated our urban design. It dismisses or underestimates the spatial rules (especially local rules) that generate urban patterns.

Urban designers can enrich their top-down global viewpoints on patterns with detailed bottom-up regeneration of those patterns. This also helps to raise questions like how much information exists in urban patterns. For example, we can count rules or measure the length of the program generated the pattern. We used 8 rules here that are programmed in less than 100 lines. An expected result would be that our urban design approaches not only dismissed some important (spatial) rules, but also carries a large amount of redundancies[12]. This knowledge can help to design better cities and improve the quality of life in the existent cities.

Acknowledgments

We acknowledge Dr. Mansouri, Dr. Sabri, Eng. Ebrahimi, and Eng. Najian for participating in rules validation. We used NetLogo 3.1.2 (http://ccl.northwestern.edu/netlogo/) to create the simulation model (Wilensky, 1999). We also carried out the analyzes using python 2.4, extending the network library developed by Aric Hagberg (hagberg@lanl.gov), Dan Schult (dschult@colgate.edu), and Pieter Swart (swart@lanl.gov). You can contact the authors for flow diagram of the simulation model implemented in this paper.

References

Alexander C (2002a) The Nature of Order: The Phenomenon of Life. Taylor & Francis, London

Alexander C (2002b) The Nature of Order: The Process of Creating Life. Taylor & Francis, London

Batty M, M Dodge, Jiang B, and Smith A (1998) GIS and Urban Design. Technical Report, Center for Advanced Spatial Analyses – CASA

[12] Redundancies are not useless necessarily if they bring stability and do not have significant effect on the results.

Carvalho R; Penn A (2004) Scaling and universality in the micro-structure of urban space. Physica A 32:539-547

Epstein JM, Axtell R (1996) Growing Artificial Societies: Social science from the bottom up. Brookings Institution Press, Washington D.C.

Hillier B (1989) The Architecture of the Urban Object. Ekistics, 56:5-21.

Hillier B, (2001) A Theory of the City as Object: How Spatial Laws Mediate the Social Construction of Urban Space. In: 3rd International Space Syntax Symposium Proceeding, Atlanta, USA

Lechner T, Watson B, Wilensky U, Felsen M (2006) Procedural Modeling of Land Use in Cities. In: ACM SIGGRAPH Conference proceeding

Lynch K (1981) Good City Form. Massachusetts Institute of Technology.

Salingaros NA (1998) Theory of the Urban Web. Journal of Urban Design 3:53-71

Salingaros NA (2001) Remarks on a City's Composition. In: Resource for Urban Design Information (RUDI)

Salingaros NA (2003) Connecting the Fractal City. In: 5th Biennial of towns and town planners in Europe, Barcelona, Spain

Tobler W (1979) Smooth Pycnophlactic Interpolation for Geographical Region. Journal of the American Statistical Association 74:519-530

Watts DJ, Strogatz S (1998) Collective dynamics of 'small-world' networks. Nature 393:440–442

Wilensky U (1999) NetLogo. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL

Conceptual Neighborhoods of Topological Relations Between Lines

Rui M.P. Reis[1], Max J. Egenhofer[2], João L.G. Matos[3]

[1] Instituto Geográfico Português, Direcção de Serviços de Investigação e Gestão de Informação Geográfica, Rua Artilharia Um, 107, 1099-052 Lisboa, Portugal; e-mail: rui.reis@igeo.pt

[2] National Center for Geographic Information and Analysis, Department of Spatial Information Science and Engineering, Department of Computer Science, University of Maine, Orono, ME 04469-5711, USA
e-mail: max@spatial.maine.edu

[3] Departamento de Engenharia Civil e Arquitectura Instituto Superior, Técnico, Av. Rovisco Pais, 1049-001 Lisboa, Portugal
e-mail: jmatos@civil.ist.utl.pt

Abstract

Conceptual neighborhood graphs form the foundation for qualitative spatial-relation reasoning as they capture the relations' similarity. This paper derives the graphs for the thirty-three topological relations between two crisp, undirected lines and for the seventy-seven topological relations between two lines with uncertain boundaries. The analysis of the graphs shows that the normalized node degrees increases, from the crisp to the broad-boundary lines, roughly at the same degree as it increases for crisp lines that are transformed from R^1 into R^2.

1. Introduction

Spatial databases need semantically rich geometric data types to describe appropriately spatial configurations. The definitions of such data types include the specification of their data structures, the identification of operations on instances of that type, as well as the identification of the *spatial relations* between such instances. Often these spatial relations are qualitative in nature, abstracting away quantitative details. Topological relations have been the predominantly studied field of such qualitative spatial relations with the identification of binary relations between regions (Egenhofer and Franzosa 1991), lines, and points (Egenhofer and Herring 1991).

The elements of such a set of possible relations are essentially on a *nominal* scale (Stevens 1946), which enables the distinction when two relations are the same or different, but the categorization of relations *per se* yields no further information about the relationships *among* non-equal relations. Such relationships, however, are germane for deciding about higher-level concepts about the spatial relations, such as order, partial order, or similarity. To relate qualitative relations one typically constructs the relations' *conceptual neighborhood graph* (Freksa 1992), which captures explicitly all those pairs of relations that are most similar. Such an organization of a set of relations brings these relations onto a level where more than nominal comparisons can be made, enabling richer analyses.

Nodes in a conceptual neighborhood graph represent spatial relations, while edges are created to connect the relations with the least differences (Bruns and Egenhofer 1996). Since some relations are closer to each other than others, the differences among the relations offer an opportunity to determine the relations' similarities. Paths in the graph refer to sequences of different spatial relations as they result from continuous deformations of the objects. The conceptual neighborhood graph of a set of relations provides a rationale for answering three types of questions:

- Given two relations find the intermediate relations and possible alternative paths, if they exist.
- Given a relation and a particular change process (such as a translation, a rotation, or a scaling), determine set of next possible relations.
- Given two consecutive relations, determine the possible change processes that were involved.

This paper studies the conceptual neighborhood graph of topological relations between *undirected lines*. This graph has been a missing piece in the puzzle of conceptual neighborhood graphs. We pursue a similar approach as for the topological relations between two regions, exploring the graph for crisp lines as well as for lines with broad boundaries. The resulting graphs form a rationale for qualitative similarity reasoning and may also serve in the future as a framework for identifying the semantics of natural-language relations, similar to earlier approaches for relations between lines and regions (Mark and Egenhofer 1994; Kurata and Egenhofer 2007).

The remainder of this paper is structured as follows: Section 2 reviews work on conceptual neighborhood graphs, particularly those for intervals, regions, and directed lines. Sections 3 and 4 derive the conceptual neighborhood graphs for the thirty-three topological relations between two simple, undirected lines and the seventy-seven topological relations between two lines with broad boundaries. Section 5 analyzes these graphs and compares them with the graphs for intervals and regions. Section 6 offers conclusions and discusses items for future work.

2. Conceptual Neighborhood Graphs

The earliest developments of conceptual neighborhood graphs applied to binary relations between intervals in R^1 (Freksa 1992) and to binary topological relations between regions in R^2 (Egenhofer and Al-Taha 1992). For Allen's (1983) thirteen interval relations a type of similarity is established by moving one of the two ends of an interval while keeping the other end fixed. All possible transitions of this kind are then captured in the conceptual neighborhood graph (Figure 1a). Other types of deformation, such as moving both ends at the same time, leads to somewhat different links, although the overall structure of the graph is preserved.

For topological region-region relations the conceptual neighborhood graph (Figure 1b) was derived from the relations' 9-intersection matrices, considering those pairs of relations as neighbors that feature the least non-zero difference in their matrix elements. In analogy to the interval graph, different types of continuous deformations of the related objects may lead to slightly different graphs, adding occasionally additional links between some nodes. Overall, however, the general framework of the conceptual neighborhood graph remains the same.

Such conceptual neighborhood graphs have been developed for most every set of spatial relations studies, including relations between two cyclic intervals (Hornsby et al. 1999) and between an interval and an interval with a gap (Egenhofer 2007), topological relations between regions on the sphere (Egenhofer 2005) and in Z^2 (Egenhofer and Sharma 1993), for topological relations between regions and lines (Egenhofer and Mark 1995), for topological relations between minimum bounding rectangles (Papadias et al. 1995), convex hulls (Clementini and Di Felice 1997), regions with broad boundaries (Clementini and

Di Felice 1996; Cohn and Gotts 1996), and for the orientation of two lines in the plane (Schlieder 1995).

Most relevant for the development of the conceptual neighborhood of line-line relations is, however, the neighborhood graph for topological relations between two directed lines (Kurata and Egenhofer 2006), which features two parallel layers for relations, one for relations whose lines' interiors do not intersect and another for relations whose interiors do intersect (Figure 1c). In this depiction the nodes along the top and right fringes are repetitions of the nodes along the bottom and the left; therefore, a less redundant conceptual neighborhood graph that shows a single node for each relation warps around the surfaces of two tori, one inside the other, with spike-like connections between the tori, connecting corresponding relations that differ only by their interior-interior intersections.

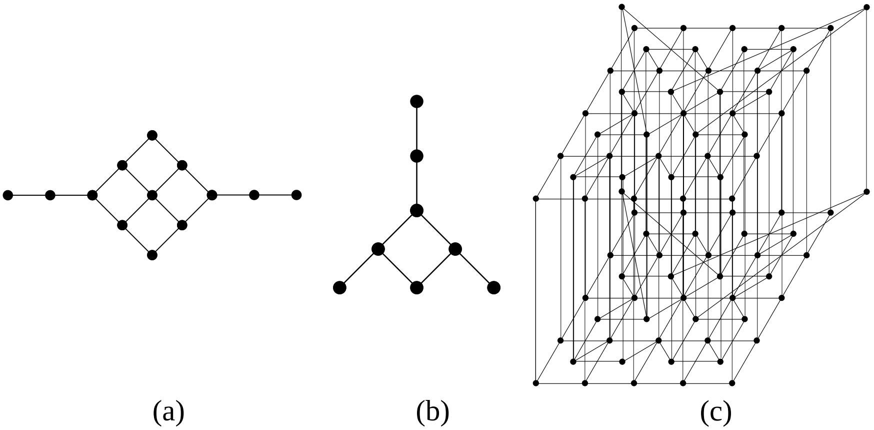


Fig. 1. The conceptual neighborhood graphs of (a) the thirteen interval relations in R^1, (b) the eight topological relations between two regions in R^2, and (c) the sixty-eight topological relations between two directed lines in R^2.

3. Conceptual Neighborhood Graph For Topological Relations Between Two Undirected Lines

The 9-intersection distinguishes 33 different topological relations (Figure 2) between two simple, undirected lines (Egenhofer 1993). Such simple lines have exactly two distinct end nodes and feature no self-intersections, bifurcations, or loops. A major difficulty in the development of these relations' conceptual neighborhood graph is the lack of a clear specification of what establishes similarity between two pairs of topological line-line relations. Since these lines are one-dimensional features that are embedded in a two-dimensional space, they have a higher degree of freedom than two lines in R^1 or two regions in R^2, for which continuous deformations, which establish the rationale for neighborhood, are much more constrained and can be carried out in a much more controlled fashion. For example, in R^2 with a single movement two lines that are *equal* (LL 22 in Figure 2) can be transformed so that they are *disjoint* (LL 1). Such an atomic change of relations would be impossible with a continuous transformation for two lines in R^1 (Figure 2a), because in order to become disjoint the lines would need to migrate through intermediate several steps where, at one point, they *overlap* and later *meet*. The analog holds for the transitions of two regions from being *equal* to *disjoint* in R^2 (Figure 1b).

It is this higher degree of freedom that requires a more lenient use of the typical method that has been used to determine the similarity of topological relations, namely to count the differences in all pairs of the relations' 9-intersection matrices and to consider, for each relation, those as neighbors that feature the least differences (Egenhofer and Al-Taha 1992). While this rationale is still valid to determine prime candidates for conceptual neighbors, the single difference count in the matrices is too restrictive at times. For instance, the transition from LL 15 to LL 24 can be accomplished by an atomic deformation, pulling the end of the line from the other line's interior to its boundary. This change, however, implies a difference of two in their matrices, because the moved line's boundary moves not only out of the interior (one change) but also into to other line's boundary (a second change). Since both LL 15 and LL 24 have other neighbors that differ by only one count, the least difference in matrix elements does not lead appropriate to identifying neighboring relations.

Therefore, we start deriving the lines' conceptual neighborhood graph from the graph of directed lines. Both sets of relations have relation pairs that differ only by their interior intersections, which creates for the line-line relations a similar dichotomy as for the relations between two directed lines. Another tempting inference cannot be made, however. Although each line can be refined with two orientations—one from a start node to the end node, and the other in the reverse direction—this distinction does not double the number of relations, because in a number of cases the distinction of two orientations is immaterial for the directed lines' topological relation, while in others it means that for each line-line relation there are four refinements due to the lines' orientations.

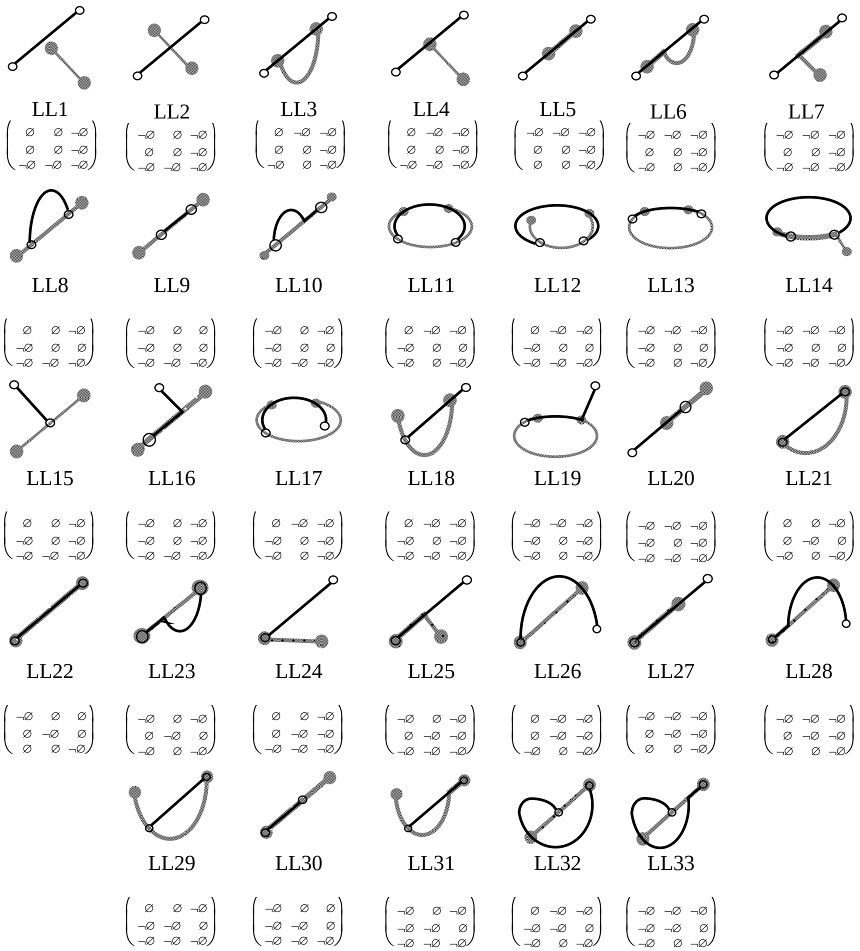


Fig. 2. The thirty-three topological relations between two undirected lines identified by the 9-intersection (Egenhofer 1995).

Within a layer of the directed-line neighborhoods, this impact of the orientation on the line can be observed directly (Figure 3a). The graph's horizontal and vertical axes partition it into four quadrants, each of which captures for the same line-line relation a different aspect of the directed lines' orientations. At the intersection of the two axes is a relation for which the lines' orientations have no impact. If for all relations the orientation of one directed line is ignored, then those relations merge with the mirror images in the graph, either along the horizontal or the vertical axis. If subsequently the other line's orientation is ignored as well, then another merger occurs, this time of the remaining relations along the second axis, leaving as the relations between two undirected lines those that fall into a single quadrant (plus those that coincide with the mirror axes). Therefore, the framework of the conceptual neighborhood graph for the relations between two undirected lines resorts to thirteen relations located in one of the four quadrants (Figure 3b). The parallelism of line-line relations with empty and non-empty interior-interior intersections yields two parallel layers, in which each corresponding pair of line-line relations is connected (Figure 3c).

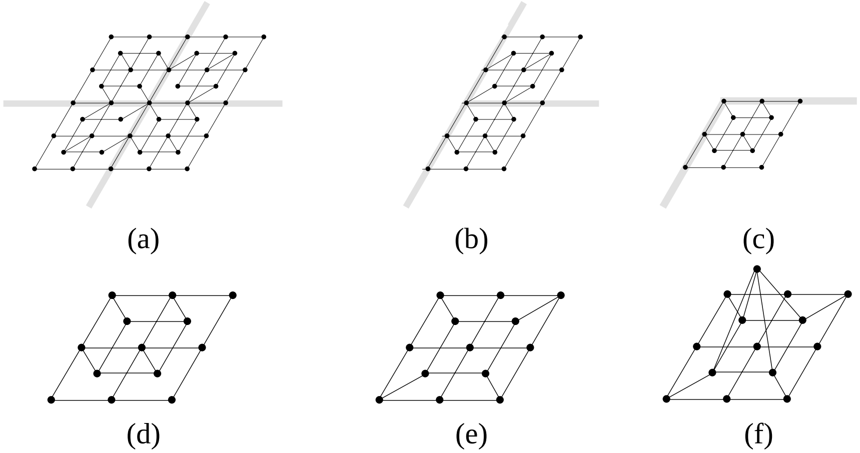


Fig. 3. Deriving a layer in the conceptual neighborhood graph of the line-line relations from the graph for directed lines: (a) the mirror axes in the directed lines' graph; (b) the graph after ignoring the orientation for one of the two lines; (c) the graph after ignoring the orientation of the second line; (d) the derived framework for the conceptual neighborhood graph of the line-line relations; (e) the same framework after adjusting for differences due to the consideration of the lines' exteriors; and (f) the pyramid formed to connect the relation with coinciding boundaries to the layer.

The conceptual neighborhood of the directed-lines' relations was derived from matrix differences of the *head-body-tail intersection* (Kurata and Egenhofer 2006), however. Unlike the 9-intersection it ignores intersections with exteriors, therefore, some of the connections within a quadrant need to be adjusted for the 9-intersection-based line-line relations (Figures 3d and 3e). For the line-line relations' conceptual neighborhood graph, all connections across the two layers now reflect a single difference in the corresponding 9-intersection matrices.

Special attention needs to be paid to two irregularities, however. First are the relations where the two lines' nodes coincide are not integrated into the regularity of the two layers, but—for directed lines—yield two spikes that come out of the layer. When ignoring the lines' orientations, these spikes are merged into a single relation. From this relation, the least matrix difference to any of the relations in the layers is two, not one as among the other relations in a layer, which provides further evidence for that relation's location in the graph. There are four such connections of length 2 in a layer, distributed equally around the central node, which yields a pyramid that emerges from a layer (Figure 3f).

The second exception is the remaining set of five relations (i.e., 33 line-line relations minus 2 layers, each at 13 relations, minus 2 times one spiked relations), which do not yet fit into the overall framework. These five relations include equal (LL 22), the direction-independent versions of starts and finishes (LL 27 and LL 30), and the true subset and its converse relation (LL 5 and LL 9). All five have a non-empty interior-interior intersection, which puts them to the layer with the non-empty interior-interior intersections. Relations LL 5, LL 9, LL 27, and LL 30 each have a neighbor with a one-unit difference to the layer with the non-empty interior-interior intersections. Also LL 27 and LL 30 are within one unit from LL 5 and LL 9, respectively.

These considerations lead to a conceptual neighborhood graph (Figure 4) that strongly resembles the graph of the directed-line relations (Figure 1c). It features again two connected layers, each with a spiked node. In addition, there is a reduced top layer connected to the relations with non-empty interior-interior intersections, including a spiked node. The line-line graph, however, lacks the repetition of relations along the fringes so that no warping into a higher space is suggested.

This conceptual neighborhood graph of the topological line-line relations exhibits some of the properties that have been found with other relations' graphs that were derived from the matrix differences of their 9-intersections.

- Pairs of relations that differ by one entry in their relation matrices form the connections between the extreme landmarks of the conceptual neighbors.
- A one-unit difference is not necessarily possible for all relations within a connected graph. In some cases (e.g., from LL 21 to LL 24), the smallest difference between two neighbors may be two units in the matrix differences. It confirms the insight first gained with the conceptual neighborhood graph for region-region relations where the matrices of the neighbors *equal* and *covers*, as well as *equal* and *coveredBy*, are three units apart (Egenhofer and Al-Taha).

- The least number of differences is not necessarily symmetric. For instance, from LL 23 the nearest neighbors are LL 25 and LL 21, because their matrices differ from LL 23 by one unit. On the other hand, LL 22 differs from LL 23 by two matrix elements so it would not have the least number of differences from LL 23 to be considered a neighbor. Reversely, however, the two-unit difference between LL 22 and LL 23 is the least difference for any pair involving LL 22 so that this pair forms a conceptual neighbor. Again this property is occurs with the region-region graph as well, where the nearest neighbor for *meet* is *disjoint* (distance 1), while *overlap* (with distance 3) would not qualify as a neighbor. In the reverse direction, however, *overlap*'s nearest neighbors have all distances of three, among them *meet*, so that *meet* and *overlap* are neighbors in the graph.
- If only symmetric least differences are accounted for neighbors, then the conceptual neighborhood graph would be disconnected, preventing comparisons across the separations. For example, relations LL 22 would be isolated in the graph since all its neighbors—LL 23, LL 27, and LL 30—themselves have neighbors with smaller matrix differences.

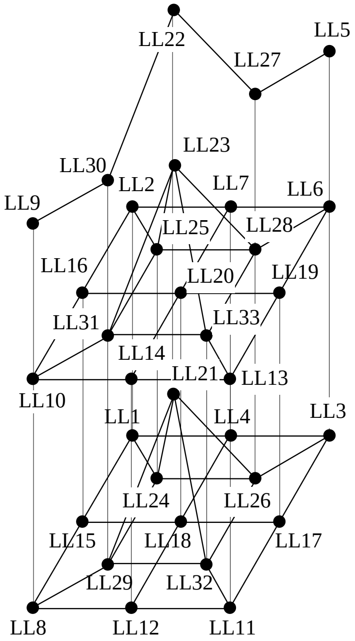


Fig. 4. The conceptual neighborhood graph of the thirty-three topological relations between two lines.

4. Conceptual Neighborhood Graph For Topological Relations Between Two Broad-Boundary Lines

The thirty-three topological line-line relations rely on a crisp representation of the lines. Often, however, this requirement cannot be guaranteed so that more ambiguous line representations should be considered when analyzing their topological relations. This corresponds to the transition of region-region relations from crisp regions to regions with broad boundaries (Clementini and Di Felice 1996; Cohn and Gotts 1996).

Several models are in practice for uncertain lines. *Broad lines* attach uncertainty uniformally around a line (Chrisman 1982), *broad-boundary lines* (Clementini and Di Felice 1997) consider uncertainty only for a line's end points, and *uncertain lines* (Clementini 2005) propagate uncertainly from the boundaries to the interior (Dutton 1992). While the geometric differences of these models may be minor with respect to individual lines, their impact on the possible topological relations is significant: broad lines give rise to five different topological relations (Reis et al. 2006), uncertain lines yield 146 topological relations (Clementini 2005), and broad-boundary lines feature 77 different relations (Reis et al. 2006). We compare the 33 crisp line-line relations with the 77 broad-boundary lines, because their numbers of relations come closest to each other. The broad-boundary relations have been formally derived with the 9-intersection and geometrically verified, which is a process that confirms their existence (Figure 5).

The model for broad-line relations maps onto a subset of the eight region-region relations (Egenhofer and Franzosa 1991) that is obtained by merging partial and complete containments, both with respect to interiors and exteriors, so that the differences between the three pairs of relations of *disjoint–meet*, *covers–contains*, and *coveredBy–inside* are ignored. Therefore, the conceptual neighborhoods can be derived directly from conceptual neighborhood graph of the region-region relations (Egenhofer and Al-Taha 1992). The 146 relations between uncertain lines have been determined computationally, but lack to date a geometric verification.

To determine their conceptual neighborhoods, we employ the snapshot model (Egenhofer and Al-Taha 1992) as well as the smooth-transition model (Egenhofer and Mark 1994). Both methods provided the same pairs of neighbors for the seventy-seven relations. The conceptual neighborhood graph for the broad-boundary relations features two subgraphs—one capturing the 25 relations with empty interior-interior intersections (Figure 6) and another one the 52 relations with non-empty interior-interior intersections (Figure 7).

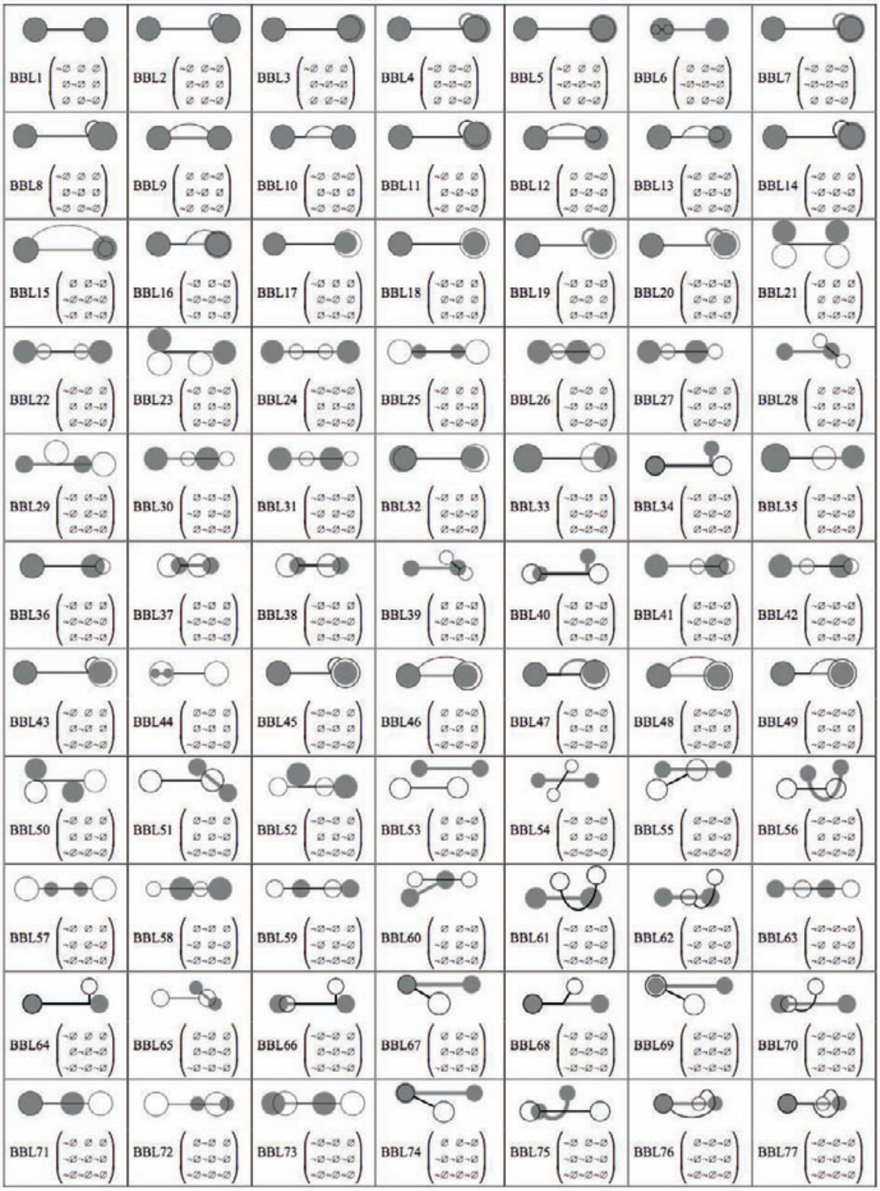


Fig. 5. Seventy-seven topological relations between two broad-boundary lines derived from the 9-intersection (Reis et al. 2006).

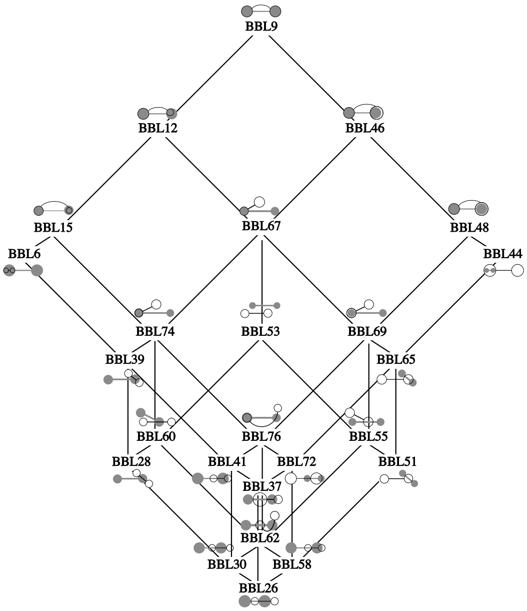


Fig. 6. Conceptual neighborhood graph of broad-boundary lines: layer of relations with empty interior-interior intersections.

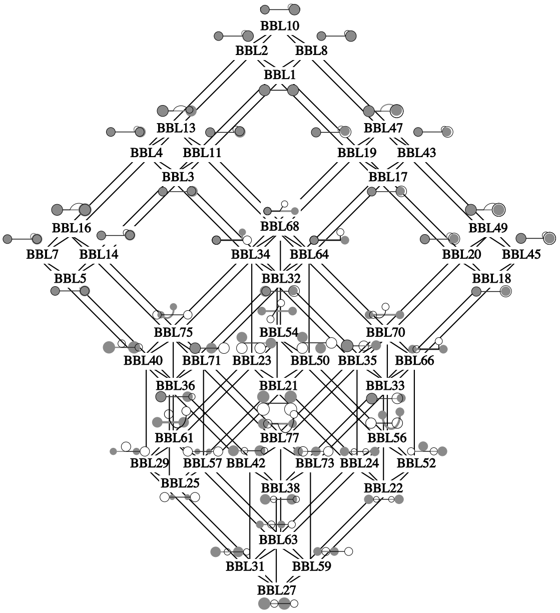


Fig. 7. Conceptual neighborhood graph of broad-boundary lines: layer with non-empty interior-interior intersections.

Both subgraphs have the same overall pattern with two rhombi—a larger one and a smaller one—that are interleaved but do not intersect. The graphs emphasize that 17 of the 77 relations are symmetric (they are all located along the vertical centerlines) and that the remaining 60 relations are made up of 30 pairs of converse relations. Each pair of converse relations is located as a mirror-image along the graph's vertical center line. The 9-intersection matrices of each converse pair are also mirror images of each other, taken along the main diagonal of the matrices. The graphs' structure also highlights fifteen 4-tuples of relations, which form a rhombus based on the four relations' 9-intersections. Within each rhombus the matrices follow the same pattern (Figure 8), that is, seven intersections are fix, while the interior-exterior and exterior-interior intersections cycle through all possible empty and non-empty combinations.

$$\begin{pmatrix} a & b & \neg\varnothing \\ c & d & e \\ \neg\varnothing & f & h \end{pmatrix}$$

$$\begin{pmatrix} a & b & \neg\varnothing \\ c & d & e \\ \varnothing & f & h \end{pmatrix} \qquad\qquad\qquad \begin{pmatrix} a & b & \varnothing \\ c & d & e \\ \neg\varnothing & f & h \end{pmatrix}$$

$$\begin{pmatrix} a & b & \varnothing \\ c & d & e \\ \varnothing & f & h \end{pmatrix}$$

Fig. 8. The repeated pattern of the 9-intersection matrices in each rhombus of the conceptual neighborhood graph of the regions between two lines with broad boundaries

Each of the 25 nodes in the empty-interior-interior subgraph (Figure 6) has a corresponding node in the non-empty-interior-interior subgraph (Figure 7). The 9-intersection matrices of these 25 pairs differ by one unit; therefore, each pair is a neighbor as well and their connections (as vertical links) yield a single, connected conceptual neighborhood graph with the familiar 2-layered structure (Figure 9).

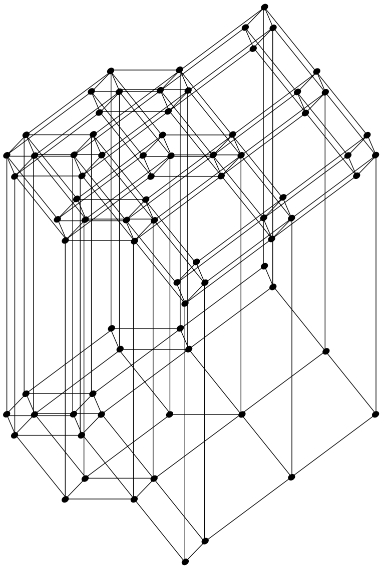


Fig. 9. The conceptual neighborhood graph for the seventy-seven topological relations between two broad-boundary lines.

5. Comparisons of Conceptual Neighborhood Graphs

Beyond the pure visual comparison of these conceptual neighborhood graphs, we offer some quantitative means and apply them to the graphs of two regions (Figure 1b), two lines in R^1 (Figure 1a), two lines in R^2 (Figure 4), and two broad-boundary lines in R^2 (Figure 8).

The *number of nodes* in each graph corresponds to the number of relations for which the graph captures their neighborhood. The *node degree* captures, for each relation, the number of conceptual neighbors. Relations with a node degree of 1 have exactly one conceptual neighbor, relations with a node degree of 2 have two neighbors, and so on. We count in each graph how many relations have a particular node degree. The *node degree sum* is the total node degree of a graph (which is always equal to twice the number of edges in a neighborhood graph). Since graphs with more nodes have a tendency to have a higher node degree, we also normalize the node degree by the number of nodes in a graph, which yields the graph's *normalized node degree*.

Each conceptual neighborhood graph has a lower and upper bound for the node degree. The *lower bound* (Eq. 1a) is the node degree sum of a linear graph that connects all nodes such that there are two end nodes, each with a node degree of 1, and each of the remaining nodes has a node degree of two. The *upper bound* (Eq. 1b) is the node degree of a complete graph (i.e., each pair of nodes in connected by exactly one edge).

The *minimality* (Eq. 1c) and the *saturation* (Eq. 1d) are two measures that compare a graph's node degree to its lower and upper bound, respectively.

$$lowerBound(G) = 2*(\#nodes(G)-2)+2 \tag{1a}$$

$$upperBound(G) = \#nodes(G)*(\#nodes(G)-1) \tag{1b}$$

$$minimality(G) = \frac{nodeDegreeSum(G)}{lowerBound(G)} - 1 \tag{1c}$$

$$saturation(G) = \frac{nodeDegreeSum(G)}{upperBound(G)} \tag{1d}$$

Table 1 summarizes the measures for the four conceptual neighborhood graphs.

Table 1. Summary of quantitative comparison of the conceptual neighborhood graphs between two regions in R^2, two lines in R^1, two lines in R^2, and two broad-boundary lines in R^2

	regions in R^2	lines in R^1	lines in R^2	broad-boundary lines in R^2
node cardinality	8	13	33	77
node degree 1	3	2	0	0
node degree 2	2	4	2	0
node degree 3	3	6	3	3
node degree 4	0	1	24	13
node degree 5	0	0	1	31
node degree 6	0	0	1	24
node degree 7	0	0	0	5
node degree 8	0	0	0	1
node degree sum	16	32	120	403
norm. node degree	2.00	2.46	3.64	5.23
lower bound	14	24	64	152
upper bound	56	156	1,056	5,852
minimality	14%	33%	47%	165%
saturation	28.6%	20.5%	11.4%	6.9%

We establish two baselines for assessing the graphs' node degrees: (1) the region-region relations, (2) the line-line relations in R^1, and (3) the line-line relations in R^2 (Table 2). The first case shows that the graphs' normalized node degrees increase from regions to lines. Case 1 and 2 also show that the graphs' normalized node degrees increase progressively more for broad-boundary lines. The increase in the normalized node degree, however, is less than the increase in the number of nodes. Finally the cross-comparison shows that the transition from lines in R^1 to lines in R^2 has roughly the same impact on the graphs' normalized node degrees as the transition from lines to broad-boundary lines (a 48% increase vs. a 44% increase).

Table 2. Comparisons of the increases in number of nodes and normalized node degree for the conceptual neighborhood graphs of two regions in R^2, two lines in R^1, two lines in R^2, and two broad-boundary lines in R^2

	regions in R^2	lines in R^1	lines in R^2	broad-boundary lines in R^2
nodes	100%	+63%	+313%	+863%
norm. node degree	100%	+23%	+82%	+162%
nodes		100%	+153%	+492%
norm. node degree		100%	+48%	+113%
nodes			100%	+133%
norm. node degree			100%	+44%

7. Conclusions and Future Work

The conceptual neighborhood graphs for topological relations between two crisp lines in R^2 between two lines broad-boundary lines in R^2 have a similar structure due to the parallel occurrence of line relations that share an interior-interior intersections and those that do not. Both graphs are not planar, but still show a highly regular structure. The quantitative comparison showed that the increase in the normalized node degree from crisp to broad-boundary lines is roughly at the same degree as the increase from the mapping of lines from R^1 into R^2.

It is worthwhile to analyze in the future whether a similar behavior can be found for the relations of regions and broad-boundary regions. Beyond the node degree, additional measures on the graphs should be considered to analyze the distribution of the maximum path lengths in the graphs, because the maximum path lengths are used as a measure to normalize

8. Acknowledgments

Max Egenhofer’s work is partially supported by the National Geospatial-Intelligence Agency under grant numbers NMA201-01-1-2003 and NMA401-02-1-2009.

References

Alexandroff P (1961) Elementary concepts of topology. Dover: Mineola, NY

Allen J (1983) Maintaining knowledge about temporal intervals. Communications of the ACM 26(11): 832-843

Bruns HT, Egenhofer M (1996) Similarity of spatial scenes. In: Kraak MJ, Molenaar M (eds), Seventh international symposium on spatial data handling pp 173-184

Chrisman N (1982) A theory of cartographic error and its measurement in digital data bases. Fifth international symposium on computer-assisted cartography. Bethesda, MD, pp 159-68

Clementini E (2005) A model for uncertain lines, *Journal of Visual Languages and Computing* 16: 271-288

Clementini E, Di Felice P (1996) An algebraic model for spatial objects with indeterminate boundaries. In: Burrough P, Frank A (eds) Geographic objects with indeterminate boundaries. Taylor & Francis, London, pp 155-169

Clementini E, Di Felice P (1997) Approximate topological relations. International journal of approximate reasoning 16: 173-204

Cohn A, Gotts N (1996) The ‘egg-yolk’ representation of regions with indeterminate boundaries. In: Burrough P, Frank A (eds) Geographic objects with indeterminate boundaries. Taylor & Francis, London, pp 171-187

Dutton G (1992) Handling positional uncertainty in spatial databases. Spatial data handling symposium. Charleston, SC, vol 2, pp 460-469

Egenhofer M (1993) Definitions of line-line relations for geographic databases. IEEE data engineering bulletin 16(3): 40-45

Egenhofer M (1997) Query processing in Spatial-Query-by-Sketch. Journal of visual languages and computing 8(4): 403-424

Egenhofer M (2005) Spherical topological relations. Journal of data semantics III: 25-49

Egenhofer M (2007) Temporal relations of intervals with a gap, 14th international symposium on temporal representation and reasoning (TIME 2007*)*. Alicante, Spain, IEEE Computer Society, pp 169-174

Egenhofer M, Al-Taha K (1992) Reasoning about gradual changes of topological relationships. In: Frank A, Campari I, Formentini U (eds) Theory and methods of spatio-temporal reasoning in geographic space. Lecture Notes in Computer Science vol 639, pp 196-219

Egenhofer M, Franzosa R (1991) Point-set topological spatial relations, International journal of geographical information systems 5(2): 161-174

Egenhofer M, Herring J (1991) Categorizing binary topological relationships between regions, lines, and points in geographic databases. Technical Report, Department of Surveying Engineering, University of Maine, Orono, ME, (http://www.spatial.maine.edu/~max/9intreport.pdf)

Egenhofer M, Mark D (1995) Modeling conceptual neighborhoods of topological line-region relations, International journal of geographical information systems 9(5): 555-565

Egenhofer M, Sharma J (1993) Topological relations between regions in R^2 and Z^2. In: Abel D, Ooi BC (eds) Advances in spatial databases-third international symposium on large spatial databases, SSD'93. Lecture Notes in Computer Science vol 692, pp 316-336

Freksa C (1992) Temporal reasoning based on semi-intervals. Artificial intelligence 54: 199-227

Hornsby K, Egenhofer M, Hayes P (1999) Modeling cyclic change. In: Chen P, Embley D, Kouloumdjian J, Liddle S, Roddick J (eds) ER workshops. Lecture Notes in Computer Science vol 1727, pp 98-109

Kurata Y, Egenhofer M (2006) The head-body-tail intersection for spatial relations between directed line segments. In: Raubal M, Miller H, Frank A, Goodchild M. (eds) GIScience 2006. Lecture Notes in Computer Science vol 4197, pp 269-286

Kurata Y, Egenhofer M (2007) The 9^+-intersection for topological relations between a directed line segment and a region, In: Gottfried B (ed.) Workshop on behaviour monitoring and interpretation, University Bremen, TZI Technical Report 42, pp 62-76

Mark D, Egenhofer M (1994) Modeling spatial relations between lines and regions: combining formal mathematical models and human subjects testing. Cartography and geographic information systems 21(4): 195-212

Papadias D, Theodoridis Y, Sellis T, Egenhofer M (1995) Topological relations in the world of minimum bounding rectangles: a study with R-Trees, SIGMOD record 24(2): 92-103

Reis R, Egenhofer M, Matos J (2006) Topological relations using two models of uncertainty for lines. In: Caetano M, Painho M (eds), Accuracy 2006 (http://www.spatial-accuracy.org/2006/PDF/Reis2006accuracy.pdf)

Stevens S (1946) On the theory of scales of measurement. Science 103: 677-680.

Schlieder C (1995) Reasoning about ordering. In: Frank A, Kuhn W (eds), COSIT'95, Spatial information theory. Lecture Notes in Computer Science vol 988, pp 341-350

Spatial Support and Spatial Confidence for Spatial Association Rules

Patrick Laube[1], Mark de Berg[2], Marc van Kreveld[3]

[1] Geomatics Department, The University of Melbourne, 3010 Parkville VIC, Australia, plaube@unimelb.edu.au
[2] Department of Mathematics and Computing Science, TU Eindhoven, P.O. Box 513, 5600 MB Eindhoven, The Netherlands, mdberg@win.tue.nl
[3] Department of Computer Science, Utrecht University, P.O. Box 80.089, 3508 TB Utrecht

Abstract

In data mining, the quality of an association rule can be stated by its support and its confidence. This paper investigates support and confidence measures for spatial and spatio-temporal data mining. Using fixed thresholds to determine how many times a rule that uses proximity is satisfied seems too limited. It allows the traditional definitions of support and confidence, but does not allow to make the support stronger if the situation is “really close”, as compared to “fairly close”. We investigate how to define and compute proximity measures for several types of geographic objects—point, linear, areal—and we express whether or not objects are “close” as a score in the range $[0, 1]$. We then use the theory from so-called fuzzy association rules to determine the support and confidence of an association rule. The extension to spatio-temporal rules can be done along the same lines.

Keywords: Spatial data mining, spatial association rule mining, fuzzy association rules, support, confidence.

1 Introduction

Association rule mining (ARM) is one of the defining operations of data mining. The idea is best illustrated by the example of mining frequent item sets in market-basket data (1). The task is finding sets of items that co-occur in user purchases more than a user-defined number of times (6). A classical example of such an association rule is “If a transaction includes bread, then it includes butter.” Of course, an association rule need not always be satisfied to be interesting: even though some transactions including bread do not include butter, it is still interesting to know that the rule holds for many transactions. This leads to the concepts of *support* (the number of transactions including bread and butter) and *confidence* (the fraction of all transactions including bread that also include butter) of association rules. Interesting rules are the ones where both the support and the confidence are high.

In a number of data-mining applications one has to deal with spatial and/or spatio-temporal data: crime hot-spot analysis, optimization of location-based services (LBS), public health and geomarketing applications (6; 22) are important examples. The number of such applications is only growing because of the inexorable fusion of previously separate spatio-temporal data sources. When mining spatial data, spatial association rules are needed. A spatial association rule (sAR) is an association rule where at least one of the predicates is spatial (20). Figure 1 illustrates such a spatial association rule. The rule captures a relation between *location* and *price* for items *houses*: "If a house is close to the river, then it is expensive."

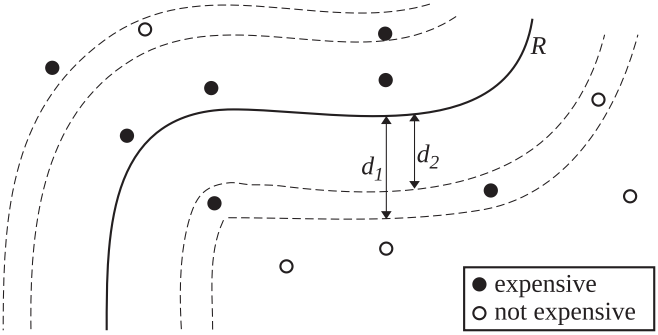


Fig. 1. A set of expensive houses and a set of inexpensive houses, a river R, and two thresholds d_1 and d_2 to express "closeness". The rule "If a house is close to the river, then it is expensive" accounts for a support of 6 and a confidence of $6/8 = 0.75$ for d_1, and a support of 3 and a confidence of $3/3 = 1.0$ for d_2 respectively.

Just as with conventional association rules, the quality of spatial association rules is determined by their support and confidence. When the rule uses proximity, one thus has to decide when objects are deemed "close". This is typically done by a Boolean distance buffer (9; 13): objects are close if their distance is less than some user-defined threshold d. In Figure 1, for example, the given threshold d_1 leads to a support of 6 and a confidence of 0.75. Note that a slightly smaller threshold d_2 would have given a support of 3 and a confidence of 1.0. Clearly, such strong sensitivity to a user-defined threshold is undesirable.

The thresholding problem not only arises for spatial relations. For example, the attribute "expensive" in the rule "If a transaction includes expensive wine, it also includes French cheese." is essentially non-binary. The same holds for attributes like "old", "tall", and so on. Hence, there has been some work on so-called *fuzzy association rules* (10; 3). Here, instead of having a threshold that, for instance, defines whether wine is expensive, the wine price is mapped to a score in the range $[0, 1]$. Then fuzzy-logic theory is used to measure support and confidence of an association rule—see the next section for details.

Given the variety and complexity of proximity relations between spatial entities, it should be clear that simple thresholding is often not the best approach for defining support and confidence in spatial association rules. Unfortunately, it seems that the much more appropriate concept of fuzzy association rules has been largely overlooked in spatial data mining: it is still standard practice to use thresholding when it comes to proximity relations. The main contribution of our paper is to investigate how fuzzy association rules can be applied in a spatial context. In particular, in this paper

- we explore the use of fuzzy association rules to spatial data mining,
- we investigate how to define suitable distance measures between various types of spatial objects—point, linear, areal—, we show how to compute these distance measures efficiently, and we discuss how to map these distance measures to scores in the range $[0, 1]$ for use in fuzzy spatial association rules,
- we extend our ideas to spatio-temporal association rules.

2 Background

2.1 Spatial data mining

Knowledge Discovery in Databases (KDD) is the overall process of discovering useful knowledge from data, and data mining is its most prominent step (5). Data mining is more specifically defined as “the application of specific algorithms for extracting patterns from data” (5, p.39). The need for spatial data mining and geographic knowledge discovery techniques has been widely acknowledged in both the GIScience (13; 17; 18) and data-mining communities (22). Spatial data mining is defined as the process of discovering interesting and previously unknown, but potentially useful patterns from large spatial data sets (22).

The challenges arising when mining spatial data include geographic measurement frameworks (formal and computational representations of geographic information requires the adoption of an implied topological and geometric measurement framework), spatial dependency (tendency of attributes at nearby locations in space to be related), spatial heterogeneity (an intrinsic degree of uniqueness at all geographic locations that makes global comparisons difficult), and the complexity of spatio-temporal objects and rules (whereas objects in non-spatial datasets normally are points in information space, spatial objects often have size, shape, and boundary properties) (13). Important output patterns for spatial data mining are spatial outliers (21; 12; 15), spatial clusters (16; 19), movement patterns (11; 7), spatial co-location patterns (20) and spatial association rules (9).

In conclusion, spatial data mining is in many respects more complex than conventional data mining due to the complexity of spatial data types, spatial relationships, and spatial autocorrelation (22).

2.2 Spatial, spatiotemporal, and fuzzy association rules

Association rule mining has been an active research area ever since the seminal work by (1). Less work has been seen extending association rules into the spatial domain. In an early work, (9) investigated *spatial association rules* (sAR) among a set of spatial and possibly some non-spatial predicates. They present optimisation techniques for association rules with a spatial antecedent and a non-spatial consequent ($is\_a(x, house) \wedge close\_to(x, river) \rightarrow is\_expensive(x)$) and with a non-spatial antecedent and a spatial consequent ($is\_a(x, gasStation) \rightarrow close\_to(x, highway)$). More recently, (6) describe an approach transforming the spatio-temporal rule-mining task to the traditional market-basket analysis task for the improvement of a location-based service application. Both approaches rely on support and confidence measures that are based on counts of pattern frequencies.

Association rules have also been used for the mining of object mobility patterns. (24) and (25) present a definition of *spatio-temporal association rules* (STARs) that specifically describe how objects move between regions over time, motivated by a scenario of mobile phone users moving in a cell-phone network. Although rather specific in their orientation toward mobility patterns between sets of cells, their inclusion of the spatial semantics of the cell sets into their support measure is relevant for our work. As they define support for their STARs in terms of transition counts from one cell to another, and since these cells can be very different in size, they suggest to include the size of the cells into their spatial support measure. Rules expressing transitions between small and restrictive cells are stronger than rules describing transitions between large and inclusive cells (24).

In traditional market-basket analysis, when considering binary association rules $Ant \rightarrow Cons$, each transaction either completely satisfies Ant or it does not satisfy Ant at all, and the same is the case for $Cons$. This can also be applied to *quantitative association rules*, where attribute *intervals* are used (3). An example for such a quantitative association rule is “If an employee is between 35 and 45 years of age, then his/her income is more than $100,000”. Sometimes, however, one wishes to be less precise, and work with rules like “If an employee is middle-aged, then he/she has a high income." To this end, (10) introduced *fuzzy association rules*, where crisp intervals are replaced by fuzzy intervals. In other words, strict membership functions are replaced by fuzzy membership functions giving *scores* in the range $[0, 1]$. Hence, the concepts of support and confidence must be redefined. Now suppose that for a database item x, the score functions $s_{Ant}(x)$ and $s_{Cons}(x)$ determine to what extent x satisfies the antecedent *Ant* and consequent *Cons*, respectively. Then the support and confidence of the rule $Ant \to Cons$ can be expressed using a so-called t-norm (3). This is a function $\otimes : [0, 1] \times [0, 1] \to [0, 1]$ that is commutative, associative, monotone, and has 1 as its identity element (8). The support of $Ant \to Cons$ is now given by

$$\text{support} = \sum_x s_{Ant}(x) \otimes s_{Cons}(x) \tag{1}$$

and the confidence is given by

$$\text{confidence} = \frac{\sum_x s_{Ant}(x) \otimes s_{Cons}(x)}{\sum_x s_{Ant}(x)}. \tag{2}$$

A t-norm that is used often is the minimum t-norm, which takes the minimum of its two arguments; another possibility is the product t-norm, which multiplies its arguments (3). Note that when $s_{Ant}(x)$ and $s_{Cons}(x)$ are either 0 or 1, then $s_{Ant}(x) \otimes s_{Cons}(x) = 1$ when $s_{Ant}(x) = s_{Cons}(x) = 1$ and $s_{Ant}(x) \otimes s_{Cons}(x) = 0$ otherwise. (This holds for both the minimum t-norm and for the product t-norm.) Hence, the definitions for support and confidence given above reduce to the standard definitions for the non-fuzzy case.

2.3 Proximity, nearness, and fuzzy neighbourhoods

An sAR may include spatial predicates such as *close_to*, *adjacent_to*, or *inside*. Most sARM approaches use thresholds such as "within a distance of 80km" when modeling proximity (9). Revolving around Tobler's first law of Geography, claiming that "close things are more likely to be related than distant things" (23), the field of Geographical Information Science developed a wide range more sophisticated models to express and measure proximity relations. See (14) for an introductory text on geographic relationships. (26) discusses distance and proximity relationships between entities in geographic spaces, putting an emphasis on representations that go beyond metric spaces, allowing, for instance, asymmetry in distance relations (e.g. travel time, uphill vs. downhill). In a later piece of work, (27) explores the vagueness of the spatial relation "near". He specifically suggests the extension of sets with broad boundaries as nearness neighbourhoods to fuzzy neighbourhoods with continuous measures of nearness between 0 and 1 . Worboys's work focusses on how people perceive and reason with vague concepts such as nearness, so it is different from our approach in Section 4 where we try to quantify proximity automatically.

Most work on the vague spatial relations *proximity* and *nearness* agree on the context dependency of such concepts. However, if context knowledge can be provided by expert users, concepts such as fuzzy neighbourhoods and the like can be powerful tools for modeling proximity and nearness in a sARM context.

3 Quality measures for spatial association rules

In many non-spatial applications the numerical values of the attributes are already given and it is rather straightforward to map these values to scores in the range $[0, 1]$. In the spatial domain, however, this is not the case. In this paper we discuss the application of concepts from fuzzy-association-rule theory to spatial association rules, in particular to rules involving proximity.

3.1 Scoring proximity

In section 2.2 we have seen that in order to apply the theory of fuzzy ARM, we need to define score functions in the range [0,1] for the antecedent and for the consequent of the association rule. The task of defining score functions for proximity can be split into two subtasks. Consider as an example the condition “close to the river”. Then the first task is to determine how to quantify distance to the river, and the second task is to convert this distance to a score in the range $[0, 1]$. Suppose for the moment that we have a suitable distance function $dist$. Traditionally, a user would define a threshold parameter d and one would say that an object o is close to a river R if $dist(o, R) \leq d$. Instead, we use two parameters, d_c and d_f, with $d_c < d_f$, and define

$$score(o) = \begin{cases} 1 & \text{if } dist(o, R) \leq d_c \\ \frac{d_f - dist(o,R)}{d_f - d_c} & \text{if } d_c < dist(o, R) \leq d_f \\ 0 & \text{if } dist(o, R) > d_f \end{cases} \tag{3}$$

Typically, the value of d_c would be somewhat smaller than the strict threshold d would be, while d_f would be somewhat larger. Figure 2 illustrates this definition.

The best values for the parameters d_c and d_f depend on the application, and can be determined by the user. Note however, that even though we now require two parameters rather than one, the score function is much less sensitive

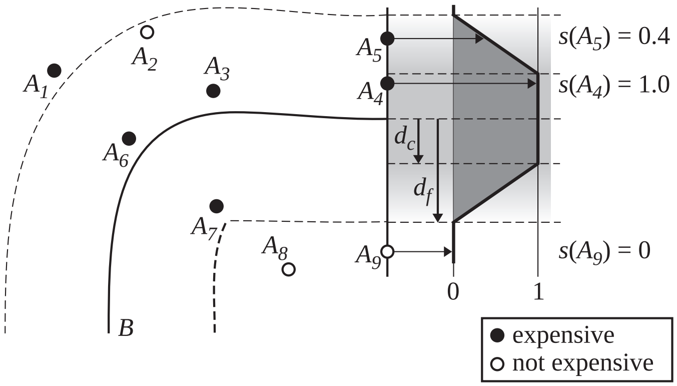


Fig. 2. A linear score function allows the computation of spatial support for the antecedent "A is close to B". The score function rises from 0 at distance d_f to 1 at distance d_c. Example point A_9 provides support 0, A_5 accounts for support 0.4, and A_4 for 1.0. Note that both A_4 and A_3 provide the same support 1.0 even if A_3 is closer to B than A_4.

to the exact value of the parameters than was the case for simple thresholding (see Figure 1). Indeed, the score function is continuous in the parameters d_c and d_f, so changing their values slightly will result in only a slight change in the score function.

Above we have used a piecewise linear score function. When so desired, one could also use a smooth Gaussian-like score function. Furthermore, the score function can easily be adapted to capture other proximity concepts such as "far". All that we need is a suitable distance function *dist*. In the Section 4 we investigate how one can define distance in various different settings.

3.2 Spatial support and spatial confidence

First we discuss how to define support and confidence in the simplest case, namely where both the antecedent and the consequent consist of a single predicate. Recall from Section 2.2 that in order to define support and confidence, we need to choose a suitable t-norm. Although the minimum t-norm is often used, we believe that in spatial association rules the product t-norm is more appropriate. As an example, consider the rule "If a house is close to the sea, then it is expensive." Suppose we have a score $s_{Ant}(H)$ in the range $[0, 1]$ that captures to what extent a house H is close to the sea, how to obtain such a score is discussed in the next section. We also map the price of H to a score $s_{Cons}(H)$ in the range $[0, 1]$ to capture to what extent a house is expensive. For example, houses that cost \$1,000,000 or more get a score of 1, houses that cost less than \$500,000 get a score of 0, and in between we interpolate the score. Using the product t-norm in the definition of support then gives us the spatial support

$$\text{spatial support} = \sum_{H} s_{Ant}(H) \times s_{Cons}(H), \tag{4}$$

where the sum is over all houses H. Similarly, spatial confidence is defined as

$$\text{spatial confidence} = \frac{\sum_H s_{Ant}(H) \times s_{Cons}(H)}{\sum_H s_{Ant}(H)}. \tag{5}$$

We mentioned that in a spatial context the product t-norm is usually the most appropriate. Indeed, in the example above, houses that are rather close to the sea (say, score 0.7) and very expensive (score 0.9) should give more support than houses that are rather close (score 0.7) to the sea and somewhat expensive (score 0.7), as is given with the product t-norm ($0.7 * 0.9 = 0.63 > 0.7 * 0.7 = 0.49$).

Note that the antecedent and/or the consequent of a sAR can be composed of several predicates. In the next subsections we discuss two examples of this.

A score for two antecedents combined by AND

Consider the sAR "If a house is close to the sea and close to a big city, then it costs at least $800,000.", and suppose we have a score for "close to the sea" and a score for "close to a big city". From these two scores we must then compute an overall score for the antecedent. For this we need another t-norm. We believe that also here the product t-norm is the most appropriate. Indeed, if an expensive house is very close to the sea (score 1) and somewhat close to a big city (say, score 0.5) then it should support the rule somewhat, while if an expensive house is somewhat close to the sea (score 0.5) and somewhat close to a big city (score 0.5), then it should support the rule less. We get this behavior by multiplying the scores: in the first case we then have a score for the antecedent of 0.5, and in the second case we have a score of 0.25. Note that in standard fuzzy logic, the AND-operator gives the minimum (instead of the product) of the two scores. The semantics of the product appears more suitable in our case than the semantics of the minimum, as we argued briefly.

A score for two antecedents combined by OR

Consider the sAR: "If a house is close to an airport or close to a highway, then it has good sound insulation." If a well-insulated house is already very close to a highway, then proximity to an airport is irrelevant for the support of the rule, which simply should be 1, no matter how close or far any airport is. But if the house is somewhat close to a highway and somewhat close to an airport (both with score 0.5), then it supports the rule more than when only one of these antecedents was present. A t-conorm—this is similar to a t-norm, except that it has 0 as identity element—that nicely captures this behavior is the Einstein sum, defined as $\frac{s_1+s_2}{1+s_1 s_2}$ for two scores s_1 and s_2.

4 Proximity measures

In this section we discuss how to quantify proximity or, in other words, how to define distance measures for various types of geographic objects. Since spatial association rules have most use in two-dimensional space, we will limit ourselves to this case; the ideas, however, easily extend to three-dimensional objects.

In the object view, one generally distinguishes three types of geographic objects in two-dimensional space: zero-dimensional objects (points), one-dimensional objects (linear objects, typically polylines), and two-dimensional objects (areas, typically polygons). When measuring the distance between such objects A and B, we thus have a number of different cases: point-to-point, point-to-polyline, and so on. (Note that, depending on the application, the point-to-polyline case need not be the same as the polyline-to-point case.) The next three subsections discuss the point-to-point, point-to-polyline, and point-to-polygon cases, respectively; the last subsection then comments on some other cases.

4.1 The point-to-point case

We begin with the case of point-to-point proximity. As an example, consider the rule "If a street-crime incident is close to an ATM, then it is a pickpocket case." Here both the location of the incident and the location of the ATM are points. Note, however, that the rule speaks of *an* ATM. Thus we are not measuring the distance to a specific ATM, but to any ATM. In more abstract terms, the point-to-point situation is often as follows. We are given a point A and a set S_B of points, and we are interested in $\mathit{dist}(A, S_B)$, the distance from A to the set S_B. In our example, A would be a street-crime incident and S_B would consist of all ATM locations. In this example, a natural interpretation of the rule is to consider the distance from the incident to the nearest ATM. This corresponds to setting

$$\mathit{dist}(A, S_B) = \min_{B \in S_B} d(A, B),$$

where $d(A, B)$ denotes the distance between the points A and B. The distance $d(A, B)$ can be the Euclidean distance, the travel time on the road, or any other distance measure between individual points, but in any case $\mathit{dist}(A, S_B)$ would be defined by the nearest point $B \in S_B$—see Figure 3.

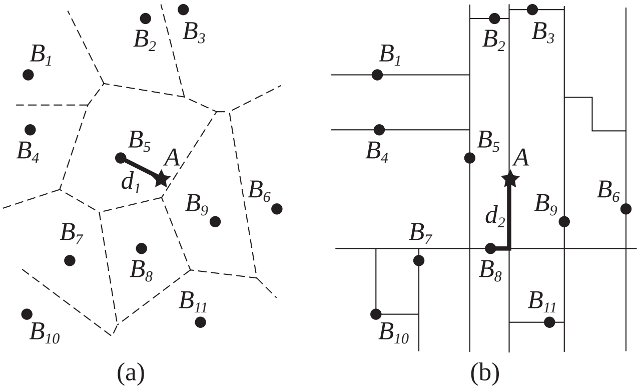


Fig. 3. Score function for the point-to-point case. (a) The Voronoi diagram is used to define the nearest ATM location B_5 to a crime spot A. (b) If travel time is assessed on a street network a different ATM is closest, B_8. The Euclidean distance d_1 and the Manhattan distance d_2 are used when computing distance $d(A, B)$.

If S_A is a set of n points, S_B is a set of m points, and $d(A, B)$ denotes the Euclidean distance, then we can compute $dist(A, S_B)$ for each $A \in S_A$ in $O(m \log m + n \log m)$ time in total. To this end we compute the Voronoi diagram of S_B, preprocess it for efficient planar point location, and then perform a query with each point $A \in S_A$ to find its closest point in S_B. The first two steps take $O(m \log m)$ time, and each of the point-location queries takes $O(\log m)$ time (2). The network version—here the points from S_A and S_B lie on a network with E edges, and distances are measured along the network—can be solved in $O((n + m) \log(n + m) + E)$ (4).

Sometimes not only the closest point of S_B matters. For example, consider the rule: "If a hotel is close to castles, then it is used mostly by tourists." In this case, not just the nearest castle matters, but also other nearby castles and their proximity: the more castles in the vicinity, the more the antecedent of the rule applies. One possibility to handle this is as follows. First, for two points A and B—a hotel and a castle—we define $w(A, B)$ to be a value in the range $[0, 1]$ that expresses to closeness of A and B. We do this in such a way that $w(A, B) = 0$ if A and B are really close to each other, $w(A, B) = 1$ if A and B are really far from each other, and $w(A, B)$ increases linearly in between. (Note the similarity to the way we mapped distance to score in the previous section.) Then we define

$$dist(A, S_B) = \sum_{B \in S_B} w(A, B).$$

Observe that $\mathit{dist}(A, S_B)$ can range from 0 (if all castles in S_B are really close to A) to $|S_B|$ (if all castles are really far from A); this distance[1] will be mapped to a score in the range $[0, 1]$, as usual. Computing $\mathit{dist}(A, S_B)$ for each $A \in S_A$ now takes $O(nm)$ time in total, since the distance to all other points in S_B can play a role. However, one would expect that only few points B are relevant—that is, have $w(A, B) > 0$—for a given A. Hence, one can speed up the computation in practice by using spatial index structures such as R-trees to quickly find for each $A \in S_A$ the relevant points $B \in S_B$.

4.2 The point-to-polyline case

Next we discuss how to measure the distance between a point and a single polyline. As in the point-to-point case, we may want to extend this definition to the case of multiple polylines. This can be done by considering the closest polyline (as in the street-crime/ATM example) or in some more involved manner (as in the hotel/castles example). For the point-to-polyline case, however, these considerations already play a role when considering the distance to a single polyline. Consider for example the rule "If a house is close to the river, then the occupants own a boat." Here what matters could be how quickly the occupants can reach the river. For a point A and polyline B—the house and the river—we would then consider the minimum distance from A to any point on B:

$$\mathit{dist}(A, B) = \min_{p \in B} d(A, p),$$

where $d(A, p)$ denotes (for instance) Euclidean distance. But now consider a house and its proximity to a highway, for the purpose of studying problems due to noise: "If a house is close to the highway, ...". Clearly, a house that is within 500m from the highway over a stretch of 1.2km of that highway, suffers more noise pollution than a house that is within 500m over a stretch of 0.7km—see Figure 4. Hence, minimum distance does not seem an appropriate distance measure in this example. One possible solution is to proceed similarly to the hotel/castles example. Thus we first define for points $p \in B$ a function $w(A, p)$ that is 0 (resp. 1) if A and p are really close to (resp. far from) each other, and that increases linearly in between, and we define

$$\mathit{dist}(A, B) = \int_0^1 w(A, p(x))dx$$

Now $\mathit{dist}(A, B)$ can vary between 0 (if the entire highway is very close to A) to $\mathit{length}(B)$ if the entire highway is far from A.

[1] Note that adding another point to S_B that is far away from A will increase $\mathit{dist}(A, S_B)$. If this is undesirable, one may reverse the definition of $w(A, B)$ so that $w(A, B) = 1$ for points that are close to (instead of far from) A. This also means one should change the mapping from distance to score, as larger "distance" now implies more castles that are closer, which should lead to a higher score.

Let S_A be a set of n points whose proximity to a polyline B consisting of m line segments is needed. For a score based on the minimum distance only, similar to the point-to-point case, we use Voronoi diagrams of line segments to compute $dist(A, B)$ for all $A \in S_A$ in $O(m \log m + n \log m)$ time in total. If we use the definition using the integral, then we cannot use Voronoi diagrams to reduce the running time and we will need $O(nm)$ time in the worst case. As before, index structures like R-trees can be used to speed up the efficiency in practice.

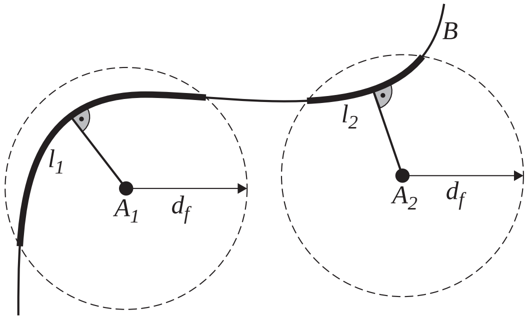


Fig. 4. The point-to-polyline case for "if a house is close to the highway". Even though in terms of absolute distance A_1 and A_2 are equally close to B, when assessing the proximity with respect to the stretch l_x covered by a disc of radius d_f, then A_1 is closer to B than A_2.

4.3 The point-to-polygon case

Also in the point-to-polygon proximity case, there are several ways to quantify proximity. Depending on the application, we can use the shortest distance from the point to the polygon, the total length of the boundary of the polygon within a certain distance from the point, or the total area of the polygon within a certain distance from the point—see Figure 5. The shortest distance could be appropriate when scoring access to a water in an agricultural property evaluation. Here the only fact that matters might be the shortest distance, as this is directly linked to development costs. When evaluating potential sites for a new beach-side hotel, the actual length of the beach stretch within some walking distance might be more important than the area of the water body itself. Finally, areas could be important in a bird ecology study evaluating potential nesting sites. In Figure 5, A_1 and A_2 could be nesting sites for birds that need access to the forest B for food supply. Even though A_1 is in absolute distance further away from forest B, it has much more forest within distance d_f than A_2 and, hence, may be the better nesting site.

Note that the definition of the distance function $dist(A, B)$ for a point A and polygon B in the latter case (the nesting example) is very similar to the definition in the highway example in the previous subsection. Hence, we can define the distance in a similar way, using a (double) integral over the polygon B.

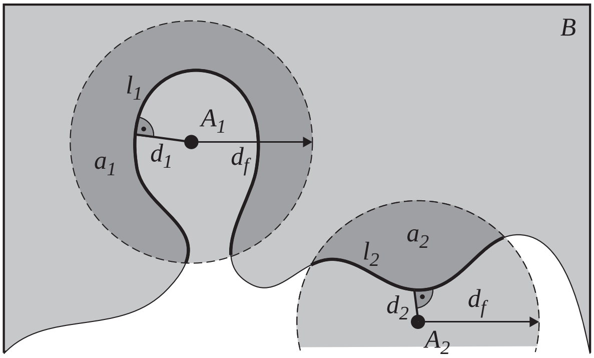


Fig. 5. The point-to-polygon case. The proximity of point A with respect to area B is assessed by several measures derived from a circular template around A. In the most simple case minimal distances d_1 and d_2 are used. Alternatively, the length of the boundary cut out by the circular template might define the proximity, with l_1 outperforming l_2. Finally, if the covered area matters, proximity depends on the areas a_1 and a_2.

4.4 Other cases

There are several more cases, namely those where A is of a linear or areal type. Examples are: "If a forest is close to highways, then it contains only few deer", and: "If an urban development area is close to existing urban areas, then the houses will have small gardens".

First consider the case where we want to quantify the distance from a polygon to a set of points: "If a lake is close to dump sites, then its water is polluted." For this rule, we want the polygon-to-point proximity to take all nearby dump sites and their distances into account. Thus we should define distance similar to some examples given before. Note that this way we already defined a score in the range [0,1] to capture proximity, so the machinery of section 3.1 to convert distance to score is not necessary. The same is true for the examples discussed next.

Now let's consider polygon-to-polygon proximity. Observe first that the polygon types can be such that overlap is not possible ("If a lake is close to forests, ...") or such that overlap is possible ("If a forest is close to municipalities with many elderly people, ..."). These cases give rise to different score functions. If overlap is not possible, then the simplest definition of the score is based on the minimum distance and is like the point-to-point case, see Figure 6.

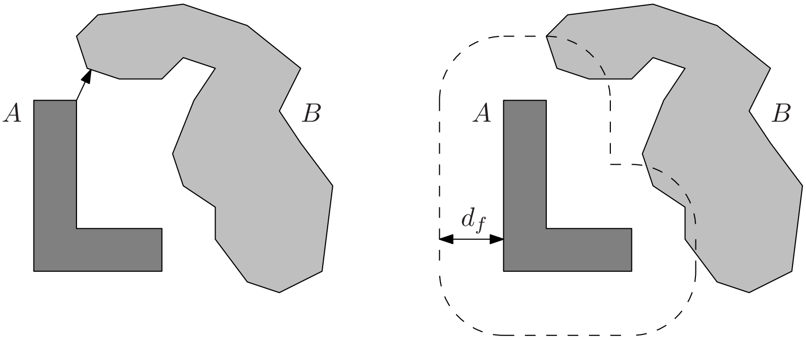


Fig. 6. Two simple ways to define polygon-polygon proximity.

Another simple definition is based on the area of B inside a buffer of width d_f around A, see Figure 6. For example, if R denotes the buffer region, then one could say that if more than, say, 30% of R is covered by B, then the score is 1. If 0% of R is covered, then the score is 0. In between these percentages, we can let the score depend linearly on the percentage covered, similar to some examples given before.

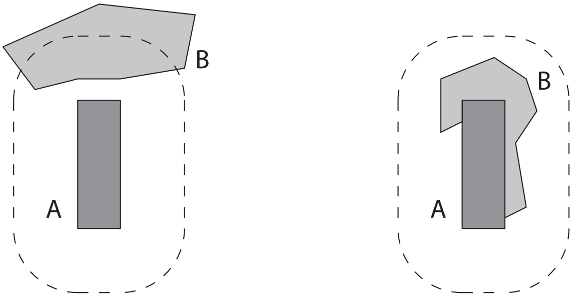


Fig. 7. Discriminating two cases where 15% of the area of the buffer of A is covered by B.

Note that such a definition may sometimes be too limited. It does not discriminate on the distance from A, and also not on which parts of A are actually close to B. Consider Figure 7. In both cases, about 15% of the area of the buffer of A is covered by B, but in the right case, A appears closer to B, because for every point in A, the polygon B is closer than in the left case.

We can refine our definition by using a weighted buffer and taking integrals, similar to what we did before. For example, define the weight $w(p)$ of a point p to be 0 if the distance from p to A is at least d_f, and let its weight be $(d_f - dist(p, A))/d_f$ otherwise. We can now define the score depending on

$$\iint_B w(p)\, dxdy.$$

We can divide this value by the integral of the weight in the whole buffer region of A, and if this is more than, say, 0.3, let the score be 1. The score can decrease linearly in the outcome of the division.

5 Spatio-temporal support and confidence

This section investigates the extension of our fuzzy spatial association rules into the temporal domain. We start our consideration of spatio-temporal association rules by extending the basic point-to-point case with a temporal antecedent T: "If A is close to B, around time T, then *Cons*". This rule has two antecedents, namely one capturing spatial proximity, "A is close to B", and one assessing temporal proximity, "A is present around time T". An example could be "If mobile phone users are close to the city centre at noon, then they are businessmen". Note that combining scores for the two antecedents with the product t-norm appears appropriate.

If we perceive the time as a one-dimensional space, we can model temporal proximity similar to spatial proximity. Time stamps of events and episodes may be located on a time axis, allowing the assessment of various temporal relationships. Just as with spatial proximity, also temporal relations may be extended beyond discrete classes. Predicates such as "at noon", "in the morning", or "at night" may be assessed using score functions allowing for a refined concept of temporal proximity. Unlike spatial relations, temporal relations are in general directed. A relation such as "after sunset" is directed and this direction has to be included when modeling temporal proximity.

Figure 8 illustrates examples for a score function for temporal proximity for a linear and a cyclical time axis. The first example refers to the "baby boomer" generation. In order to qualify as a baby boomer, a person has to be born clearly after e_1, the end of World War II. Whereas births up until 1965 generally qualify for the baby-boomer generation (score 1), there might be a transition period that can be modeled with a directed score function. By contrast, when referring to the decade "the Seventies" one might want to also include funny hair cuts and bell-bottoms found in 1969 and 1981. Hence, the use of an extended bi-directed temporal proximity score function that leaps beyond the crisp decade e_2 might be appropriate.

Similar concepts can be applied for temporal proximity on a cyclic time scale. Whereas the event e_3 "sunset" is crisply defined, there may be leeway for the temporal predicate "after sunset". Finally, for many spatio-temporal association rule mining applications refering to the episode e_4 "in the morning", an extended temporal neighborhood function may be more adequate than a crisp interval ranging from 6am to 12pm.

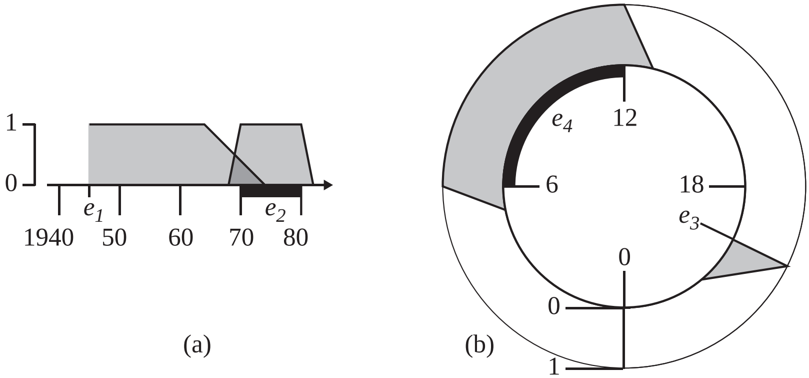


Fig. 8. Temporal proximity can be modeled with uni-directional (events e_1 and e_3) or bi-directional (episodes e_2 and e_4) score functions on both linear (a) and cyclic (b) time axes.

6 Discussion

Our exploratoy study on quality measures for spatial association rules illustrates that it is imperative to include spatial semantics when adopting association rules for assessing the complex relationships between spatial objects. This finding complies with similar conclusions from work on spatio-temporal ARs in a mobility context (24; 25). As we have argued and illustrated when defining spatial support and spatial confidence, the spatial context of specific sAR applications makes the selection of some t-norms for score combinations more suitable over others. Similarly, the semantics of the variability of interrelations amongst spatial objects leads to a wide range of sensible proximity measures.

Allowing score functions for both antecedents and consequents in various conjunctions, requires the specification of rules on how to combine separate scores. The presented t-norms and t-conorms offer a suitable framework for such guidelines.

If one wants to use fuzzy association rules in a spatial context, one has to quantify proximity. We have shown how this can be done for a number of different cases, involving proximity between features of various types (point, linear, areal). Often spatial association-rule mining is only a first step in analyzing spatial data. This step gives a number of association rules, which can then be subjected to further investigation using e.g. statistical methods. Note that these statistical methods would also need to quantify proximity. Hence, our work on proximity measures not only has applications in spatial association-rule mining, but it can also be useful in statistical methods for spatial data analysis (or, in fact, any type of quantitative analysis involving spatial data).

Being one-dimensional in nature, the temporal dimension *per se* adds little complexity to our proximity discussion. However, taken in combination with almost arbitrarily complex score functions for spatial proximity, the suggested temporal proximity score functions sum up to a powerful tool when assessing spatio-temporal association rules.

7 Conclusions

This paper contributes to the theory of spatial data mining by refining quality measures for spatial association rules. We present a conceptual framework for *spatial support* and *spatial confidence* applying concepts from the theory of fuzzy association-rule mining. The major contribution of this paper is twofold: First, we introduce fuzzy association rule quality measures into the spatial domain. Second, we explore various possibilities for defining and efficiently computing suitable proximity measures amongst objects in space-time. With a series of illustrative examples we have shown that developing spatial and spatio-temporal quality measures for association rules presents a set of interesting proximity problems, ranging from simple point-to-point constellations to rather complex polygon-to-polygon scenarios.

Acknowledgments

This research was initiated during the GADGET Workshop on Geometric Algorithms and Spatial Data Mining, funded by the Netherlands Organisation for Scientific Research (NWO) under BRICKS/FOCUS grant number 642.065.503. Patrick Laube's work is funded by the Australian Research Council (ARC), Discovery grant DP0662906. Mark de Berg was supported by the Netherlands' Organisation for Scientific Research (NWO) under project no. 639.023.301. Finally, the authors wish to thank three anonymous reviewers for their valuable comments.

References

[1] R. Agrawal, T. Imieliski, and A Swami. Mining association rules between sets of items in large databases. In *SIGMOD93*. ACM, 1993.

[2] M. de Berg, O. Cheong, M. van Kreveld, and M. Overmars. *Computational Geometry: Algorithms and Applications*. Springer-Verlag, Berlin, 3nd edition, 2008.

[3] Didier Dubois, Eyke Hüllermeier, and Henri Prade. A systematic approach to the assessment of fuzzy association rules. *Data Min. Knowl. Discov.*, 13(2):167–192, 2006.

[4] M. Erwig. The graph Voronoi diagram with applications. *Networks*, 36(3):156–163, 2000.

[5] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54, 1996.

[6] G. Gidofalvi and T. B. Pedersen. Spatio-temporal rule mining: Issues and techniques. In *Data Warehousing and Knowledge Discovery, Proceedings*, volume 3589 of *Lecture Notes in Computer Science*, pages 275–284. Springer-Verlag, Berlin, 2005.

[7] J. Gudmundsson, M. van Kreveld, and B. Speckmann. Efficient detection of patterns in 2D trajectories of moving points. *GeoInformatica*, 11(2):195–215, 2007.

[8] P. Hájek. *Metamathematics of Fuzzy Logic*. Kluwer, 1998.

[9] K. Koperski and J. Han. *Discovery of Spatial Association Rules in Geographic Information Databases*. Proceedings of the 4th International Symposium on Advances in Spatial Databases. Springer-Verlag, 1995.

[10] C.M. Kuok, A.W.-C. Fu, and M.H Wong. Mining fuzzy association rules in databases. *SIGMOD Record*, 27:41–46, 1998.

[11] P. Laube, S. Imfeld, and R. Weibel. Discovering relative motion patterns in groups of moving point objects. *International Journal of Geographical Information Science*, 19(6):639–668, 2005.

[12] C.-T. Lu, D. Chen, and Y. Kou. Detecting spatial outliers with multiple attributes. In *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence 2003 (ICTAI'04)*, pages 122–128, 2003.

[13] H. J. Miller and J. Han. Geographic data mining and knowledge discovery: An overview. In H. J. Miller and J. Han, editors, *Geographic data mining and knowledge discovery*, pages 3–32. Taylor and Francis, London, UK, 2001.

[14] H. J. Miller and E. A. Wentz. Representation and spatial analysis in geographic information systems. *Annals of the Association of American Geographers*, 93(3):574–594, 2003.

[15] R. T. Ng. Detecting outliers from large datasets. In H. J. Miller and J. Han, editors, *Geographic data mining and knowledge discovery*, pages 218–235. Taylor and Francis, London, UK, 2001.

[16] D. O'Sullivan and D. J. Unwin. *Geographic Information Analysis*. John Wiley and Sons, Hoboken, NJ, 2003.

[17] J. F. Roddick, K. Hornsby, and M. Spiliopoulou. An updated bibliography of temporal, spatial, and spatio-temporal data mining research. In J. F. Roddick and K. Hornsby, editors, *Temporal, spatial and spatio-temporal data mining, TSDM 2000*, volume 2007 of *Lecture Notes in Artificial Intelligence*, pages 147–163. Springer, Berlin Heidelberg, DE, 2001.

[18] J. F. Roddick and B. G Lees. Paradigms for spatial and spatio-temporal data mining. In H. J. Miller and J. Han, editors, *Geographic data mining and knowledge discovery*, pages 33–49. Taylor and Francis, London, UK, 2001.

[19] Y. Sadahiro. Cluster detection in uncertain point distributions: a comparison of four methods. *Computers, Environment and Urban Systems*, 27(1):33–52, 2003.

[20] S. Shekhar and Y. Huang. Discovering spatial co-location patterns: A summary of results. In *Advances in Spatial and Temporal Databases, Proceedings*, volume 2121 of *Lecture Notes in Computer Science*, pages 236–256. Springer-Verlag, Berlin, 2001.

[21] S. Shekhar, C. T. Lu, and P. S. Zhang. A unified approach to detecting spatial outliers. *Geoinformatica*, 7(2):139–166, 2003.

[22] S. Shekhar, P. Zhang, Y. Huang, and R. R. Vatsavai. Trends in spatial data mining. In H. Kargupta, A. Joshi, K. Sivakumar, and Y. Yesha, editors, *Data Mining: Next Generation Challenges and Future Directions.* MIT/AAAI Press, 2003.

[23] W. R. Tobler. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(2):234–240, 1970.

[24] F. Verhein and S. Chawla. Mining spatio-temporal association rules, sources, sinks, stationary regions and thoroughfares in object mobility databases. In *Database Systems for Advanced Applications*, pages 187–201. 2006.

[25] F. Verhein and S. Chawla. Mining spatio-temporal patterns in object mobility databases. *Data Mining and Knowledge Discovery*, 16(1):5–38, 2008.

[26] M. F. Worboys. Metrics and topologies for geographic space. In M. J. Kraak and M. Molenaar, editors, *Advances in Geographic Information Systems Research II: Proceedings of the International Symposium on Spatial Data Handling, Delft*, pages 365–376, London, UK, 1996. Taylor & Francis.

[27] M. F. Worboys. Nearness relations in environmental space. *International Journal of Geographical Information Science*, 15(7):633–651, 2001.

A Primer of Picture-Aided Navigation in Mobile Systems

Robert Laurini[1], Silvia Gordillo[2], Françoise Raffort[1], Sylvie Servigne[1], Gustavo Rossi[2], Nan Wang[1], Andrés Fortier[2]

1 LIRIS, INSA de Lyon, France
email: {Robert.Laurini, Francoise.Raffort, Sylvie.Servigne, Nan.Wang}@insa-lyon.fr,
2 LIFIA, Universidad Nacional de La Plata, Buenos Aires, Argentina
email: {Gordillo, Gustavo, Andres}@lifia.info.unlp.edu.ar

Abstract

The goal of this paper is to present a new concept regarding the way of explaining itineraries based on pictures in mobile systems. Instead of presenting a map, or a list of words, an original method is proposed based on a picture database and handheld devices. Knowing the exact position of the user, the system will sent him regularly pictures of the way where to go. This system is especially targeted not for drivers but for pedestrians essentially in tourist cities. In order to reach this objective, pictures were taken along each street both ways. Thus, a picture database was built for which the more common query is to compute the minimum path for going from one place to another. The result format is initially presented as a sequence of nodes and edges; then this sequence is transformed into a sequence of pictures. Those pictures are then sent to the user according to his position and his pace. In order to help the user, the pictures are decorated with arrows.This presentation is based on two prototypes made within a French-Argentinean collaboration.

Keywords: GIS, images, explaining itineraries, Location-based Services, Physical Hypermedia

1. Introduction

Suppose you are lost in a Chinese city and you want to go to a precise place: textual or oral descriptions are not very efficient, and similarly a map is not of interest because you cannot read the names of the streets. More generally, there exist several modes of describing itineraries, each of them having their own advantages and drawbacks. The objective of this paper is to propose a new mode of describing itineraries, i.e. based on sequences of pictures.

If somebody gives you, or a computer generates a text explaining how to go from one place to another, you need first to understand the text, and second to build your own visual representation of the space in your mind in order to follow the explanations.

In the case of a vocal description, it is common to forget what has been said, especially when the itinerary is complex. In the case of maps, you can follow the itinerary easily, even if you are lost; however a lot of people have difficulties in reading maps.

We consider that giving an itinerary by a sequence of pictures can be a powerful mean to help pedestrians go from one place to another. Our vision is that for tourists or pedestrians equipped with a handheld device (PDA, smart phones, etc.) such a description mode can be useful equally outdoor and indoor provided that a lot of pictures have been taken, and that the pedestrian is correctly positioned with a mobile system, thus creating a sort of picture-aided navigational system (PANS). To do so, we assume the existence of a communication infrastructure covering the whole territory, for instance a tourist city, both indoor and outdoor. In other words, the initial goal of such a system is not oriented to car-drivers for long itineraries, but rather for short itineraries in a city.

In addition, let us mention that such a system can be used as a component for several types of Location-Based Services [Küpper, 2005], [Schiller-Voisard, 2004] or physical hypermedia [Rossi et al., 2006].

The considerations presented in this paper are based on two early prototypes made within Franco-Argentinean cooperation [Wang, 2007].

The paper is organized as follows. After a comparison of the various modes of describing itineraries, a comparison table based on several criteria will emphasize the advantages of picture-based itinerary descriptions. Then the main characteristics of picture-aided navigational systems will be detailed, essentially based of an ordered sequence of arrow-decorated pictures. We will conclude the paper by giving some final remarks and presenting future works.

2. Different modes for describing itineraries

Presently there exist several modes of describing itineraries: text-based, voice-based, map-based and picture-based. Immediately, we must notice that sometimes such modes can be used separately or in combination. In the next sections we will detail each of them.

2.1 Text-based mode

Description. In this mode a written text describes how to reach the desired place and it is up to the user to match the written indications with the existing objects in the real world. In an automatic system, the text is usually generated once at the preparation of the itinerary.

Advantages. The user can keep the written indications in his mobile device and refer to it when he finds it necessary.

Disadvantages. One of the main drawbacks of this mode arises when the user is lost. In this case, the user must ask somebody else in the street who usually gives him a vocal description to return the planned itinerary, which cannot be recorded as text. Another big problem arises when the user has to match the textual description with the real objects: the road names or numbers may not be written at each cross-road or they may be hidden by other obstacles, like trees or traffic lights. Finally, if the application does not support internationalization or the user's language it becomes useless.

2.2 Voice-based mode

Description. Voice-based description can be considered as the older system, and it is still very used.

Advantages. In an automatic system, usually the spoken commands are not given initially, but during the navigation when necessary.

Disadvantages. The main problem is that when the itinerary is complex, the user can forget or mix the indications. The classical consequence is the necessity of completing the indications by asking somebody else. Sometimes the user is not persuaded to be on the planned itinerary. Another major drawback is when the user does not understand the used spoken language.

2.3 Map-based mode

Description. With a map description, the itinerary is usually presented as a moving line showing the roads to follow.

Advantages. In automatic systems, global maps are generated initially; but only a small moving map piece is presented to the user when navigating, reduced to his nearby future, usually toward the successive node. When the user is lost, but knowing exactly where he is in the map, he can reconstruct himself the itinerary to reach the planned route.

Disadvantages. One of the big drawbacks is that many persons have difficulties in reading maps and identifying exactly where they are on a map; and especially when they are lost.

Maps are essentially 2D representations. However, some systems offer the possibility of a perspective angle, so giving a sort of 3D feeling; but these are not truly 3D systems, because in our understanding, 3D visualization will imply the storing of all existing buildings with a 3D representation.

2.4 Picture-based mode

Description. Consider that practically all landmarks have not their name written in huge letters; take for instance a mountain. With only the name you can have some difficulties of identification, but not with pictures.

Advantages. Pictures are independent from languages and can help easily identify places. They can be used equally outdoor and indoor.

Disadvantages. The big drawback is that a database of pictures must be created and regularly updated. Another aspect is that such a system cannot be used in the case of dense fogs.

We claim that most people can understand a picture and match it easily with the real world unlike text, oral indications or maps. As far as we know, none automatic systems based on pictures exist. At the preparation of the itineraries, the sequence of pictures can be generated and then sent one-by-one when necessary.

2.5 Criticisms of existing navigation systems

The existing navigational systems are usually designed for car-drivers and based on a positioning system such as GPS. They are all map-based accompanied sometimes by textual indications or spoken commands when arriving at the vicinity of a turning place. In other words, those systems are using a combination of the previous modes only for outdoor itineraries.

In the other hand, there exist systems for desktop or laptop computers to prepare itineraries, or to make a sort of virtual visit of a city. Among those systems, we can quote:

- the French system of yellow pages with photos: http://photos.pagesjaunes.fr/, showing streets and façades for a large variety of cities in France and in Spain,
- and the system "street view" of Google maps for some American cities.

From the French system, we have borrowed the idea of pictures decorated by arrows as given in Figure 1.

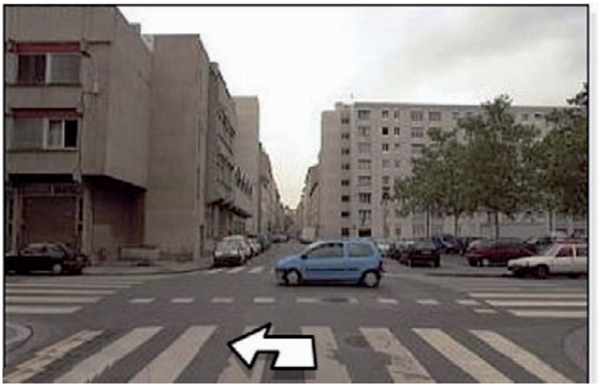


Fig 1. Example of a picture decorated with an arrow to show where to go.
Source: http://photos.pagesjaunes.fr/.

The conclusion of this rapid analysis is that those systems are used essentially outdoor for car-drivers: they cannot be used directly for walkers and pedestrians, and as far as we know none are based on pictures.

2.6 Comparison table

Bearing all that considerations in mind, a comparison among the different modes explained earlier can be made based on several criteria.

Among them, let us mention the necessity of abstracting the space, the memory burden, the local and the global vision of the itinerary, language independence, clarity and complexity of the given description and the robustness of identifying landmarks. Table 1 summarizes this comparison.

Finally, the consequence is that picture-based systems can be an interesting candidate as well outdoor as indoor for pedestrians equipped with some handheld device.

Table 1. Comparison of different modes of describing itineraries.

Modes/Criteria	**Text-based**	**Voice-based**	**Map-based**	**Picture-based**
Abstraction of space	High	High	High	Low
Memory burden	Medium	High	Medium	Low
Space representation	Difficult	Difficult	Good	Excellent
Global vision	No	No	Yes	No
Local vision	Possible	Possible	Yes	Excellent
Language independence	No	No	Generally yes	Yes
Clarity of description	Low	Low	Medium	High
Complexity of the description	High	High	Medium	Low
Landmark recognition robustness	No	No	By location	Yes

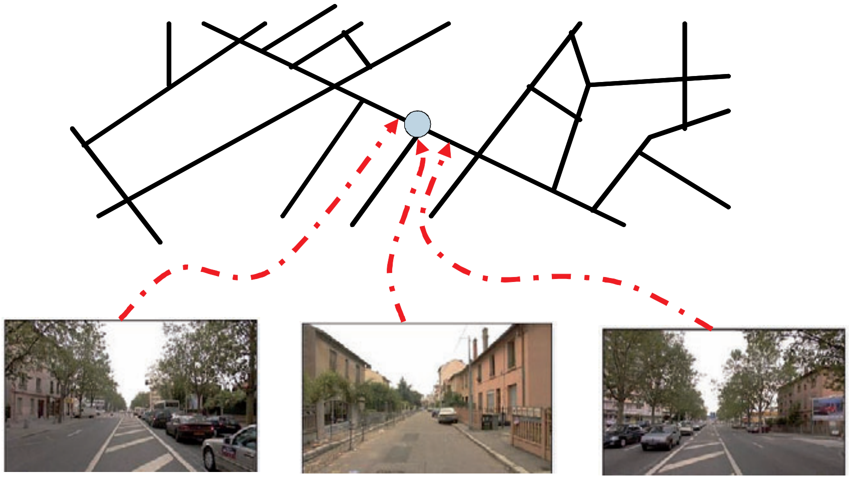


Fig. 2. Assigning pictures to arcs.

3. Basic considerations for picture-based itinerary description

In order to design such a system, the pre-requisites are:

- not only to model the road network as a graph, but also all places which can be walked by pedestrians, outdoor and indoor;
- to take pictures of all the required views and structure them into a picture database;
- and to model the itinerary description as an ordered sequence of pictures decorated by arrows.

From now on we will call this graph the pedestrian graph. Since there is no one-way path for mobile pedestrians, the edges become arcs in both directions. For each arc, let us speak of the origin node and the destination node.

3.1 Picture acquisition

Pictures must be taken for each arc and with a precise protocol (Figure 2). More exactly, the photographer must stand and be positioned in the origin node (N_1), and the picture must be directed along the arc such as the destination node (N_2), must be more or less in the centre of the picture. In order to position exactly the arrow, the coordinates (in pixels) of the destination node must be stored together with the picture.

If the distance between two nodes (N_1N_2) is very long, it could be interesting to create additional nodes. A threshold could range between 50 and 100 meters.

The starting phase is composed of:

- the pedestrian graph in which edges in both senses are materialized,
- the outdoor paths and indoor corridors,
- and the pictures that are associated to all edges in both senses.

Two experimental databases have been made, the first one in the city of La Plata in Argentina, and the second in La Doua Campus in Lyon, France. But presently only outdoor pictures are taken and stored.

3.2 Picture DB and itinerary query answering

The minimum database will be organized as follows (Figure 3).

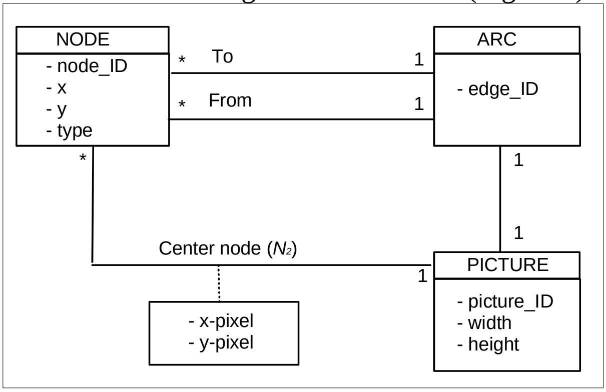
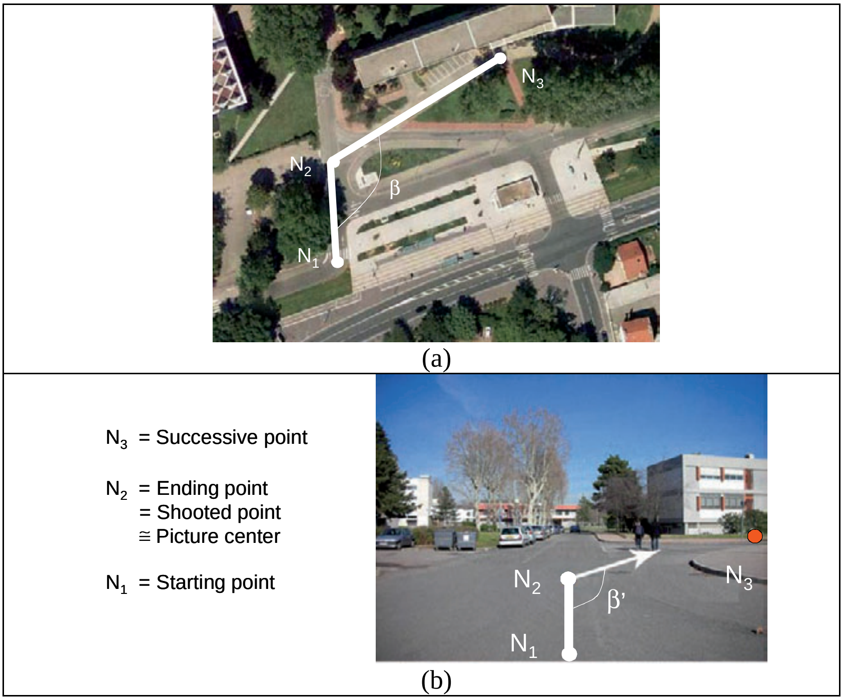


Fig. 3. Class model of the picture database.

Once the pictures have been taken according to the previous protocol, a database must be created. In essence, it is a graph network in which pictures are attached to arcs.



(a)

(b)

Fig. 4. Decorating pictures with arrows. (a) aerial photo. (b) picture of the same place taken according to our protocol.

The more important query linked to our purpose can be specified as follows: “give me the ordered sequence of pictures for going from one place to another place and set the results according to my position and my pace?” This query must use a minimum path algorithm such as Dijkstra or A* [Zhan, 1997] the result of which is an ordered sequence of nodes and arcs, which will be immediately transformed into a sequence of pictures decorated with arrows (Figures 4 and 5).

3.3 Arrow decoration

This arrow must be understandable for the user without hiding important features of the pictures (Figure 4).

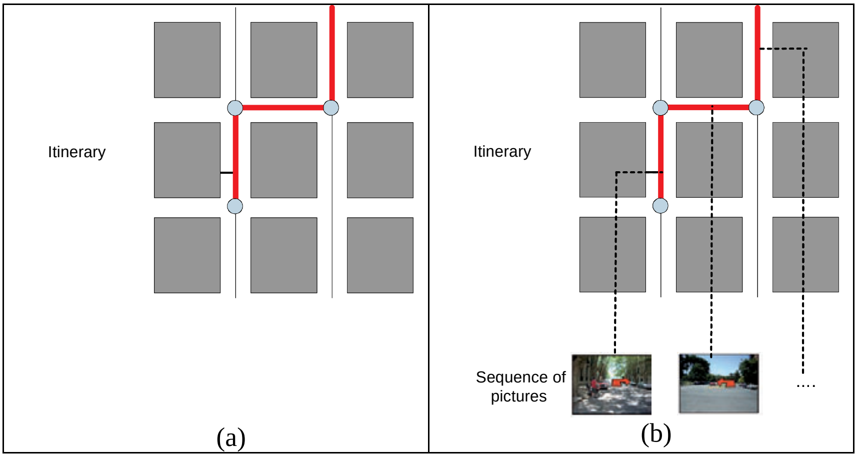


Fig. 5. From itinerary (a) to a sequence of pictures.

As a consequence, two arcs and three nodes must be considered (Figure 4):

- the origin node from which the picture was taken (N_1)
- the destination node which is approximately in the centre of the picture (N_2)
- and the successive node of the itinerary (N_3).

So, the current picture will correspond to the arc from N_1 to N_2 (as stored in the database), whereas the decorating arrow will correspond to the arc from N_2 to N_3.

One first solution should be to compute those arrows at the creation of the database; but the main drawback is the multiplication of pictures to store (about three times more). Instead, we prefer compute those arrows on the fly. Three problems have to be solved (1) the position of the arrow, (2) its direction and (3) its color.

Once a solution is defined, some cognitive studies must be performed to examine user’s reactions.

Arrow positioning

The arrow will be positioned as follows: its source will be located at the N_2 pixel coordinates as stored in the picture database, its length and width can be parameters of the system, for instance 20 % of the picture size for the length.

Angle computation

To determine the arrow, we need to know the coordinates of three nodes (Figure 6):

- the starting node, N_1, in which the observer is standing to take the picture,
- the shooted node, N_2 which will appear approximately in the centre of the picture,
- and the successive node N_3, knowing that the arrow will begin in N_2 and targeting N_3.

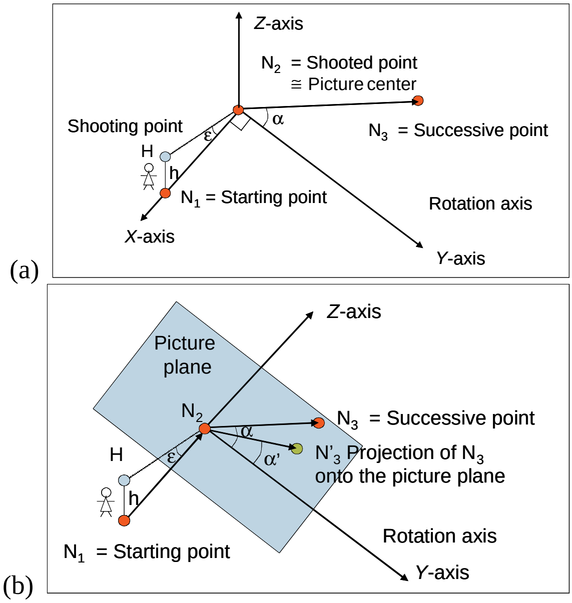


Fig. 6. Context for arrow angle computation.

First let us suppose a linear transformation in the plane so that the new axes have their origin in N_2, and the x axis is directed to N_1. The observer is standing with the feet in N_1, and his eye is located in a point H, with the elevation h.

The picture plane *Im* is supposed to be perpendicular to the line of sight, i.e. to the vector HN_1. The objective is to find the projection of N_3, named N'_3 onto the picture plane *Im*.

As a consequence, the main points are defined as follows:

$$N_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, N_1 = \begin{bmatrix} x_{21} \\ 0 \\ 0 \end{bmatrix}, N_3 = \begin{bmatrix} x_{23} \\ y_{23} \\ 0 \end{bmatrix}, H = \begin{bmatrix} x_{21} \\ 0 \\ h \end{bmatrix}.$$

And the following important angles:

- the shooting angle $\varepsilon : tg\varepsilon = h / x_{21}$
- and the initial arrow angle $\alpha = \beta - \pi / 2 : tg\alpha = y_{23} / x_{23}$.

First let us determine the picture plane equation *Im*. Since it is perpendicular to HN$_2$, the following dot product must be null: $x_{21}X + 0Y + Zh = 0 \Rightarrow X = -Zh / x_{21} \Rightarrow X = -Ztg\varepsilon.$ So the *Im* plane equation is $X = -Ztg\varepsilon$.

Now, we have to determine the equation of the line *L* supporting the vector *HN$_3$*. We can see that it is at the intersection of two planes, the first one passing through the points: *H*, *N$_3$*, *N$_2$*, and the second one through *H*, *N$_3$*, and *N$_1$*.

a/ Since the origin *N$_2$* belongs to the plane *HN$_3$N$_2$*, its generic equation is in which *a*, *b*, and *c* are coefficients to be determined from the general equation: $aX + bY + cZ = 0$.

Since the points *H* and *N$_3$* belong to the *Im* plane, we have:

$H : +ax_{21} + ch = 0 \Rightarrow c = ax_{21} / h = -a / tg\varepsilon.$

$N3 : ax_{23} + by_{23} = 0 \Rightarrow b = -ax_{23} / y_{23} = atg\alpha.$

By substituting those values in the general equation, we obtain the first equation of *L*: $X + Ytg\alpha - Z / tg\varepsilon = 0$.

b/ For the second plane *HN$_3$N$_1$*, we need to consider a more general equation: $aX + bY + cZ = d.$ So:

$H : ax_{21} + ch = d.$

$N3 : ax_{23} + by_{23} = d.$

$N1 : ax_{21} = d.$

By rearranging, we obtain:

$c = 0$

$a = d / x_{21}$

$$b = d\frac{x_{21} - x_{23}}{x_{21}y_{23}}$$

So the second equation for L is $X + Y\dfrac{x_{21} - x_{23}}{y_{23}} = x_{21.}$

This line *L* will cross the *Im* plane in *N'$_3$* the coordinates of which must be the solution of the following system of three equations:

$$(1): X + Ytg\alpha - Z / tg\varepsilon = 0$$

$$(2): X + Y\frac{x_{21} - x_{23}}{y_{23}} = x_{21}$$

$$(3): X + Ztg\varepsilon = 0.$$

By applying (1) and (3), we obtain:

$$(X + Ytg\alpha - Z / tg\varepsilon = 0) \Rightarrow (X + Ytg\alpha + X / tg^2\varepsilon = 0) \Rightarrow (X / tg\alpha = -Y\sin^2\varepsilon)$$

Reporting in (2) gives:

$$(X + Y\frac{x_{21} - x_{23}}{y_{23}} = x_{21}) \Rightarrow (Y(\frac{x_{21} - x_{23}}{y_{23}} - tg\alpha\sin^2\varepsilon) = x_{21})$$

$$\Rightarrow Y = \frac{x_{21}}{\frac{x_{21} - x_{23}}{y_{23}} - tg\alpha\sin^2\varepsilon} = \frac{y_{23}x_{21}}{(x_{21} - x_{23}) - y_{23}tg\alpha\sin^2\varepsilon}$$

To simplify, let us denote $T = (x_{21} + x_{23}) - y_{23}tg\alpha\sin^2\varepsilon$; so giving:

$$Y = \frac{y_{23}x_{21}}{(x_{21} - x_{23}) - y_{23}tg\alpha\sin^2\varepsilon} = \frac{y_{23}x_{21}}{T}$$

$$\Rightarrow X = -Y\sin^2\varepsilon tg\alpha = -\frac{y_{23}x_{21}tg\alpha\sin^2\varepsilon}{T}$$

$$\Rightarrow Z = -X / tg\varepsilon = \frac{y_{23}x_{21}tg\alpha\sin^2\varepsilon}{Ttg\varepsilon}$$

So the result is:

$$N'_3 = \begin{bmatrix} x'_3 \\ y'_3 \\ z'_3 \end{bmatrix} = \begin{bmatrix} -\frac{y_{23}x_{21}tg\alpha\sin^2\varepsilon}{T} \\ \frac{y_{23}x_{21}}{T} \\ \frac{y_{23}x_{21}tg\alpha\sin^2\varepsilon}{Ttg\varepsilon} \end{bmatrix} = \frac{y_{23}x_{21}}{T}\begin{bmatrix} -tg\alpha\sin^2\varepsilon \\ 1 \\ tg\alpha\sin\varepsilon\cos\varepsilon \end{bmatrix}$$

Now, let us transform the *N'*$_3$ coordinates in the *Im* plane. We can pass to it by a rotation of angle ε along the *y* axis (*XX*, *YY*, *ZZ* being the coordinates in the plane); in the same time, we can take into consideration the focal length *F* and the scale (from meters to pixel width):

$$XX = F \times (x \times \cos\varepsilon + z \times \sin\varepsilon)$$

$$YY = F \times y$$

$$ZZ = F \times (-x \times \sin\varepsilon + z \times \cos\varepsilon)$$

But in reality, we are concerning by the angle α‘ in the picture. This angle can be defined from for the *N’₃* point in the *Im* plane:

$$tg\alpha' = \frac{ZZ}{YY} = tg\alpha(\sin^3\varepsilon + \sin\varepsilon\cos^2\varepsilon) = tg\alpha\sin\varepsilon(\cos^2\varepsilon + \sin^2\varepsilon) = tg\alpha\sin\varepsilon$$

So,giving the result $tg\alpha' = tg\alpha\sin\varepsilon.$

We can check this result very easily in two important cases:

1 – when the photographer is vertical or in case of aerial photo, we obtain $\varepsilon = 90°$, and sin $\varepsilon = 1$ so giving $\alpha' = \alpha$; this means that the arrow is positioned along the future edge without any modification;

2 – when the photographer is lying flat on the ground, we have ε = 0° and sin $\varepsilon = 0$ so giving $\alpha' = 0$ *or* $\alpha' = \pi$ *;* this means that the various possible directions are unrecognizable.

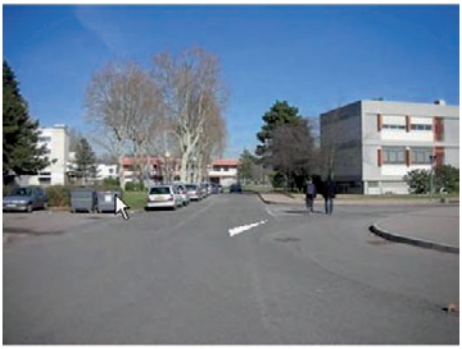


Fig. 7. Image decorated by an arrow.

Color selection

According to Itten´s theory about colors [Itten, 1997], it is convenient to select the more contrasting color for the arrow. Two solutions can be examined. The first one is a priori to select one fixed color which is used for all pictures; the second is to compute the optimal color for each picture.

The solution based on a fixed color for the arrow is that the contrast could not be sufficient enough to be understandable. Suppose green is selected for the color of all arrows it can be a good solution when the arrow is laid on tar color, but not when the itinerary is a footpath in a forest.

The variable color corresponding to second solution (Figure 7) can be computed by considering the histogram of colors in the Hue-Lightness-Saturation system, around the centre of the picture. For instance we consider the minimum value of this histogram which is used for the arrow as the most convenient color. The main drawback is that the color evolves from one picture to another, perhaps disturbing the user. For instance, a red color or a green color can be interpreted differently [Bertin, 1973]. In order not to have difficulties with color representation, only 8 buckets of colors are selected. The main steps of the algorithm are as follows:

i – consider the Itten circle of colors, and divide it into eight regular buckets,
ii – make a circular histogram with those eight buckets,
iii – select the maximum peak of this circular histogram (which represents the dominant color),
iv – select the opposite (which is the more contrasting color, i.e. the arrow's color).

4. Main characteristics of a picture-aided navigational system

Taking all previous considerations into account, a navigational system based on pictures can be defined. This mobile system must be based on the following components:

- a server storing the pedestrian graph, the picture database, running the minimum path algorithm, and sending the sequence of pictures,
- a handheld device for the user always in connection with the server,
- a communication infrastructure perhaps based on WIFI systems allowing the service roaming,
- and a positioning system such as GPS for outdoor, or based on beacons indoor.

Remark that the arrow decoration can be made either at server or at client level.

Several of these components are outside the scope of this paper. However, we will present some more important issues to be solved.

4.1 Visual Interface

With his handheld device, the user must specify the place where to go and give his preferences. So a path query is sent to the server which runs the minimum path algorithm and returns a set of pictures (Figure 8). The decoration phase can be made either on the server or in the client (handheld device).

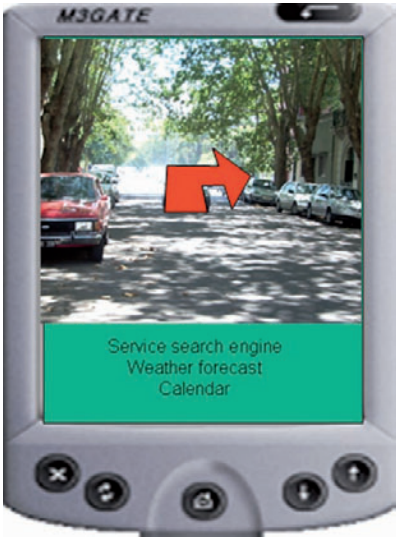


Fig. 8. Example of visual interface.

4.2 User Positioning

Since the system is targeted to run outdoor and indoor, two different positioning systems must be used. When outdoor, the usual GPS system will be used, whereas indoor, some other positioning systems will be used such as beacons or RFID systems. Those functionalities are yet totally operational, especially considering the continuity of service.

4.3 User Disorientation

When the user is lost, it means that the position of the user is different from a planned position. In this case, the path is recalculated from the current point to the defined target, and a new picture set is computed, and sent to the user.

5. Conclusions

In this paper, we have presented a new concept concerning the explanation of itineraries based on a sequence of arrow-decorated pictures, which is the basis for any pictured-aided navigation systems. We do think that in the future, such a system will be considered as the kernel as Location-Based Services and physical hypermedia.

Let's thank the anonymous referee who gives us references of the LOCUS project (http://www.locus.org.uk/) of the City University of London, the objectives of which bear some similarities with our project.

Presently, two prototypes were made in this direction, one in Argentina and one in France, and we think that the concept is now validated. However, several problems must be solved such as the design of a totally operational prototype, working outdoor and indoor, and the continuity of services.

Another important aspect is the cognitive side of the problem in order to test the reactions of the users.

References

[1] Bertin J. (1973). "*Sémiologie graphique*", Mouton/ Gauthier-Villars, 2nd edition, 1973.

[2] Dijkstra E. W. (1959) "*A note on two problems in connection with graphs*". Numerische Mathematik. 1 (1959), S. 269–271

[3] Itten J. (1997). "*The Art of Color*". Wiley. (December 1997).

[4] Küpper A. (2005) "*Location-Based Services: Fundamentals and Operation*" Wiley, 386 p.

[5] Rossi G., Gordillo S., Challiol C., Fortier A. (2006) "*Context-Aware Services for Physical Hypermedia Applications*". OTM Workshops 2006, Proceedings, Part II. Lecture Notes in Computer Science 4278 Springer 2006, ISBN 3-540-48273-3 pp. 1914-1923.

[6] Schiller J. and Voisard A. (2004), "*Location-Based Services*", Morgan Kaufmann Publishers, 255 p.

[7] Wang N. (2007) "*PHOTOWAY, Describing itineraries by pictures with arrows in pervasive systems*", Master Dissertation in Computing, INSA de Lyon, June 2007.

[8] Zhan F B. (1997) "*Three Fastest Shortest Path Algorithms on Real Road Networks*". Journal of Geographic Information and Decision Analysis, vol.1, no.1, pp. 69-82.

Road Network Model for Vehicle Navigation using Traffic Direction Approach

Yang Yue[1,2], Anthony Gar-On Yeh[1], Qingquan Li[2]

[1] Centre of Urban Planning and Environmental Management,
The University of Hong Kong, Hong Kong SAR, China
e-mail: {yangyue,Anthony.Yeh}@hku.hk

[2] The Transportation Research Center at the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing,
Wuhan University, Wuhan, 430072, People's Republic of China
e-mail : qqli@whu.edu.cn

Abstract

In conventional GIS, road links are represented by a centerline-based modelling schema which raises many difficulties in lane-specific transport applications, such as navigation, simulation, and visualization. To meet such lane-specific requirements, road network data model has evolved from centerline-based model, to carriageway-based model, and furthermore, to lane-based model. Varying in representation scale, each schema has its advantages and disadvantages, and no consensus has been reached on which one is more proper or efficient. The evaluation and selection of data model should be closely related to the application purposes and requirements. This paper examines the modelling strategies of the three schemas in the context of vehicle navigation. Based on navigation characterists, this paper further puts forward a GIS road network model which is not based on road object scale, but the driving direction that each road link provides. Compared with the other three modelling schema, this schema is more effective on topology maintenance, multi-cartographic representation, and data compaction.

Keywords: Road network data model; Vehicle navigation

1. Introduction

Although GIS is increasingly used in transportation field, it still has many limitations in modelling road network, especially in the representation and analysis of lane-specific information which is very important for traffic management and vehicle navigation. Because as the basic element of roadway, lanes often have different attributes, such as lane availability, driving direction and dedicated lane use (for example, bus/HOV (high occupied vehicle) lane), and therefore, different traffic conditions. As a result, many transport applications, such as transportation planning and congestion management, often require at least a bidirectional centerline if not dual carriageways or even individual lanes, either in basic geometry or by attribution (Dueker and Bender, 2003). For vehicle navigation, lane-specific information is more important because once a driver enters the wrong lane, he/she may not be able to get back to the correct lane because of lane driving restrictions, such as double-white line.

Figure 1 shows a road section that illustrates lane-based driving directions (represented by lines with arrows) caused by the variation of lane connectivity, which is a case of a typical section with on- and off-ramps. It could be even more complicated in many high-density areas where road systems are much more complex due to the construction of large number of intersections and interchanges.

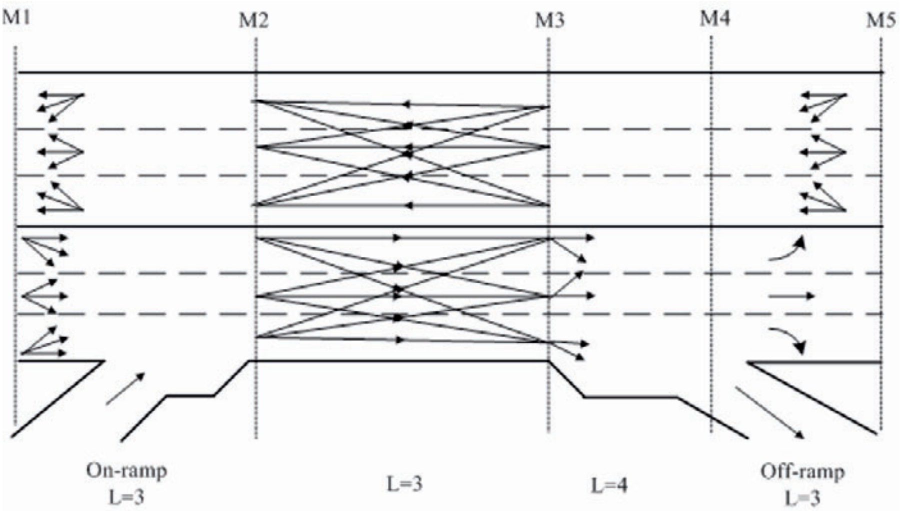


Fig. 1. Traffic flow direction and lane connectivity.

To represent lane-based diversities and disseminate them effectively to traffic control centres and individual drivers, it is necessary for the road network model to have the ability to describe both physical and traffic attributes of each lane. However, due to the heterogenous lane features within a single roadway, the description, i.e., the modelling of lane information, also has several schemas.

A roadway is often represented as a polyline in GIS, i.e., by its centerline. Lane information, if available, is stored as attributes of the centerline. Therefore, both the database and interface operations are centerline-based and they are extended to lane level only when needed. A lane in this schema is represented as a virtually existed cartographic symbol rather than a real database object. This is inefficient when frequent lane-based operations are involved. From an object-oriented perspective, although ideally each lane should be expressed as a primary geometry and visualization element, such lane-based representation schema could suffer from a significant data redundancy, and result in decreases in operation and maintenance efficiency when inappropriately designed. To solve this dilemma, carriageway-based model is developed which is in between of centerline-based and lane-based schemas. However, each of the three road network modelling schema has pros and cons and no consensus has been reached on which one is more proper or efficient in representing lane-based spatial and attribute information. This paper first reviews the characteristics of the three data modelling schemas, and then puts forward a new modeling schema based on traffic direction instead of lane, carriageway or centerline to support vehicle navigation.

The remainder of the paper is organized as follows. Section 2 investigates the modelling strategies of the three road network modelling schemas. Based on the analyses, the traffic direction-based model is proposed in section 3. Conclusions are given after model comparsion is conducted in section 4. The assumption of this paper is that details of lane-related attributes and topology information are available.

2 Existing modeling schemas

Most lane-specific applications are related to lane topology for representing inter-lane connectivity and multi-cartographic representation (Vonderohe et al., 1997). Lane topology is determined by both lane spatial connectivity and lane attribute, such as dedicated lane use (for example, specific vehicle mode and property), and temporally lane-specific objects and events (for example, temporal lane closures and reversible lane system). Under different application purposes and map scales, the ability to provide appropriate cartographic representation of road object is also necessary, such as from centerline to carriageway and to lane level, and vice versa. To meet such requirements, implementation strategies vary among different models, such as GDF – ISO/TC 204 (TC 204, 2004), Kiwi (Kiwi-W Consortium, 2001), NCHRP 20-27 (Vonderohe et al., 1997), NCHRP 20-64 (TransXML, 2005), NSDI Framework Transport Identification Standard (FGDC, 2001), UNETRANS data model (Curtin et al., 2003), and MDLRS (Koncz and Adams, 2002). MDLRS is basically developed on the concept of NCHRP 20-27, and UNETRANS references the NSDI standards. Generally, these models can be classified into two categories: centerline-based model and carriageway-based model. As for lane-based model, the one proposed by Malaikrisanachalee and Adams (2005) is a representative of such development.

This section examines the three main road modelling schemas by their primitive element, i.e., centerline, carriageway and lane, concentrating on road geometry representation, lane-level topology and analysis, while taking the efficiency of data organization and operation into consideration.

2.1 Centerline-Based Modelling Schema

Centerline-based model is the most conventional road representation method in which each roadway is stored as a polyline, and lane information is stored as attributes of the polyline with the spatial information for lanes being limited to start points and end points along the relevant roadway as illustrated by Fohl et al. (1997). Therefore, the road section in Figure 1 is represented as a single line R1 and because the number of lanes various along the road section, five points M1 to M5 are needed as referencing positions to mark the variation ---- M1 and M2 are the *From* and *To* positions of the first lane section; lanes starting from M2 to M3 belong to a second group; similar, M3 and M4, M4 and M5 are used to maintain the third and forth group of lanes, respectively. And lanes within the same group are numbered uniquely, for example, 1 to *n* from right to left. Figure 2 illustrates how the relationship is organized to represent the first lane group, i.e., lanes (L1~L6) between M1 and M2.

In this schema, the method of representing lane information is analogous to the way how linear attributes are modeled using linear reference or dynamic segmentation which causes dependency of lanes on roadways and restricts the representation of heterogeneous lane attributes (Malaikrisanachalee and Adams, 2005). It also generates difficulty for lane-level visualization, for example, the display of different traffic conditions of the opposite directions, or on each individual lane. Other inadequacies of the method have also been identified. They are mainly related to its inability to represent multi-cartographic/spatial topology (Demirel, 2004). However, multi-cartographic feature can be achieved by using of a hierarchical structure as in the Kiwi format (Kiwi-W Consortium, 2001). For example, in a higher level, an intersection can be a single point, while in a lower level the intersection can be a composition of a set of nodes and links according to the intersection type and the level of user and application purpose as shown in Figure 3. The implementation of the hierarchical structure can be either graphically or abstractly derived by topology relationship.

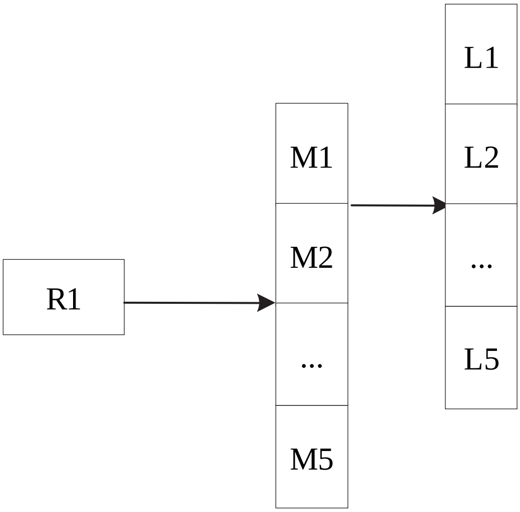


Fig. 2. Relationship among road, reference points and lanes in centerline-based model.

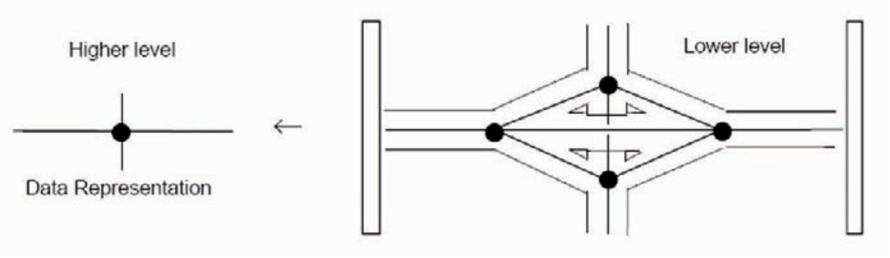


Fig 3. An example of the hierarchical structure of Kiwi (Kiwi-W Consortium, 2001)

In terms of topology maintenance, although the centerline-based model provides a compact and efficient means to maintain road geometry, it requires massive additional points and turntables to maintain lane-specific topology. For example, the lane-level topology is maintained by two tables together: point turntable and linear turntable. The former marks the point where a turn begins while the latter records the linear range that the turn is available. As a consequence, at each point where lane connectivity changes for every lane, the point and turntables have to be recoded. Therefore, both the building and updating work are very tedious in the centerline-based model.

2.2 Carriageway-Based Modelling Schema

Carriageway-based model to a certain extent is an extension of centerline-based model. It uses carriageway as primitive element for road network representation and analysis. Therefore, carriageway-based model has smaller granularity and can be more effective from both lane-based cartographic and analysis perspectives. A typical example of carriageway-based model is GDF, an European standard used to describe and transfer road networks and road related data, whose outcome formed the major input to the world standard ISO GDF 4.0 created by ISO/TC204 WG3. Comparing with centerline-based model, it considers each carriageway independently and consequently, a road with two carriageways is often treated as two objects. Meanwhile, same as that in centerline model, lanes are maintained as attributes of carriageway. Using this model to represent the road section shown in Figure 1, one more layer, i.e., carriageway is added. As shown in Figure 4, C11 and C12 are the center of the relationship, which represent the northbound and southbound carriageway respectively. And three lanes, for example L1~L3 are linked to C12 through referencing points M2~M3.

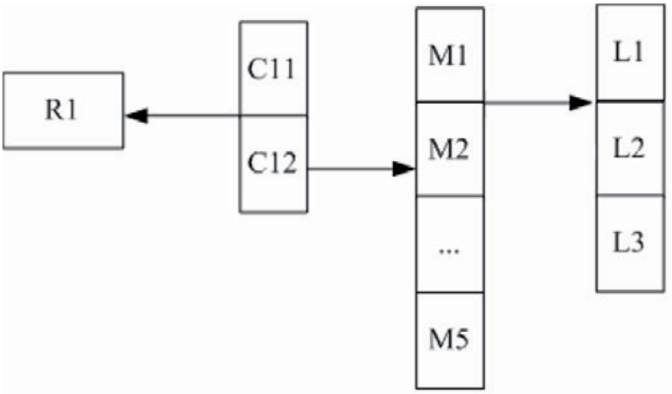


Fig 4. Relationship among road, reference points and lanes in carriageway-based model

To fulfill multi-representation and topology requirement, a three-level structure is adopted in GDF, as shown in Figure 5 that, Level 1 is the core of the structure whose elements receive a real world significance, and Level 2 are composed of a group of simple or complex features and thus represents abstract topology (TC 204, 2004).

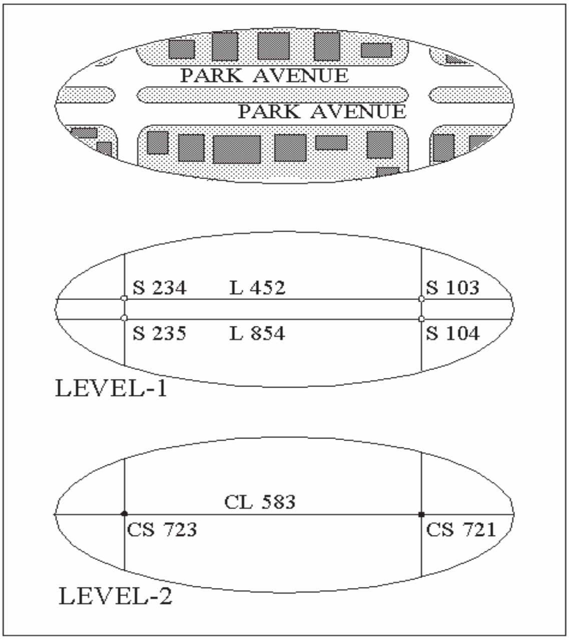


Fig 5. A road can be represented as two objects in Level 1 or one object in Level 2 (TC 204, 2004)

Therefore, similar to the hierarchical structure in Kiwi, this structure enables the display of complex road network at different granularity levels, from a single line to the details of carriageways, especially that it enables a more direct data representation and analysis at carriageway level, thus can meet the requirements of traffic management and analysis better than centerline-based models. As to lane information, this model relaxes the dependency of lane from centerline to carriageway, which is more feasible and accurate in visualization, analysis, and data update as well. Although it is inevitable that this model increases the database size, it is more transport-oriented, therefore, has a chance to outperform centerline-based models in terms of transport-related query, topology maintenance, and facility management.

However, both centerline and carriageway-based lane model dwell on the roadway geometry, and lanes are referred to centerline or carriageway by certain links such as lookup tables. In the both models, lane-level visualization and analysis are still indirect.

2.3 Lane-Based Modelling Schema

Realizing the importance of lane in transport applications, the idea of lane-based road model emerged. Fohl et al. (1997) put forward the concept of lane-based traffic data model in which ideally lanes would be used as display features as well as primary modelling elements. A representative lane-based model was proposed by Malaikrisanachalee and Adams (2005), which maintains lanes as independent geometry and topological objects, and road network is modeled as a directed graph where traffic flow connectivity between lanes is defined by nodal adjacency. To avoid a significant increase of data size, the model concentrates on the aggregation of lanes and nodes.

Following the idea of the lane-based model, the southbound carriageway of Figure 1 can be modelled as:

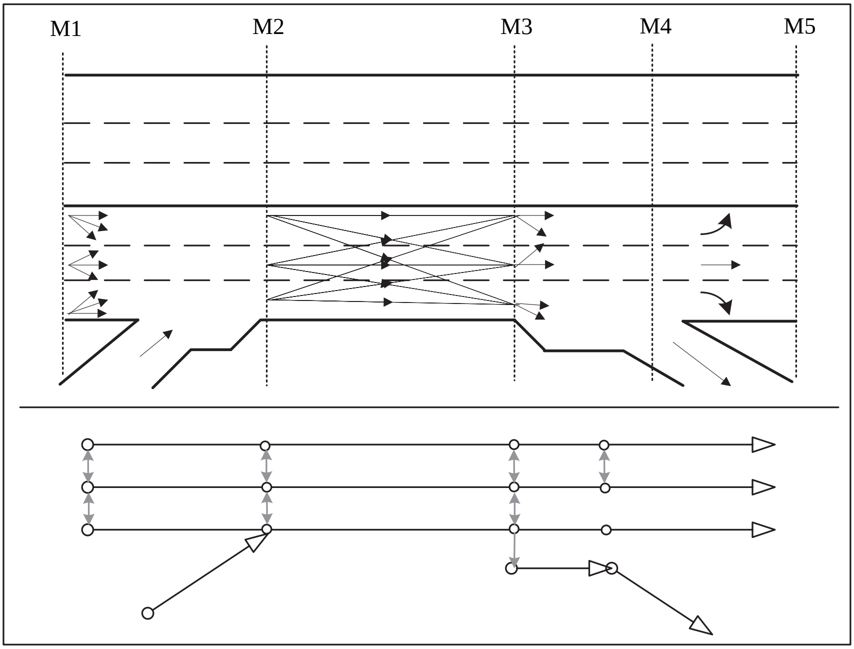


Fig 6. Lane-based road network model proposed by Malaikrisanachalee and Adams (2005)

In this model, lane-specific attributes and events are maintained by both linear offset and the horizontal offset, which can be point, linear, or area-based. This is a more feasible way to manage lane information compared with centerline and carriageway-based modelling schemas.

Because lane connectivity can be derived from geometric relationship, this mode eliminates the encoding and maintenance costs of turntables. Meanwhile, two main strategies are used in the model to express lane connectivity. One is “lateral sequence” that specifies the relative lateral position of the lane among parallel lanes. The other is discretization of the continuous lateral connectivity among parallel lanes by using extra virtual nodes section by section to connect parallel lanes. Therefore, lane connectivity can be obtained not only from geometric relationship, but also the lane lateral sequence and discretization. However, although discretizaiton method reduces the size of database by eliminating the linear turntables required in Fohl’s model, it is achieved at the expense of the management of increase of geometry objects and extra nodes and links as well.

Similarly, a kind of hierarchy linkage is also used in this model among lane, carriageway and centerline. But Malaikrisanachalee and Adams do not explicitly explain in which way the hierarchical structure is organized and they only tell how lane-based nodes and links are associated with centerline node and carriageway links. The problem is when lane-based geometry entities are available, how should carriageway and centerline geometry be organized? Should they be derived from the chain coordinates of child links, or stored as concrete geometry entities, like the method used in UNETRANS? As a transport-oriented data model, UNETRANS was compatible with ArcGIS 8.1. It maintains both centerline and carriageway objects to meet the multi-level representation requirement, therefore triples the size of geometry database. This is one of the reasons why UNETRANS model has not been widely accepted (Curtin et al., 2003).

In general, each of the above road network modelling schema uses certain strategies to meet lane-specific transport requirements in one way or the other which has both advantages and disadvantages. In the next section, an innovative modelling schema that models road network using traffic direction on each roadway instead of roadway geometric element is proposed.

3. Traffic direction based modeling schema

The function of a road network is to provide traffic accessibility. Therefore, the form of road network from another angle can be abstracted as a composition of traffic flows on it. In other words, traffic flows presents the shapes of road network. Vehicle driving direction is the focus of vehicle navigation. It is determined not only by the road network physical form, but also the traffic regulations applied to it, such as lane changing and turning restriction. Therefore, it may be more effective to model road network using roadway traffic direction as a primitive modelling element. Based on this perspective, the proposed new road network model is based on roadway traffic directions.

The traffic direction-based model can be considered as a model that is in-between the carriageway-based and lane-based model. Its primitive element is not a physical existing object, but an abstract vehicle driving direction. In carriageway-based model, several permitted vehicle driving directions are generlized into one modelling object; while in lane-based model, vehicle driving directions are usually splitted into lane level, and as a result, overlapped driving directions often exist. The general principle of the new modelling schema is that only permitted vehicle driving directions are recorded, such as straight driving, right turn and left turn. Therefore, if every lane in a carriageway has the same driving direction, the model is actually generalized into a carriageway model; if every lane has different driving direction, the model is the same as lane-based model in terms of spatial representation. Using the principle to model the road section in Figure 1, a single line is generated to represent the northbound carriageway as in the carriageway-based model since all the lanes in this bound, i.e., carriageway, have the same driving direction. As for the southbound carriageway where five traffic directions exist, then five modelling objects are used according to the above rule, i.e., on-ramp flow, off-ramp flow, straight ahead flow, right-turn and left-turn flow as shown in Figure 7.

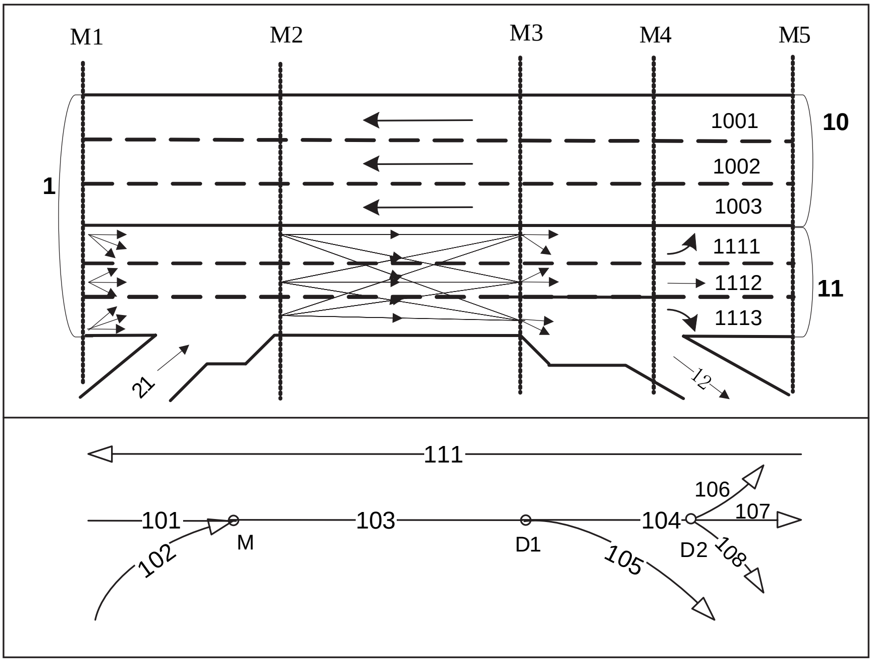


Fig. 7. Flow-direction road network model

This modeling schema populates parallel lanes with lateral connectivity and the same driving direction into one traffic flow object. Associated lane attributes of each individual lane, such as HOV, speed limit, and temporal events, can still be modelled by the use of dynamic segmentation and linear referencing system (LRS). This method meets the basic requirement of graph theory which requires that there is only one directed link between two nodes. This requirement is also the main concern of lane-based model in order to use existing network algorithms.

To establish network toplogy, topological points are added to represent the merging and diverging of traffic flow directions as shown by points *M*, *D*$_1$ and *D*$_2$ in Figure 7. In general, the southbound carriageway of the road section are modelled using eight polylines which correspond to the five traffic flow directions on it while taking the road network topology into consideration. The location of merging point can be the exact merging point where two flows meet. While the location of diverging point should be some distance (for example, 100 or 200 meters) ahead of the physical diverging point for navigation purpose, such as for guiding the change of lane. Because of the turning restrictions applied on each lane, for example, 106, 107 and 108 in this example, it makes this model different from carriageway-based model. Otherwise, the flow-dirction road network model will be equal to carriageway-based model in terms of spatial objects stored in database. In this way, this modelling schema represents both turns at point and turns between parallel lanes.

To meet multi-cartographic representation requirement, a similar hierarchical structure, i.e, a lookup table is also maintained. There is a linkage between traffic flow direction and lane, which is a one-to-many relationship because each traffic flow direction can correspond to more than one physcial lanes as shown in Table 1. Based on the same reason, the relationship between traffic flow direction and carriageway is many-to-one or at least one-to-one. Maintaining this table can fulfill the multi-topology requirement as well.

Table 1. Relationship among flow direction, lane, carriageway and centerline

Flow Direction ID	*Lane ID*	*Carriageway ID*	*Centerline ID*
111	1001	10	1
	1002		
	1003		
101	1111	11	1
	1112		
	1113		
102	21	21	21
103	1111	11	1
	1112		
	1113		
104	1111	11	1
	1112		
	1113		
105	12	12	12
106	1111	11	1
107	1112	11	1
108	1113	11	1

Figure 8 is the UML diagram of this traffic-flow based model.

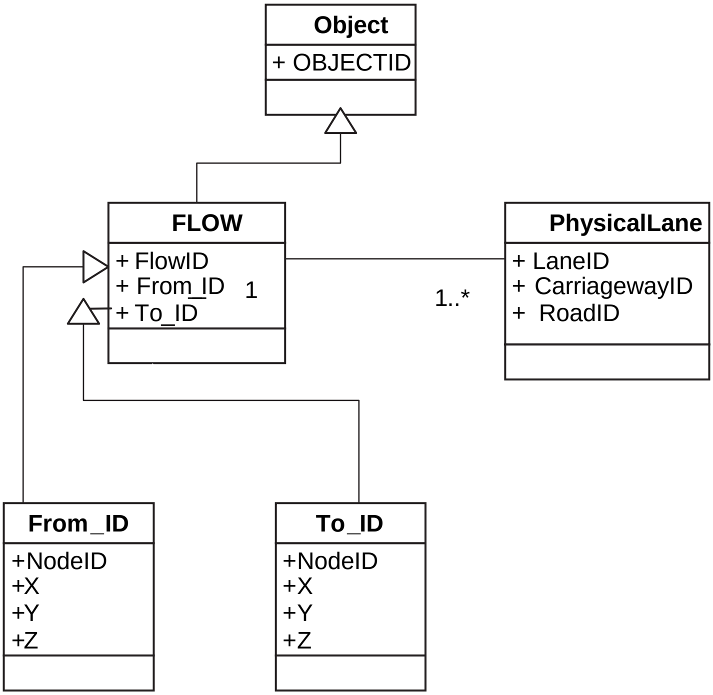


Fig. 8. UML diagram of traffic direction-based model

Compared with lane-based modelling schema, this proposed schema to a certain extent can also circumvent geometry data size problem by avoiding discretizing links. Furthermore, different from other models that rely on turntable to maintain lane-based topology, the traffic direction-based approach eliminiates turntables by physically representing the relationship into concrete entities. As shown in Figure 9, conflicting points in the intersection have three types: diverging, merging and crossing. In this modelling schema, point objects are created when lines cross at all diverging and merging points, and lines at crossing points cross without breaking. In general, it projects both point turntables and linear turntables into spatial objects. The network topology therefere is spatially deducible.

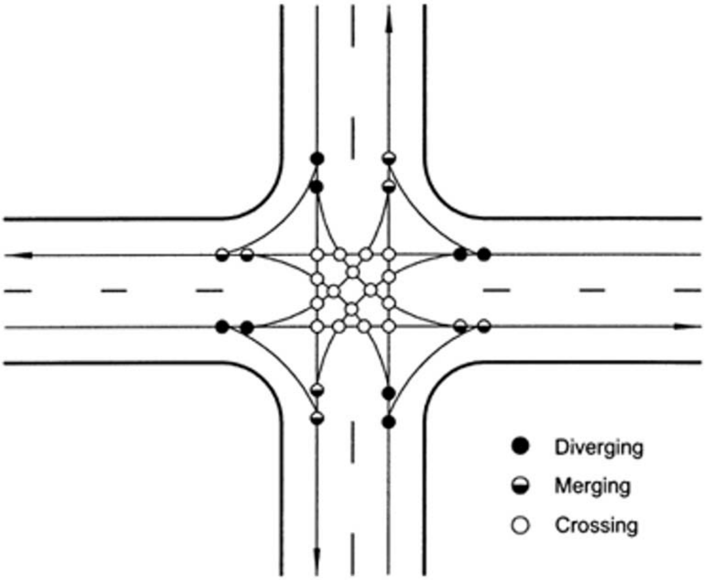


Fig. 9. Representation of topology of traffic direction model at an intersection

This model can benefit vehicle navigation because it has the ability to include lane-related information that is necessary for vehicle navigation in a relatively simplified and thus more efficient way. The next section compares the general performance of the proposed model with existing schemas.

4. Model comparison

Due to the difference in primitive modelling element granularity, the four road modelling schemas have different modelling strategies to meet lane-specific transport requirements, such as lane-based topology, multi-cartographic representation. Table 2 is a brief summary of the comparison.

Table 2 Comparison of centerline-based, carriageway-based, lane-based Model and traffic direction-based model

Model Requirements	Centerline	Carriageway	Lane	Traffic Direction
Object-level	Centerline	Carriageway	Lane	Traffic direction
Lane-based topology	Turntable	Turntable	Direct	Direct
Multi-cartographic & topological representation	Centerline→Carriageway→Lane	Centerline ←Carriageway→Lane	Lane→ Carriageway→ Centerline	Traffic flow→Lane→ Carriageway→ Centerline
Size of geometry data	+	+ +	+ + + +	+ + +

The distinctive abilities of these models and hence their performances are primarily attained by the primitive modelling element they maintained. The lane-based model maintains lanes as its primitive objects and captures the complete lane topology at the expense of exponentially increase in data size not only in its geometry but also LRS. It does not rely on turntable to maintain lane-based topology, however, the maintenance of lane connectivity is tedious. Comparatively, the traffic direction-based modelling schema can support lane-based spatial and temporal topology more directly with less extra artifical broken points and parallel lines. The centerline and carriageway-based models have to achieve lane-level granularity by using some other strategies, for example, turntable, that inevitably increase the cost of query and maintenance of lane information. To achieve multiple-cartographic ability, lookup tables are usually adopted to maintain a hierarchical structure, whose costs are similar for the four models. From the data size perspective, we use "+" to represent the size of geometry data in centerline-based model, then carriageway-based road model may double the number of road objects compared with centerline-based model, lane-based model may further triple the size of carriageway model. As for the traffic direction-based model, empirically it creates moderate geometry data size compared with lane-based model.

5. Conclusions

Conventionally, roads are represented as centerlines, i.e., link-node form in GIS. But in transport applications, this modeling schema is insufficient in lane-specific situations where the complexity of navigation, visualization and analysis increases. This requires new approaches and rules for representing road network. According to the geometric granularity of road network models in representing lane-specific information, at present, there are three types of modelling schemas --- centerline-based, carriageway-based, and lane-based models. In general, the three schemas vary in their sophistication, complexity and data requirement, and show mixed degrees of performance. Centerline-based model is compact and efficient to maintain road geometry data, but requires massive turntables to maintain lane related information. Comparatively, carriageway-based model can access lane-specific information more directly; however, both lane geometry and attributes dwell on carriageway representations. Therefore, the maintenance of lane-specific information is tedious in both centerline-based and carriageway-based models. Theoretically, the lane-based model schema conforms to object-oriented modeling principle in transport applications. However, fragmenting road and its related objects so thoroughly may lead to data redundancy and decrease the efficiency of database management.

This paper puts forward a traffic direction-based modelling schema that uses traffic flow directions on roadways as primitive modelling element. As it can concretely represent road network form and functions, traffic flow directions can describe road network features more directly, in both spatial and attribute terms. They are more useful to drivers and vehicle navigation. This modelling schema overcomes the disadvantage of lane-based model that maintains a large number of parallel lanes and their connectivity at great cost. In the mean time, traffic direction-based model has the advantage that both cenerline-based model and carriageway-based do not have, i.e., network topology is maintained directly without the need of turntables. The proposed traffic direction approach is at its early stage of development. Further works are needed to develop it into an effective and efficient operational system for mutli-lane vehicle navigation.

Acknowledgement

We would like to thank the Research Grants Council of Hong Kong (HKU 7466/06H), the National Natural Science Foundation of China (No. 40571134) and the National 973 Basic Research Program of China (No. 2006CB705500) for their support of the research.

References

Curtin, K., Noronha, V., Goodchild, M. & Grise, S. (2003). “ArcGIS Transportation Data Model (UNETRANS)”. http://www.geog.ucsb.edu/~curtin/unetrans/UNETRANS_index.htm. Accessed 2007 19 July.

Demirel, H. (2004). “A dynamic multi-dimensinal conceptual data model for transportation applications”. ISPRS Journal of Photogrammetry and Remote Sensing, 301-314.

Dueker, K. J. & Bender, P. (2003). “Building and Maintaining a Statewide Transportation Framework”. Transportation Research Record: Journal of the Transportation Research Board, 1836, 93-101.

FGDC (2001). NSDI Framework Transportation Identification Standard, Ground Transportation Subcommittee.

Fohl, P., Curtin, K. M., Goodchild, M. F. & Church, M. F. (1997). “A Non-planar, Lane-based Navigable Data Model for ITS”. 7th International Symposium on Spatial Data Handling. Delft, Netherlands London: Taylor & Francis.

Kiwi-W Consortium (2001). Kiwi Format Ver1.22. http://www.kiwi-w.org/documents_eng.html. Accessed 2007 19 July.

Koncz, N. A. & Adams, T. M. (2002). “A data model for multi-dimensional transportation applications”. International Journal of Geographical Information Science, 16(6), 551-569.

Malaikrisanachalee, S. & Adams, T. M. (2005). “Lane-based Network for Transportation Network Flow Analysis and Inventory”. TRB 2005 Annual Meeting CD-ROM.

TC 204 (2004). ISO 14825:2004 Intelligent transport systems - Geographic Data Files (GDF) - Overall data specification.

TransXML (2005). NCHRO 20-64 XML Schemas for the Exchange of Transportation Data. http://www.transxml.org/. Accessed 2007 19 July.

Vonderohe, A. P., Chou, C. L., Sun, F. & Adams, T. M. (1997). A Generic Data Model for Linear Referencing Systems. Research Results Digest (National Cooperative Highway Research Program), 218.

Clustering Algorithm for Network Constraint Trajectories

Ahmed Kharrat[1], Iulian Sandu Popa[1],Karine Zeitouni[1], Sami Faiz[2],

[1] PRiSM Laboratory, University of Versailles
45, avenue des Etats-Unis - 78035 Versailles, France
[2] LTSIRS, Ecole nationale d'ingénieurs de Tunis
B.P. 37 – 1002 Tunis-Belvédère, Tunisie

Abstract.

Spatial data mining is an active topic in spatial databases. This paper proposes a new clustering method for moving object trajectories databases. It applies specifically to trajectories that only lie on a predefined network. The proposed algorithm (NETSCAN) is inspired from the well-known density based algorithms. However, it takes advantage of the network constraint to estimate the object density. Indeed, NETSCAN first computes dense paths in the network based on the moving object count, then, it clusters the sub-trajectories which are similar to the dense paths. The user can adjust the clustering result by setting a density threshold for the dense paths, and a similarity threshold within the clusters. This paper describes the proposed method. An implementation is reported, along with experimental results that show the effectiveness of our approach and the flexibility allowed by the user parameters.

Keywords: Spatial data mining, clustering algorithm, similarity measure, moving objects database, road traffic analysis.

1. Introduction

Trajectory database management is a relatively new topic of database research, which has emerged due to the profusion of mobile devices and positioning technologies like GPS or recently the RFID (Radio Frequency Identification). Trajectory similarity search forms an important class of queries in trajectory databases. Beyond querying such complex data, new problems motivate research on the management of moving objects in general and on the spatiotemporal data mining in particular. The clustering of trajectories is part of this research.

We advocate that discovering similar sub-trajectories density based on the network is very useful. There are many examples in real applications. We present hereafter three application scenarios.

1. Knowledge and prediction of the road traffic: Given that the numbers of vehicles increases on the roads, information related to the density on the network becomes very useful for many purposes as navigation, trip planning, etc.
2. Car-sharing: In these last years, the massive use of the private means of transport caused many problems, namely the pollution and also the raising of oil prices. Car-sharing appears as an interesting alternative. Identifying the similar trajectories or even sub-trajectories becomes very useful for such types of applications.
3. Transport planning: At the moment of its creation, each road is planned for certain utilization. Reporting trajectory groups allows assessing the suitability of the road infrastructure with its actual use.

Generally speaking, clustering is a data mining technique extensively used in applications like market research, financial analysis or pattern recognition from images, to name but a few. Several types of clustering algorithms have been proposed among which K-Means (Lloyd, 1981), BIRCH (Zhang et al., 1996), DBSCAN (Ester et al., 1996) and OPTICS (Ankerst et al., 1999). Recent researches on trajectory clustering uses these algorithms while adapting them to the studied domain (Lee et al., 2007), since trajectories are complex objects.

We borrow the idea of density based algorithms such as DBSCAN, and adapt it to trajectories. The key idea behind our approach is that the knowledge of traffic density on the network would allow guiding the clustering of trajectories. We propose a two-step approach. In a first step, we define the similarity between the road segments and use it to group them in dense paths. In a second step, we propose a similarity measure between trajectories, and then we use it to make up the trajectory clusters around the dense paths. As in Lee et al., (2007), the time factor is relaxed in our approach. Nevertheless, we take account of the trajectory orientation. Another feature is that it regroups sub-trajectories rather than the whole of trajectories. Thus, a trajectory can belong to several clusters.

In summary, the contributions of this paper are as follows:

- We propose an innovative and effective method for network constraint trajectory clustering based on the network density.
- We define new similarity functions.
- We implement this framework and conduct an extensive experimentation that validates the method and shows its usefulness.

The rest of this paper is organized as follows. In section 2, we will detail a state of the art on the similarity and the clustering of the trajectories. We will explain our clustering approach in section 3. We will present in section 4 the first phase of the algorithm - named NETSCAN - for the clustering of the road segments. We will describe the second phase of the algorithm afterwards. In section 5 we will present our experimental results. Finally, we will conclude this article in section 6 and will propose some tracks for the pursuit of this research.

2. Related work

Research on the clustering of moving objects trajectories is closely connected to three topics: trajectories representation, similarity and clustering algorithms. In an orthogonal manner, we also distinguish the following criteria: the aspect of either constraint or free movement of the trajectory, the temporal aspect, the respect of the movement orientation, and finally, the grouping of sub-trajectories or entire trajectories in the clusters. This section describes the main works related to these three topics while situating them in relation with the above criteria.

Concerning the first research topic, many studies have investigated ways that the trajectory of a moving object can be represented. It can be geometric as in Lee et al., (2007) or symbolic as in Hadjieleftheriou et al., (2005). Indeed, if we know in advance the geometry and the topology of the network, we can represent a trajectory by the list of traversed segments, and alternatively, along with the instant to which the object passed from a segment to another, if we respect the temporal aspect. This representation is very precise at the spatial level, but maybe less precise at the temporal level. Nevertheless, it can be sufficient in many cases and especially in our context where the time is relaxed.

Regarding the works related to the similarity of moving objects trajectories, we first mention those in the free moving trajectory context, and then for constrained trajectories. Yanagiswa et al. (2003) focused on the extraction of the individual moving patterns of each object from the trajectories considering both time and location. Their approach uses the shape similarity between lines to retrieve required objects. Shim and

Chang (2003) considered the similarity of sub-trajectories and proposed a distance ‘K - Warping’ algorithm. Lin et al. (2005) focused on the spatial shapes and compared spatial shapes of moving object trajectories by developing algorithms for evaluating OWD (One Way Distance) in both continuous and discrete cases of the trajectories for similarity search. We also find similar approaches in Valachos et al. (2003), Sakurai et al. (2005), and Chen et al. (2005). Valachos et al. (2002) presented an investigation for analysis of spatio-temporal trajectories for moving objects where data contain a great amount of outliers. Therefore, they propose the use of a non metric distance function that is based on the Longest Common Sub Sequences (LCSS) algorithm in conjunction with a Sigmoid Matching function to increase the performance of Euclidean and Time Warping Distance. Zeinalipour-Yazti et al. (2006) introduce a distributed spatiotemporal similarity search based on the LCSS distance measure and propose two new algorithms offering good performances.

All these methods are inappropriate for similarity calculation on road networks since they use the Euclidian distance as a basis rather than the real distance on the road network. It is this last point that motivated the proposition of Hwang et al. (2005) that were the first to propose a similarity measure based on the spatiotemporal distance between two trajectories using the network distance. The algorithm of similar trajectory search consists of two steps: a filtering phase based on the spatial similarity on the road network, and a refinement phase for discovering similar trajectories based on temporal distance. Tiakas et al. (2006) and Chang et al. (2007) also use the same spatiotemporal distance, based on the road network, in their algorithm of similar trajectory search.

Concerning the works on trajectory clustering, we mention the two next ones: Gaffney and Smyth (1999) sustained that the vector based trajectory representation is inadequate in several cases. To surmount this problem, they introduce a model of probabilistic regression mixtures and show how the EM algorithm could be used in trajectory clustering. This approach considers non constraint trajectories. Moreover, it groups similar trajectories as a whole, thus ignoring similar sub-trajectories.

Lee et al. (2007) propose an algorithm named TRACLUS that groups similar sub-trajectories into a cluster. It consists of two phases: the partitioning of trajectories as line segments and then, the grouping of these according to their similarities. Nevertheless, this work always supposes free moving objects and, moreover, the time is not considered in this work.

As previously mentioned the majority of existing methods for trajectory similarity assumes that objects can move anywhere in the underlying space, and therefore do not support motion constraints. Even works which deal with searching similar trajectories of moving objects in a spatial network, group similar trajectories as a whole. However they could miss common sub-trajectories on one hand. On the other hand, the similarity measure adopted in these works assume that, to characterize two similar trajectories, it is not necessary to share common road segments, therefore the similarity measure must take into account the closeness of the trajectories. This assumption is not appropriate for a particular type of applications like care-sharing. In order to overcome the inefficiency of the previously described methods we propose a new clustering method that groups network constrained sub-trajectories based on a similarity measure which calculates the rate of inclusion of a trajectory compared to a dense path on the network. Our method is time relaxed.

3. The Clustering Procedure

The clustering consists of creating from a database, groups of similar objects (Han and Kamber 2006). By the clustering procedure, we mean, the steps to be performed ranging from getting a sample of data to processing the results obtained after partitioning the data space into clusters. Typically, the clustering procedures are composed of the following steps: (i) data representation; (ii) defining a similarity criteria, (iii) clustering, (iv) data abstraction (v) quality clustering evaluation.

We propose a two-steps clustering approach. The first phase allows grouping the road sections. Generally, we speak about section clustering. The second phase performs concretely the clustering of the trajectories. Because these kinds of clustering are relative to complex objects, we have to precise for each step; the representation, the similarity and the specific clustering algorithm.

3.1 Data Representation

The network representation is given by the road sections set. Hence, knowing the trajectory set, we compute[1] a transition matrix relative to road network (cf. Fig. 1). This matrix give statistics about the passages across the cross-roads and the turning movements, while reporting the number of moving objects transiting from one section to another connected one. This matrix will be denoted thereafter M and a node M(i,j) will denote the occurrence of moving objects traversing from section Si to section Sj.

[1] Alternatively, one may use traffic data produced by embedded sensors in the road or cameras.

$$\begin{array}{c|ccc} & S_1 & S_2 & S_3 \\ \hline S_1 & 0 & 10 & 0 \\ S_2 & 70 & 0 & 85 \\ S_3 & 0 & 90 & 0 \end{array}$$

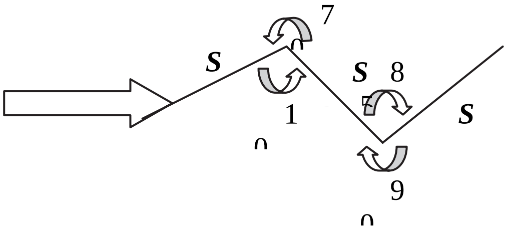


Fig. 1. Transitions Matrix assigned to the road network**.**

We adopt a symbolic representation of the trajectories (Du Mouza and Rigaux, 2004), (Savary et al., 2004). In this representation, moving objects appear as a sequence of symbols where each one refers to one road section.

$$TR=<S_{i1}, \ldots\ldots\ldots\ldots, S_{in}>$$

The symbols order shows the movement direction.

3.2 Similarity Measure

The similarity is the base of the clustering operation. We define the similarity at two levels. At the network level, the similarity is computed between two transitions as the difference of their density values. This measure concerns only the consecutive transitions.

$$\text{Sim\_segment (M(i,j)} = |M(i,j) - M(j,k)| \qquad (3.2.1)$$

At the trajectories level, we define a similarity measure between two trajectories where one is the reference. This measure reflects the resemblance to an object and it is not symmetric. It allows comparing the effective trajectories to a fictive type trajectory. To this end, the similarity is computed as the report between the common length among a trajectory and the reference from one side, and the length of the reference trajectory from another side.

$$\text{(traj\_ref) Lenght / (part\_common).Lenght = traj\_Sim} \qquad (3.2..2)$$

For the works mentioned in section 2, the similarity relies upon the Euclidian distance and/or shapes (Yanagiswa et al., 2003). This comes from the fact, that they represent the trajectories by their geometry and their forms. Our work adopts quite different criteria because it is situated in the constraint context. It uses the available information relative to the network density form one side, and the symbolic representation of the trajectories allows obtaining sequence similarities as in Chen et al. (2005), from the other side.

4. Two-stepClustering Algorithme NETSCAN

The clustering step corresponds effectively to the grouping phase and aims at deriving a database partitioning as relevant as possible. To achieve this goal, we propose a two-step clustering algorithm that we call NETSCAN. At the first step, it finds the most dense road sections, and merges them to form dense paths on the road network. The second step permits to classify the trajectories of moving objects according these dense paths. The figure 2 reports the main steps of NETSCAN. Our algorithm share the same features with the DBSCAN algorithm and it is based on two steps:

1. Section gathering: find the network paths that are the densest in terms of moving objects transiting on them.
2. Trajectory gathering: for each path, gather the trajectory similar to it.

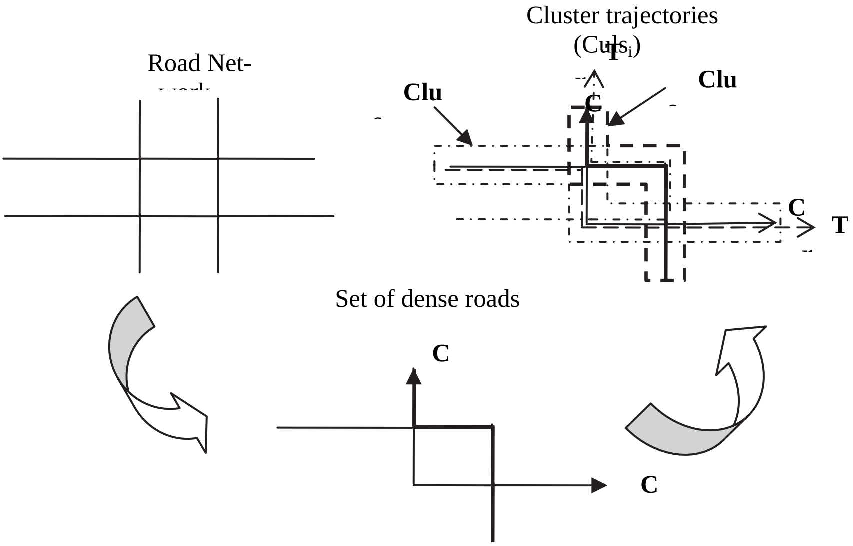


Fig. 2. Trajectory clustering Example by NETSCAN Algorithm

4.1 Segment Clustering

The first part of the proposed algorithm, NETSCAN-PHASE 1, which is described here, performs the segment clustering.

NETSCAN Algorithm - PHASE 1 /* Dense path discovery*/

Input:

- Set of road segments $S = \{S_1, S_2, \ldots, S_{no\_segments}\}$
- Transition matrix M.
- Threshold ε -- maximal density difference between neighbour segments.
- Threshold α -- minimal required density for a transition.

Output:

- Ordered dense path set O = < C1, C2, … , Cno_paths >.

Algorithm:

```
O 🡐 ∅                    -- Initialisation
k 🡐 0
While there exists non marked transitions $M(i,j) >= \alpha$
    k 🡐 k+1
    M(d,f) = max (M (i, j))
    $C_k$ 🡐 <$S_d$, $S_f$>        -- generate a new dense path from this transition
    Mark the transition M(d,f)
    While there exists u such as $M(f,u) >= \alpha$ and u not marked
      -- forward extension
        Select M(f, f_succ) such as |M(d,f) – M(f,u)| is minimum
            If |M(d,f) – M(f,_succ)| <= ε
                Insert last (Ck, Sf_succ )
                Mark M(f,  f_succ)
                d 🡐 f; f 🡐 f_succ              -- extend path
            End If
      End While
      While there exists u such as $M(u,d) >= \alpha$ and u not marked
      -- backward extension
            Select M(d_prec, d) such as |M(d,f) – M(u,d)| is minimum
            If |M(d,f) – M(u,d)| <=  ε
                Insert first (Ck, Sd_préd)
                Mark M(d_prec, d)
                f 🡐 d; d 🡐 d_prec              -- extend path
            End If
       End While
       O =  O ∪ Ck
```

```
End While
Return O
    ------------------------------------------------------------
```

Fig. 3. NETSCAN Algorithm–Phase 1: clustering segments as dense paths.

This phase is inspired from the density based clustering principle introduced with the DBSCAN algorithm (Ester et al., 1996), while applying it to road segments. It takes as input the set of segments that constitute the road network, the transition matrix (as in section 3.1), a density threshold α and a similarity threshold between the transition densities ε. In this phase, the algorithm firstly finds the dense transitions, i.e. those having maximum value. Afterwards, for each dense transition, it groups the connected segments and transitions that have similar densities, thus creating dense paths.

The process begins with the transition having the maximal density. Following, it begins searching the connected transitions in both ways in order to find those with a density ε near to the maximal density. To insure the non reuse of transitions that are included in dense paths, they are marked at the first assignment.

The extension of a dense path is done in both ways if the constraints are verified, i.e., the candidate transition is only marked if it respects the density (α) and similarity (ε) thresholds. The obtained segment clusters correspond to the densest paths in the network. The figure 3 presents the first phase of the NETSCAN algorithm.

The dense paths are represented as a sequence of segments, the same as with the trajectories (section 3.1). Each segment is identified by an associated symbol.

4.2 Trajectory Clustering

In this section, we present the second phase of the NETSCAN method, which corresponds to the trajectory clustering. This part uses the results obtained in the first phase as presented in the above section. Indeed, the dense paths are considered as natural cluster centres for the MO trajectories.

The trajectory clustering algorithm consists in grouping the trajectories according to their similarity to each dense path generated in phase 1 of NETSCAN. We use the similarity measure as defined in section 3.2. The input of the algorithm consists of the set of the dense paths, the set of trajectories and the threshold σ. For each dense path, it computes the similarity with each trajectory. If the similarity is above the threshold value, then the trajectory is kept in the cluster. More precisely, the common part between the trajectory and the dense path is added to the cluster. The number of returned clusters is equal to the number of dense paths.

```
---------------------------------------------
NETSCAN Algorithm - PHASE 2
---------------------------------------------
Input:
 -  Set of paths returned from NETSCAN algorithm - PHASE 1
        $O = < C_1, C_2, \ldots, C_{no\_paths} >$
 -  Set of MO trajectories $TR = < TR_1, TR_2, \ldots, TR_{no\_trajectories} >$
 -  Minimal similarity threshold $\sigma$

Output:
 -  Set of clusters $Clus = \{Clus_1, \ldots, Clus_{no\_paths}\}$

Algorithm:
For each path $C_i$ in O
    For each trajectory $TR_j$ in TR such as $TR_j$ overlaps $C_i$
        Compute $Sc = C_i \cap TRr_j$ -- set of common segments $Sc_k$
        Compute the sum of common segment lengths :
                $L_c = \sum_i length\ (Sc_i)$
        Compute the similarity between $C_i$ and $TR_j$ as:
                $Sim = L_c / length(C_i)$
        If Sim >= $\sigma$
            Add Sc to $Clus_i$
        End If
    End For
    Add $Clus_i$ to Clus
End For
Return Clus
-----------------------------------------------------------------
```

Fig. 4. NETSCAN Algorithm – Phase 2: Trajectory Clustering.

5. Experimental Evaluation

In this section, we evaluate the effectiveness of our trajectory clustering algorithm NETSCAN. We describe the experimental data and environment in section 5.1. We discuss the impact of parameter values in sections 5.2 and 5.3. The last section addresses the optimization issue.

5.1 Experimental Setting

NETSCAN algorithm has been implemented on a PC running under Windows XP Professional. The hardware configuration is as follows: a 2.0 GHz AMD Athlon ™ 64 Dual Core processor, 1.5 GB main memory, and 80 GB HDD. We use Oracle 10g as data server.

The trajectories may be obtained from several sources, such as Floating Car Data (FCD). However, it is preferable to use for validation and test a public data source. In the context of constraint moving objects, the generator developed by Brinkhoff is the mostly used for benchmarking and test (Brinkhoff, 2002). Based on the road network of San Joaquin bay which consists of 18496 nodes and 24123 edges (i.e. road sections), we apply the generator to produce 2064 trajectories of moving objects on the road network. The figure bellow shows the road map of San Joaquin on the left, and the locations of different moving objects displayed from the generated (virtual) GPS log.

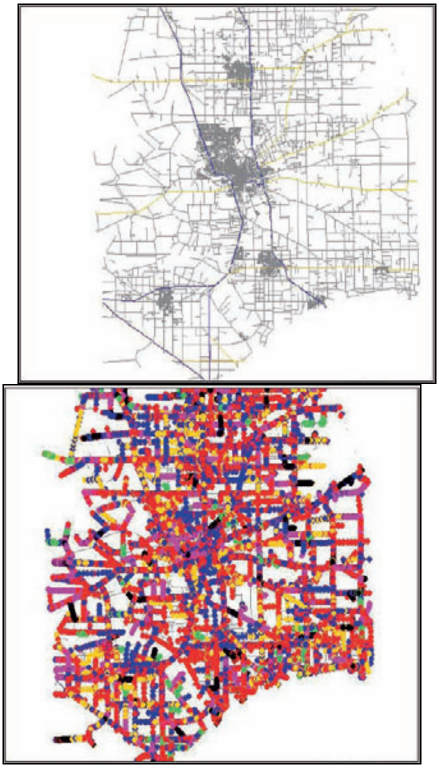


Fig. 5. The road map of San Joaquin.

Based on the above information, it becomes possible to compute the density matrix appropriate to this network. Precisely, for each transition (i,j) in this matrix, we count the occurrences of moving objects traversing it. Figure 6 shows the distribution of the transition densities (i.e. the distribution of the number of moving object going from section i to section j). As shown here, there are 225 transitions that contain 10 moving objects. As one can expect, the number of transitions having a dense traffic (a high transition value, e.g. transition value = 243) is very limited compared to those where it is less dense or either null. Notice that we did not draw the values under 10 for the transition value because the corresponding transitions where too numerous (over 2000), which would render the curve less significant.

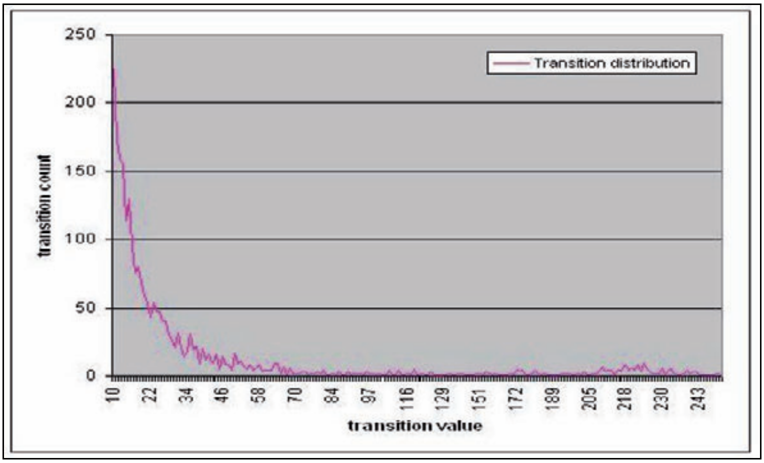


Fig. 6. Transition distribution

5.2 Experimental Results – Phase 1

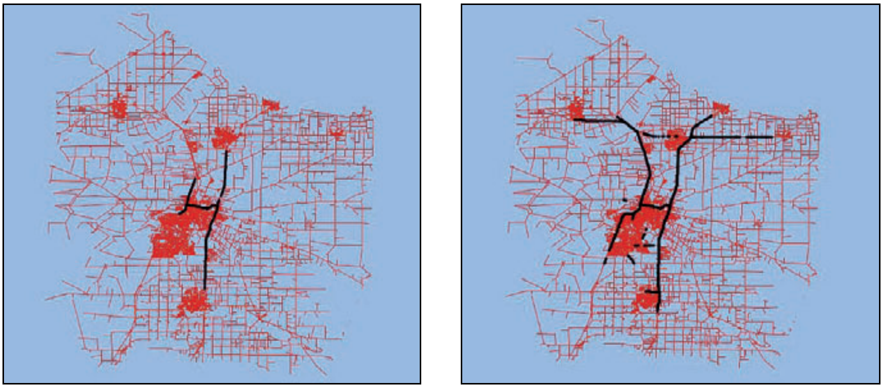
To test the first phase of NETSCAN algorithm, we change the input parameter specifying the density threshold α. We successively test the first phase by setting the following values for α: 10, 50, 100, 200, and 230. We measure the impact of those parameters on the results.

For each test, we evaluate the number of dense transitions, i.e. where the density is above the threshold α, as well as the number of dense paths resulting from Phase 1 of NETSCAN. The experimentation results are summarized in table 1. As an example, for α = 200, there are 145 dense transitions, and the algorithm generates 91 dense paths. Figure 7.a visualizes 91 dense paths on the road network of San Joaquin for α = 200, while Figure 7.b shows 239 dense paths when α = 50.

Actually, the advantages of the phase 1 of NETSCAN is, on one hand the computation of dense paths that facilitate the decision making, and on other hand an initialization of the clusters by aggregating trajectories around the obtained dense paths. This is achieved in the second phase of NETSCAN.

Table. 1. Impact of the density threshold

α	#dense_paths	#transitions >α	σ	Min (#traj. /cluster)	Mean (#traj. /cluster)	Max (#traj. /cluster)
10	1303	2231 (5,7%)	0.5	11	74,82	486
50	239	416 (1,06 %)	0.7	38	147,84	261
100	143	252 (0.64 %)	0.7	97	205,01	267
200	91	145 (0.37 %)	0.7	209	233,868	267
230	27	37 (0.09 %)	0.7	239	248,92	267



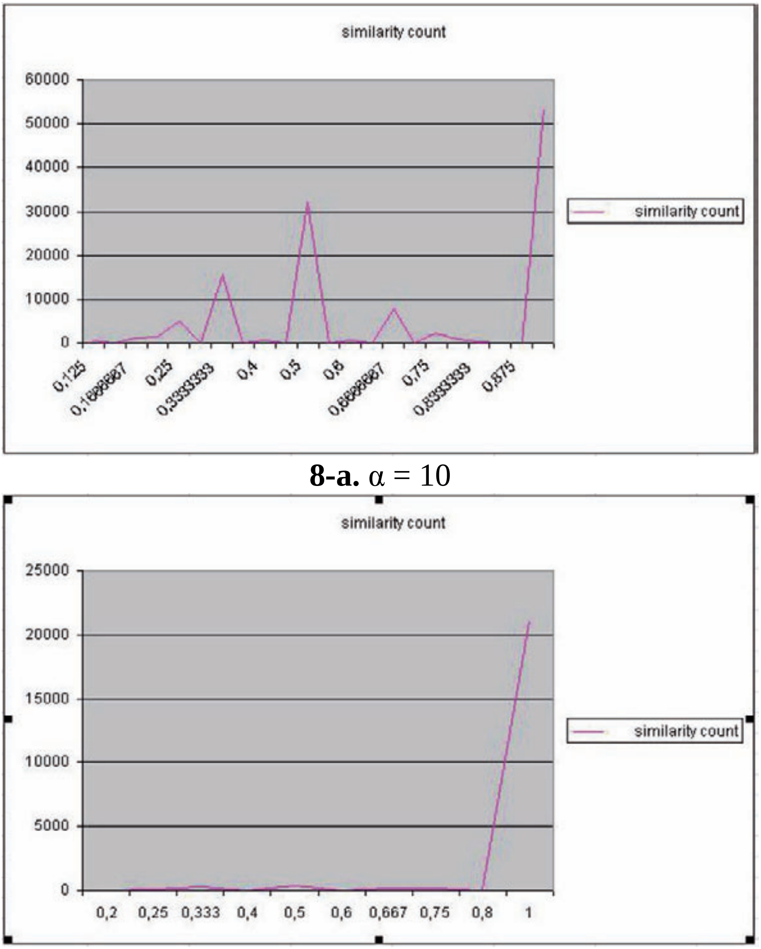
a- α =200 **b-** α = 50

Fig. 7. Mapping Dense Paths

5.3 Experimental Results – Phase 2

In this test, we evaluate the impact of the similarity threshold σ. When σ is equal to 1, this means that the cluster only groups the trajectories including the whole dense path. The other cases correspond to a degree of similarity.

This value mainly impacts the number of trajectories belonging to each cluster. As the similarity threshold increases, the number of trajectory tends to decrease and vice versa. In table 1 above, we measure the minimun, the average, and the maximum number of trajectories by cluster.



8-a. α = 10

8-b. α = 200

Fig. 8. Tuning the similarity threshold

In order to tune the similarity threshold σ, we provide the user another distribution curve as in Figure 8 bellow. This last draws the number of trajectories corresponding to each similarity value between trajectories and (intersecting) dense paths. As an example, as shown in Figure 8.a., for α = 10, many trajectories include half to 80% of the dense paths. However, in

Figure 8.b. where α = 200, most trajectories include the whole dense paths or contain at least 80% of their length, that is no need to choose a similarity threshold under 0.8.

5.4 Optimization

This algorithm has been first implemented without any optimization. The algorithm, mainly in Phase 2, performed too slowly (in many hours). Indeed, the trajectory database was scanned as many times as the number of dense paths. Since symbolic data representation adapts to relational database, it was possible to use conventional query optimisation techniques such as indexing. Doing so, we improved the performances to a quasi real time processing, even in the case of a great volume of data.

Concerning the first phase, the optimisation has concerned the storage costs. In fact, the size of the density matrix is theoretically the square of the number of sections (e.g. 24123^2 here). But this is a very sparse matrix. Therefore, we choose to store it in a table as (i, j, #objects) where only the transitions having non null values (#objects >0) are materialized. This reduces the storage size drastically (39130 transitions in our case).

6. Conclusion

This paper deals with spatiotemporal data mining. More precisely, it addresses the problem of moving object clustering and adapts it to network constrained moving objects. We have proposed a two-step clustering algorithm. The first one focuses on the road sections and allows obtaining the densest paths all over the network. The second step processes the trajectories in order to obtain similar trajectories classes.

The proposed algorithms have been implemented, optimized, and tested using simulated moving objects. The experimental results have allowed a preliminary test in order to validate our approach.

As future work, we aim at applying our approach to real datasets in order to validate it by experts in traffic management. Thereafter, we will study the extension to spatio-temporal clustering. Another promising issue is the combination of spatio-temporal clustering and previous proposals in spatio-temporal OLAP, as proposed by Savary et al. (2004) and Wan and Zeitouni (2006). Finally, we will explore this mining task for moving objects equipped by sensors, such as pollution or temperature sensors. To this end, we may keep our symbolic representation, i.e. by reference to the road segments, and extend it to capture the measure range. This implies the extension of the transition matrix with the temporal dimension and measures. The similarity also should be extended and the algorithms should be adapted and optimized.

Acknowledgements

The authors would like to thank Benjamin Nguyen for his reading and discussions about the ideas of this paper. This work has been partly supported by grants from Région Ile-de-France.

References

Ankerst M., M. M. Breunig, H.-P. Kriegel and J. Sander (1999), OPTICS: Ordering Points to Identify the Clustering Structure, In Proc. 1999 ACM SIGMOD Int'l Conf. on Management of Data, Philadelphia, Pennsylvania, pp. 46-60.

Brinkhoff T., A Framework for Generating Network-Based Moving Objects, GeoInformatica, Vol. 6, No. 2, Kluwer, 2002, 153-180

Chang J-W., R. Bista, Y-C. Kim and Y-K Kim (2007) Spatio-temporal Similarity Measure Algorithm for Moving Objects on Spatial Networks. ICCSA 2007, pp.1165-1178.

Chen L., M.T. Ozsu and V. Oria (2005), Robust and Fast Similarity Search for Moving Object Trajectories. In: ACM SIGMOD, pp. 491-502. ACM Press, New York.

Du Mouza C. and P. Rigaux. Mobility Patterns. *In Proc. Intl. Workshop on Spatio-temporal Databases (STDBM'04).*

Ester M., H.-P. Kriegel, J. Sander and X. Xu (1996) A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, In Proc. 2nd Int'l Conf. on Knowledge Discovery and Data Mining, Portland, Oregon, pp. 226-231.

Gaffney S. and P. Smyth, (1999) Trajectory Clustering with Mixtures of Regression Models, In Proc. 5th ACM SIGMOD Int'l Conf. on knowledge Discovery and Data Mining, San Diego, California, pp. 63-72.

Hadjieleftheriou M., G. Kollios, P. Bakalov, V. Trotras (2005) Complex Spatio-Temporal Pattern Queries. In Proc. of the 31st VLDB Conference.

Han J. and M. Kamber (2006) Data Mining: Concepts and Techniques, 2nd ed., Morgan Kaufmann.

Hwang J-R., H-Y. Kang and K-J. Li (2005) Spatio-temporal Analysis Between Trajectories on Road Networks. ER Workshops 2005, LNCS 3770, pp. 280-289.

Lee J-G, J. Han and K-Y. Whang (2007) Trajectory Clustering: A Partition-and-Group Framework. In Proc.SIGMOD'07, Beijing, China.

Lin B., J. Su (2005) Shapes Based Trajectory Queries for Moving Objects. GIS, pp. 21-30.

Lloyd S. (1982) Least Squares Quantization in PCM, IEEE Trans. on Information Theory, 28(2): 129-137.

Sakurai Y., M. Yoshikawa and C. Faloutsos (2005) FTW: Fast Similarity Search under the Time Warping Distance. In: PODS, pp. 326-337.

Savary L., Wan T., Zeitouni K., Spatio-Temporal Data Warehouse Design for Activity Pattern Analysis, DEXA Workshop on Geographic Information Management, September, 2004, Zaragoza, Spain, pp. 814-818.

Shim C-B and J-W Chang (2003) Similar Sub-Trajectory Retrieval for Moving Objects in Spatiotemporal Databases. In: Proc. of the 7th EECADIS, pp.308-322.

Tiakas E., A. N. Papadopoulos, A. Nanopoulos and Y. Manolopoulos (2006) Trajectory Similarity Search in Spatial Networks. In : Proc. of the 10th IDEAS, pp. 185-192.

Vlachos M., D. Gunopulos and G. Kollios (2002) Robust Similarity Measures of Mobile Object Trajectories. In: Proc. of the 13 th Intl. Workshop on DEXA, IEEE Computer Society Press, Los Alamitos pp. 721-728.

Vlachos M., G. Kollios and D. Gunopulos (2002) Discovering Similar Multidimensional Trajectories. In: Proc. Of the 18th ICDE. IEEE Computer Society Press, Los Alamitos pp. 673-684.

Wan T. and Zeitouni K., An OLAP System for Network-Constraint Moving Objects, ACM SAC 2007, The 22nd Annual ACM Symposium on Applied Computing, Seoul, Korea, March, 2007, pp. 13-18.

Yanagisawa Y., J. Akahani, T. Satoch (2003) Shape-Based Similarity Query for Trajectory of Mobile Objects. In : Proc. Of the 4th Intl. Conf. On MDM, pp. 63-77.

Zeinalipour-Yazti D., S. Song Lin, D. Gunopulos (2006) Distributed Spatio-Temporal Similarity Search. CIKM, pp. 14-23.

Zhang T., R. Ramakrishnan, and M. Livny (1996) BIRCH: An Efficient Data Clustering Method for Very Large Databases. In Proc. ACM SIGMOD Int’l Conf. on Management of Data, Montreal, Canada, pp. 103-114.

Author Index