Lecture Notes
in Geoinformation and Cartography

LNG&C

Arnold Bregt
Tapani Sarjakoski
Ron van Lammeren
Frans Rip *Editors*

Societal Geo-innovation

Selected Papers of the 20th AGILE Conference on Geographic Information Science

Springer

Lecture Notes in Geoinformation and Cartography

The Lecture Notes in Geoinformation and Cartography series provides a contemporary view of current research and development in Geoinformation and Cartography, including GIS and Geographic Information Science. Publications with associated electronic media examine areas of development and current technology. Editors from multiple continents, in association with national and international organizations and societies bring together the most comprehensive forum for Geoinformation and Cartography.

The scope of Lecture Notes in Geoinformation and Cartography spans the range of interdisciplinary topics in a variety of research and application fields. The type of material published traditionally includes:

- proceedings that are peer-reviewed and published in association with a conference;
- post-proceedings consisting of thoroughly revised final papers; and
- research monographs that may be based on individual research projects.

The Lecture Notes in Geoinformation and Cartography series also includes various other publications, including:

- tutorials or collections of lectures for advanced courses;
- contemporary surveys that offer an objective summary of a current topic of interest; and
- emerging areas of research directed at a broad community of practitioners.

More information about this series at http://www.springer.com/series/7418

Arnold Bregt · Tapani Sarjakoski
Ron van Lammeren · Frans Rip
Editors

Societal Geo-innovation

Selected Papers of the 20th AGILE Conference on Geographic Information Science

Springer

Editors

Arnold Bregt
Laboratory of Geo-information Science
and Remote Sensing
Wageningen University & Research
Wageningen
The Netherlands

Ron van Lammeren
Laboratory of Geo-information Science
and Remote Sensing
Wageningen University & Research
Wageningen
The Netherlands

Tapani Sarjakoski
Geoinformatics and Cartography
Finnish Geospatial Research Institute
Masala
Finland

Frans Rip
Laboratory Geoinformation and Remote
Sensing/GeoDesk
Wageningen University & Research
Wageningen, Gelderland
The Netherlands

ISSN 1863-2246 ISSN 1863-2351 (electronic)
Lecture Notes in Geoinformation and Cartography
ISBN 978-3-319-56758-7 ISBN 978-3-319-56759-4 (eBook)
DOI 10.1007/978-3-319-56759-4

Library of Congress Control Number: 2017936442

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Societal Geo-innovation is the overarching theme of the 20th AGILE Conference on Geographic Information Science, held in 2017 at Wageningen University & Research in Wageningen, The Netherlands. This theme reflects the importance of geographic information science in support of a large variety of societal challenges, such as urbanisation, food security, water scarcity, health quality, energy transition and climate adaptation. Geographic information science plays a crucial role in agenda setting, analysis and solutions regarding these challenges. The instrumental role of geo-information for realising the 17 United Nations sustainable development goals to transform our world is also widely acknowledged.

Over the years, the contribution of geographic information science to our society has changed both in content and context. Traditionally, the content was strongly data-centric. Methods, techniques and associated quality measures for data collection and storage were key research and practical challenges. Currently, data analysis and algorithms for automated information extraction receive ample attention. This trend is also reflected in the papers submitted to 20th AGILE Conference. Many of them deal with new analysis techniques and methods for spatio-temporal data. Hence, a special part in this book on ***Spatio-Temporal Analysis***.

Another notable change is the context of our discipline. Traditionally, geo-information professionals determine "what, how and why" spatial data is collected. The perception of space was mainly based on the view of the geo-information scientist and professional. This has changed in the last decade. Due to easier to use technology and interest in their environment, citizens became conscious as well as unconscious collectors of spatial data. This has resulted in a new research focus to address questions like "how can we utilise 'citizenized' spatial data?" Given this new focus, we noticed that even the perception of the space arena becomes wider.

Nowadays, multiple perceptions of the same location (space and place concept) are more common than before. Quite some papers submitted to the conference deal with the citizens' view on space; therefore, the other part of the book is about ***Spatio-Temporal Perception***.

The last 20 years, the Association of Geographic Information Laboratories for Europe (AGILE) organised conferences and associated activities acted as a mirror for the developments in our field. Hardy Pundt and Fred Toppen reflect on these developments in the last part of this book: ***20 Years of AGILE***. They reflect on AGILE as an organisation, its members, activities and research themes. We highly appreciate their contribution: a historical account of AGILE developments.

The organisation of the 20th AGILE conference, and also the publication of associated scientific papers in this book, was only possible thanks to the involvement of enthusiastic individuals and organisations. Local organisers, sponsors, authors, presenters, reviewers, AGILE council members and AGILE participants, thank you all for your valuable and indispensable contribution.

Wageningen, The Netherlands/Masala, Finland
March 2017

Arnold Bregt
Tapani Sarjakoski
Ron van Lammeren
Frans Rip

Organizing Committee

Scientific Programme Committee

Arnold Bregt, Wageningen University & Research, The Netherlands (chair)
Ron van Lammeren, Wageningen University & Research, The Netherlands
Tapani Sarjakoski, Finnish Geospatial Research Institute, Finland

Local Organizing Committee

Jandirk Bulens (workshop chair)
Ron van Lammeren (chair)
Arend Ligtenberg
Suze Looyschelder
Frans Rip (submissions manager and editor)
Karin Schipper
Antoinette Stoffers
Wies Vullings
Claudius van de Vijver

Scientific Programme Committee

Ana Paula Afonso, University of Lisbon, Portugal
Fernando Bacao, New University of Lisbon, Portugal
Marek Baranowski, Institute of Geodesy and Cartography, Poland
Melih Basaraner, Yildiz Technical University, Turkey
Giedrė Beconytė, Vilnius University, Lithuania
Itzhak Benenson, Tel Aviv Unversity, Israel

Lars Bernard, Technical University Dresden, Germany
Michela Bertolotto, University College Dublin, Ireland
Ralf Bill, Rostock University, Germany
Sandro Bimonte, IRSTEA, France
Thomas Blaschke, University of Salzburg, Austria
Arnold Bregt, Wageningen University and Research, The Netherlands
Thomas Brinkhoff, Jade University Oldenburg, Germany
Pedro Cabral, New University of Lisbon, Portugal
Sven Casteleyn, University Jaume I of Castellon, Spain
Christophe Claramunt, Naval Academy Research Council, France
Serena Coetzee, University of Pretoria, South Africa
Lex Comber, University of Leeds, United Kingdom
Joep Crompvoets, KU Leuven, Belgium
Isabel Cruz, University of Illinois at Chicago, United States
Sytze de Bruin, Wageningen University and Research, The Netherlands
Cidalia Fonte, University of Coimbra, Portugal
Anders Friis-Christensen, European Commission, Joint Research Centre, Italy
Jerome Gensel, University of Grenoble, France
Michael Gould, Esri and University Jaume I, Spain
Carlos Granell, University Jaume I of Castellón, Spain
Henning Sten Hansen, Aalborg University, Denmark
Lars Harrie, Lund University, Sweden
Francis Harvey, University of Leipzig, Germany
Roberto Henriques, New University of Lisbon, Portugal
Gerard Heuvelink, Wageningen University and Research, The Netherlands
Stephen Hirtle, University of Pittsburgh, United States
Hartwig Hochmair, University of Florida, United States
Joaquín Huerta, University Jaume I of Castellon, Spain
Bashkim Idrizi, Mother Teresa University, Republic of Macedonia
Mike Jackson, University of Nottingham, United Kingdom
Bin Jiang, University of Gävle, Sweden
Didier Josselin, University of Avignon, France
Derek Karssenberg, Utrecht University, The Netherlands
Tomi Kauppinen, Aalto University, Finland
Marinos Kavouras, National Technical University of Athens, Greece
Dimitris Kotzinos, University of Cergy-Pontoise, France
Petr Kuba Kubicek, Masaryk University, Czech Republic
Patrick Laube, Zurich University of Applied Science, Switzerland
Robert Laurini, University of Lyon, France
Francisco J. Lopez-Pellicer, University of Zaragoza, Spain
Malgorzata Luc, Jagiellonian University, Poland
Ali Mansourian, Lund University, Sweden
Bruno Martins, University of Lisbon, Portugal
Filipe Meneses, University of Minho, Portugal
Peter Mooney, Maynooth University, Ireland

João Moura Pires, New University of Lisbon, Portugal
Beniamino Murgante, University of Basilicata, Italy
Javier Nogueras-Iso, University of Zaragoza, Spain
Juha Oksanen, Finnish Geospatial Research Institute, Finland
Toshihiro Osaragi, Tokyo Institute of Technology, Japan
Frank Ostermann, University of Twente, The Netherlands
Volker Paelke, Hochschule Ostwestfalen-Lippe, Germany
Marco Painho, New University of Lisbon, Portugal
Petter Pilesjö, Lund University, Sweden
Poulicos Prastacos, FORTH, Greece
Hardy Pundt, Harz University of Applied Sciences, Germany
Ross Purves, University of Zurich, Switzerland
Viktor Putrenko, National Technical University of Ukraine, Ukraine
Martin Raubal, ETH Zürich, Switzerland
Wolfgang Reinhardt, Bundeswehr University Munich, Germany
Claus Rinner, Ryerson University, Canada
Jorge Rocha, University of Minho, Portugal
Armanda Rodrigues, New University of Lisbon, Portugal
Maribel Yasmina Santos, University of Minho, Portugal
Tapani Sarjakoski, Finnish Geospatial Research Institute, Finland
L. Tiina Sarjakoski, Finnish Geospatial Research Institute, Finland
Sven Schade, European Commission—DG JRC, Belgium
Christoph Schlieder, University of Bamberg, Germany
Monika Sester, Leibniz University of Hanover, Germany
Takeshi Shirabe, Royal Institute of Technology, Sweden
Jantien Stoter, Delft University of Technology, The Netherlands
Maguelonne Teisseire, IRSTEA, France
Fred Toppen, Utrecht University, The Netherlands
Nico Van de Weghe, Ghent University, Belgium
Ron van Lammeren, Wageningen University and Research, The Netherlands
Jos van Orshoven, KU Leuven, Belgium
Danny Vandenbroucke, KU Leuven, Belgium
Lluis Vicens, University of Girona, Spain
Luis M. Vilches-Blázquez, Pontifical Xavierian University, Spain
Kirsi Virrantaus, Aalto University, Finland
Vít Voženílek, Palacky University Olomouc, Czech Republic
Monica Wachowicz, University of New Brunswick, Canada
Gudrun Wallentin, University of Salzburg, Austria
Robert Weibel, University of Zurich, Switzerland
Stephan Winter, University of Melbourne, Australia
F. Javier Zarazaga-Soria, University of Zaragoza, Spain
Alexander Zipf, Heidelberg University, Germany

Contents

Part I Spatio-Temporal Perception

Investigating Representations of Places with Unclear Spatial Extent in Sketch Maps 3
Vanessa Joy A. Anacta, Mohammed Imaduddin Humayun, Angela Schwering and Jakub Krukar

Mining Rainfall Spatio-Temporal Patterns in Twitter: A Temporal Approach 19
Sidgley Camargo de Andrade, Camilo Restrepo-Estrada, Alexandre C. B. Delbem, Eduardo Mario Mendiondo and João Porto de Albuquerque

Residential Choices as a Driving Force to Vertical Segregation in Whitechapel 39
Shlomit Flint

Reference Resolution for Pedestrian Wayfinding Systems 59
Jana Götze and Johan Boye

3D Building Maps for Everyone—Mapping Buildings Using VGI 77
Laura Knoth, Manfred Mittlboeck and Bernhard Vockner

Follow the Signs—Countering Disengagement from the Real World During City Exploration 93
Markus Konkol, Christian Kray and Morin Ostkamp

Perspectives in Externalizations of Mental Spatial Representations 111
H. Löwen, A. Schwering, J. Krukar and S. Winter

Personal Dimensions of Landmarks 129
Eva Nuhn and Sabine Timpf

Personal Activity Centres and Geosocial Data Analysis: Combining Big Data with Small Data . . . 145
Colin Robertson, Rob Feick, Martin Sykora, Ketan Shankardass and Krystelle Shaughnessy

Part II Spatio-Temporal Analysis

Spatio-Temporal Road Coverage of Probe Vehicles: A Case Study on Crowd-Sensing of Parking Availability with Taxis . . . 165
Fabian Bock, Yuri Attanasio and Sergio Di Martino

A Big Geo Data Query Framework to Correlate Open Data with Social Network Geotagged Posts . . . 185
Gloria Bordogna, Steven Capelli and Giuseppe Psaila

Towards Automatic Large-Scale 3D Building Reconstruction: Primitive Decomposition and Assembly . . . 205
Hai Huang and Helmut Mayer

Dynamic Transfer Patterns for Fast Multi-modal Route Planning . . . 223
Thomas Liebig, Sebastian Peter, Maciej Grzenda and Konstanty Junosza-Szaniawski

The Effect of Regional Variation and Resolution on Geosocial Thematic Signatures for Points of Interest . . . 237
Grant McKenzie and Krzysztof Janowicz

Distance Measure Based on Spatiotemporal Coexistence of Residents and Its Application to Urban Analysis . . . 257
Toshihiro Osaragi and Ayaka Murakami

When Granules Are not Enough in a Theory of Granularities . . . 279
Ricardo Almeida Silva, João Moura Pires and Maribel Yasmina Santos

Assessing Accuracy and Geographical Transferability of Machine Learning Algorithms for Wind Speed Modelling . . . 297
Fabio Veronesi, Athina Korfiati, René Buffat and Martin Raubal

On Measures for Groups of Trajectories . . . 311
Lionov Wiratma, Marc van Kreveld and Maarten Löffler

Beyond Pairs: Generalizing the Geo-dipole for Quantifying Spatial Patterns in Geographic Fields . . . 331
Rui Zhu, Phaedon C. Kyriakidis and Krzysztof Janowicz

Part III 20 Years of AGILE

20 Years of AGILE . . . 351
Hardy Pundt and Fred Toppen

Part I
Spatio-Temporal Perception

Investigating Representations of Places with Unclear Spatial Extent in Sketch Maps

Vanessa Joy A. Anacta, Mohammed Imaduddin Humayun, Angela Schwering and Jakub Krukar

Abstract This study analyzes different ways of representing vaguely defined places from a set of sketch maps specifically when used in giving route instructions. A total of 30 participants who are familiar with the study area were asked to sketch a route map consisting of pre-identified set of places. The task involved two groups: intra-city route and inter-city route. Sketch maps were analyzed using a previously developed classification scheme to investigate how places with unclear spatial extent are represented. These were then classified into different category of places: *district*, *site and neighborhood*. Results showed that labels and regular shapes are the most preferred, as opposed to other types of sketch representations, regardless of the category of place. It also occurred that a specific place can be classified under one or more categories, which influences the type of sketch representation used.

Keywords Sketch representation · Spatial vagueness · Place · Sketch map · Wayfinding

1 Introduction

When receiving wayfinding instructions from people (either visual or textual), we often are required to interpret imprecise information such as '*go towards the city center*', '*you'll find the place inside the university campus*', '*it is near the castle*'.

V.J.A. Anacta (✉) · M.I. Humayun · A. Schwering · J. Krukar
Institute for Geoinformatics, University of Muenster, Muenster, Germany
e-mail: v.anacta@uni-muenster.de

M.I. Humayun
e-mail: humayun@uni-muenster.de

A. Schwering
e-mail: schwering@uni-muenster.de

J. Krukar
e-mail: krukar@uni-muenster.de

A. Bregt et al. (eds.), *Societal Geo-innovation*, Lecture Notes
in Geoinformation and Cartography, DOI 10.1007/978-3-319-56759-4_1

In this case, we are faced with questions such as '*Where does the city center start*?', '*Which part in the university campus*?', '*What did that person mean when referring to the castle*?' Hence, we often encounter vagueness or even ambiguity in such spatial information. Vagueness arises due to poor definition of the object in question or the class of object (Fisher et al. 2006). Sometimes there are cases wherein a person would represent a place by also referring to other features surrounding it. In natural language, qualifiers such as tall or big are vague because of the presence of borderline cases where it is unclear how to classify them. This is also true of places, where it is not clear whether certain locations are part of vague vernacular regions such as city centers, whose boundaries generally are not crisp. Other kinds of vague references in spatial information include natural features such as 'mountains' or 'lakes'. These are characterized by unclear spatial extents and boundaries of the referents. Three distinct categories are of interest in spatial information (Bennett 2010), with the first two being relevant for this study:

- General descriptions of places, which use count nouns such as *downtown*, *marketplace*, *lake* which in many cases have unclear extents.
- Referenced places such as *harbor* or a university *campus* which are associated with specific space but exhibit similar problem with boundaries.
- Spatial relations such as *is near*, *in front of*, *along the* etc. commonly used in qualitative route descriptions.

There may be different methods of representing spatial vagueness but there is no perfect model of visualizing these places because they have their own set of advantages and disadvantages (Humayun and Schwering 2013). This study attempts to understand how people represent places with unclear spatial extent in conveying route instructions. One way to understand how people represent such places is through sketch mapping.

Sketches are used to visualize people's abstract representation of specific places or objects for both learning and communication (Voudouris et al. 2006). With common symbols, patterns and strategies used, people are able to interpret and understand sketches drawn by others (Blaser 2001). It is through this graphic representation that we acquire ideas of how humans store, understand, and communicate information they see (Bertin 1983). This is evident in sketch maps which have been used in many studies of how people represent their environment (Metz 1990; Wise and Kon 1990; Taylor and Tversky 1992). The aspect of correctness has been extensively studied; particularly the distortions in sketch maps (Tversky 1981). Some scholars have looked at possible approaches to address cognitive errors of representations in sketch maps using qualitative methods (Wang and Schwering 2009; Chipofya et al. 2011). Although distortions are inevitable in sketch maps, there are other aspects that make it reliable and effective in communicating spatial information. Sketch maps are static and will not respond to changes in the user's context unlike dynamic maps. They also do not adhere to any standard cartographic conventions and offer a good insight into how people perceive and illustrate vague referents. The level of personalization and flexibility in sketch maps allows the person drawing to take liberties with representation of vague spatial features. Strategies and distortions involved in representing these places in sketch maps is an understudied topic in spatial cognition.

The paper aims to investigate how places with unclear spatial extent may be represented. Participants were asked to sketch a given route and include pre-identified places which have unclear spatial extent. The task involved two groups to be investigated. Group 1 is a route within the city (intra-city route) and Group 2 is a route from the study area to another city in Germany (inter-city route). The study focused on how participants represent the same pre-defined places in these two different groups as well as what type of sketch representation is used to represent individual place categories—*district*, *neighborhood*, and *site* (described in Sect. 2.2). We first classified these representations from human-generated route sketch maps and then identified the place category to which they belong. Results of this study are applicable to other research areas involving pattern recognition in sketches, generating mobile maps and computer aided drawing.

The remainder of this paper is structured as follows. Section 2 discusses the different types of sketch representations used for the analyses followed by the categorization of places. In Sect. 3, the procedure and materials used in the experiment are explained. The outcome of the experiment is presented in the Results section (Sect. 4) and followed by the Discussion section (Sect. 5). Finally, in Sect. 6, conclusions and outlook for future work are presented.

2 Sketch Representations and Category of Places

2.1 Types of Sketch Representations

Blaser (2001) analyzed sketched objects based on their type, how they are visually portrayed and their purpose. This involves properties of sketched objects such as their shape, outline, fill patterns, completeness, number of strokes, dimensionality and annotations. Our classification uses a subset of these, while focusing more on the semantic properties of vague places. This was classified based on the dimension of how abstract a type is and what visual style is used to depict it (Fig. 1). Less abstract types depict the top-down view or the facade of the object as realistically as possible whereas highly abstract shapes are simply intended as a marker to anchor where the place is situated. It is also observed that by using some visual styles, the sketcher tries to convey that the place in question has unclear extents. Based on these distinctions, the sketch representations are categorized into the following types and used later in our analysis:

(1) *Simple label*—uses text to identify a place. A distinction is made over whether the text simply serves as an annotation to other types or is the sole indicator of a place.

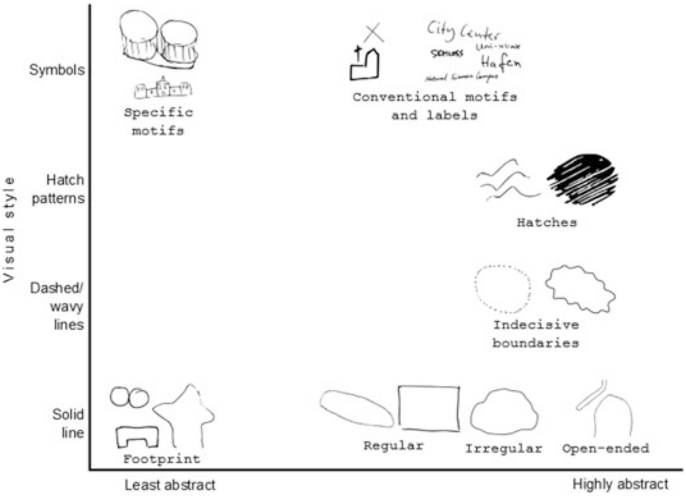


Fig. 1 Types of representations

(2) *Motifs*—the use of graphical symbols to denote the category of a place. Some motifs are specific and depict the actual appearance of a place they represent, whereas others tend to be generic and depict its type.

(3) *Footprints*—a unique pattern identifying the shape of a place such as the layout of a building or any salient feature.

(4) *Regular shapes*—clearly discernable pattern usually in regular and non-arbitrary forms such as circle, rectangle or ellipse. A shape which is outlined by street networks is also regular since the pattern is discernible.

(5) *Irregular shapes*—a representation which has no discernable regular pattern and is not a footprint.

(6) *Open-ended shapes*—a shape that is not bounded in any form and is purposely left open to indicate continuity.

(7) *Indecisive boundaries*—places are represented by wavy, dashed lines or dotted lines which serve to indicate that the drawn extent is approximate.

(8) *Hatch pattern*—series of strokes that give the impression of shaded region.

Figure 1 shows the different representations obtained from actual sketch maps. Highly abstract representations simplify the real world bearing no similarity to the shape or spatial extent of the real object. The less abstract ones attempt to imitate the real shape of the referent in a more recognizable way. Footprints are classified as least abstract since they reflect the shape of the object. Some symbols such as specific motifs are also treated as less abstract, showing a 2.5D representation of a building. Irregular, regular and open-ended shapes are treated as highly abstract.

2.2 Category of Place

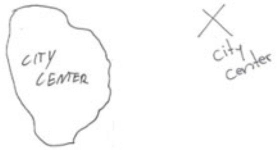
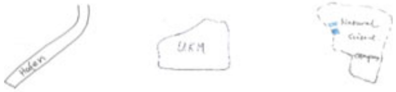
The places chosen in this study differ in how their geometry is represented in sketch maps—either as point or as region feature. To classify these representations in a more generalized way, we refer to Bennett's (2011) definition of the following place-related terms and treat each as a different category of place:

"*District*" refers to geographic regions which do not necessarily pertain to an actual unit of jurisdiction, but to a region of similar size with some (often vague) geographically related integrating principle. An example used in this study is the city center.

"*Neighborhood*" refers to part of the town with a common class of inhabitants or similar standard or buildings but it is also associated with sharing amenities such as shopping outlets and entertainment venues. Examples of places in this study that refer to this category are the Harbor, the Natural Science Campus and part of the University Hospital.

"*Site*" refers to a place where something is situated. This is typically applied to buildings and other large static artifacts. For this study, the examples of this category refer to a smaller region with a specific building wherein its surroundings are also recognized as its part, e.g. the castle and the University Hospital.

Figure 2 shows the different categories of place and some examples of how participants represented these in their sketch maps. One can observe that the type of representation used is not homogeneous, and varies even for a given category.

Category	Examples of Representation*
District	
Neighborhood	
Site	

* from actual drawings by participants

Fig. 2 Category of place

3 Methods

3.1 Participants

A total of 30 participants (15 females, 15 males) took part in the experiment and received 10€ for participation. They are between 20 and 36 years with median age of 26 (M = 26.8, SD = 4.5). Participants have been residents of the study area for minimum 6 months. Sixty-three percent (63%) had lived in the study area between 1 and 5 years and the rest (37%) lived for less than a year.

3.2 Study Area

The study area is Muenster, a mid-sized city in the northwestern part of Germany. Some vague places were identified within the city to be represented in the sketch map. These places do not have clearly defined spatial extents. Sometimes people perceive a feature to include its surroundings as well.

3.3 Design

Participants were asked to perform a sketch mapping task which required them to draw the pre-defined vague places. They were given a paper sheet of desired size (A4 or A3) and a pen and they could request additional sheets of paper if needed. Participants were equally distributed between two groups. Each person was asked to draw only one sketch map, therefore each Group produced a total of 15 sketches. In Group 1, the participants drew a route within the city. The instruction stated: *Please draw a map of the city. You may include as many landmarks and street names you can remember but please indicate the following places*: Harbor (Hafen); City center (Innenstadt); Castle (Schloss); University Hospital (Universitaetsklinikum); Natural Science Campus (Naturwissenschaftliches Zentrum or NSC). In Group 2, the participants were asked to sketch a route from the Natural Science campus to a specific place within the city center of another city in Germany they are familiar with. Except for the Harbor, all places mentioned in Group 1 were also required to be sketched for the intercity route. The Harbor, due to its location, was intentionally excluded because it may lead to confusion and difficulty in comprehending the instruction. The instruction stated: "*Please draw a route from the Natural Science Campus passing by central train station to the city center of any German town/city that you are familiar with. Please indicate where the city centers are for both cities. Please also indicate the following places in the city*:" Castle (Schloss); University Hospital (Universitaetsklinikum); Natural Science Campus (Naturwissenschaftliches Zentrum or NSC).

4 Results

4.1 Sketch Representation of Places with Unclear Extent

Participants' route sketch maps revealed differences in the representation of places. Table 1 shows the number of participants who used a specific representation type for each place in the sketch maps. The city center is represented in six types of representations—as simple label, motifs, regular shapes, irregular shape, indecisive boundary and hatches. There are no examples of sketch maps representing the city center as footprints and open-ended shape for both groups. Simple labels appeared to be the common type of representation for city center in both groups showing approximately 40% of all representations.

The University Hospital is represented as regular shapes, simple labels and footprints for both groups. For Group (Grp) 1, more than half of the representations are regular shapes and few participants used motifs and open-ended shapes to draw it. This also showed the same result for Group 2.

The Castle, on the other hand, is represented as regular shapes, footprint, motifs and simple labels. Regular shape made up 50% of the representations for castle for both Group 1 and Group 2. For this study, regular shape was the most common representation followed by footprint and motifs for Group 2 and 1, respectively.

The Natural Science Campus was mostly represented as regular shapes for both groups with 50% of the participants. However, it was not represented as hatch for both groups. In Group 1, it was not represented as motifs, footprint, and indecisive boundary. In Group 2, it was not represented as indecisive boundary.

The Harbor is represented mostly as labels (53%) but, it was also represented as simple labels, footprint, irregular shape and open-ended shape. This is similar to the results for the city center.

In general, the table shows that there are different ways of how vague places are represented. Disregarding the place, more than 40% of total representations accounted to regular shape in both groups. Simple labeling amounts to 31 and 22% of the total representations in Group 1 and Group 2, respectively. The third most common type of representation used is footprint (7% for Group 1 and 15% for Group 2).

Table 1 Participants' representations of all vague places in sketch maps

Representations	Castle		Uni hospital		NSC		Harbor		City-center	
	Grp 1	Grp 2	Grp 1	Grp 2	Grp 1	Grp 2	Grp 1	Grp 2	Grp 1	Grp 2
Simple label	2	2	3	3	3	2	8	–	7	6
Motifs	3	2	1	1	0	1	0	–	0	2
Footprint	2	4	2	4	0	1	1	–	0	0
Regular shape	8	7	9	6	8	8	4	–	6	4
Irregular shape	0	0	0	0	0	1	1	–	0	1
Open-ended shape	0	0	0	1	2	1	1	–	0	0
Indecisive boundaries	0	0	0	0	2	0	0	–	2	1
Hatch	0	0	0	0	0	0	0	–	0	1

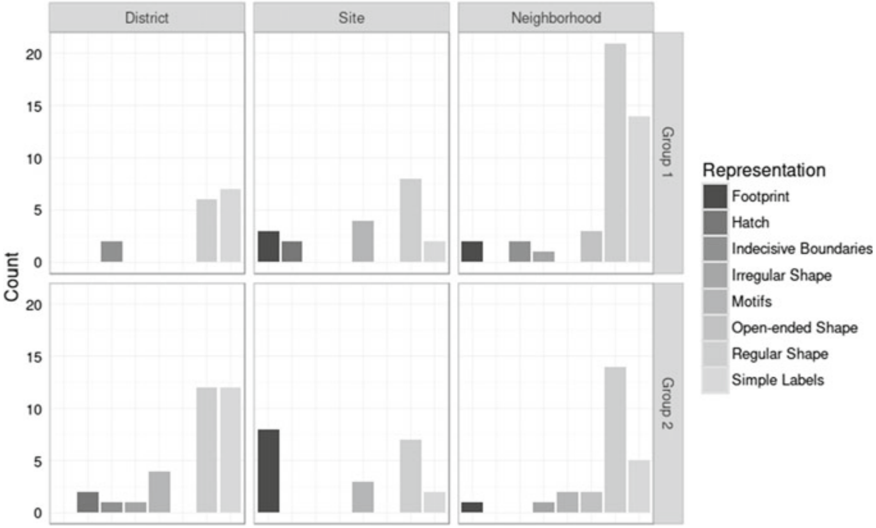
4.2 Differences in Sketch Representations of Category of Place

With regard to the differences of representations based on the categories of places, Fig. 3 shows the frequency count of how each category was represented in the sketch maps. District, which is composed of the city center, is mostly represented as labels for both groups. This is followed by regular shapes. A neighborhood, on the other hand, is frequently represented as a regular shape accounting for almost half of the representations in Group 1 (49%) and more than half of the total representations (56%) in Group 2. Regular shapes were also used more frequently (42%) to represent a site in Group 1, while in Group 2 it was footprint (40%). Looking at the other types of representations, hatches are mostly used to represent a district and site. Indecisive boundary and irregular shapes are used to represent both district and neighborhood. Open-ended shapes, on the other hand, are only used to represent a neighborhood. Irregular shapes were used to represent both district and neighborhood. Motifs are drawn to represent all three categories.

4.2.1 Representing a District

City Center

City center is considered a vague region. Inhabitants of the study area often have different perceptions of its actual extent. For example, the administrative boundary



* Representation types for each category are differentiated by shades (dark to light grey)

Fig. 3 Frequency of representation per category of place in sketch maps*

of the city center includes the train station at the lower right. But for many residents, the city center is defined by the area within the Promenade encircling the historical town (see Fig. 4).

As shown in the different sketch maps, some participants have represented it as a shaded region and dashed lines while others used a simple label. The most common representation is a bounded shape with label as annotation (Anacta et al. 2013, 2015). District is mostly represented as labels and regular shapes in both groups. It also shows that districts are not usually represented as footprint and open-ended shape. Looking at how city center is represented, participants also used dashed lines or solid lines to delineate boundary of the region.

4.2.2 Representing Neighborhood

Natural Science Campus

The boundary of the Natural Science Campus is not known to many students. From the university map, the Natural Science Campus is composed of the Physics, Chemistry, Biology, Pharmacy, and Geosciences department buildings (all red buildings in Fig. 5 including research laboratories and other facilities). Although it refers to these buildings, there are participants who include other surrounding buildings such as the Computer Science and Mathematics department buildings as part of the campus. Others refer to nearby buildings such as the University of Applied Sciences (Fachhochschule) as part of the Natural Science Campus.

Harbor

By definition, a harbor is a body of water for anchoring ships, boats and barges. However, the contemporary harbor in the study area is more of a vernacular rather than functional placename. Rather than an actual harbor, for residents and tourists alike, it is a place beside the canal with commercial buildings, restaurants and recreational facilities. Looking at the sketch maps in Fig. 6, the Harbor was

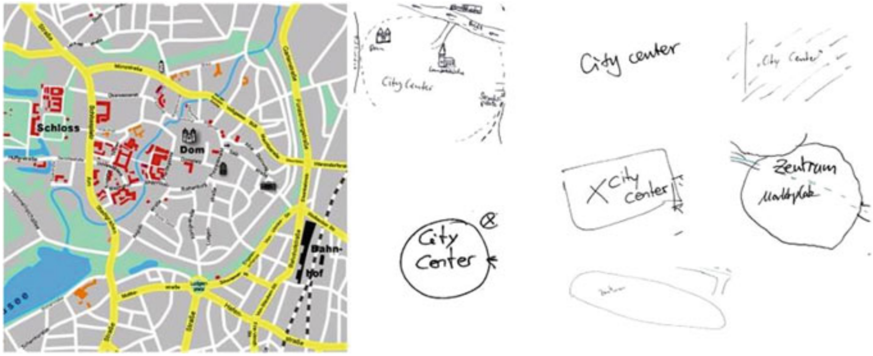


Fig. 4 Metric map (*left*) and participants' representation (*right*) of city center in sketch maps

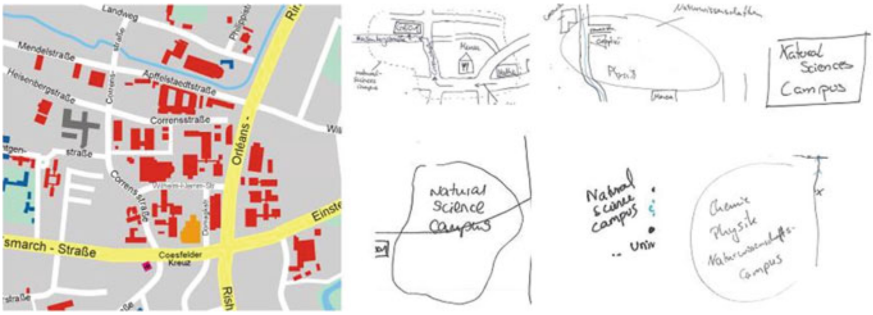


Fig. 5 Metric map (*left*) and participant's representations (*right*) of natural science campus in sketch maps

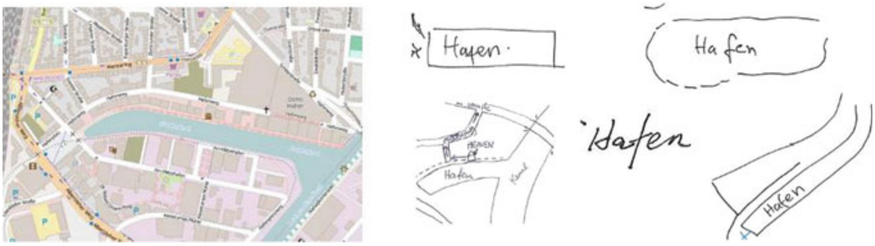


Fig. 6 Metric map (*left*) and participants' representation (*right*) of harbor in sketch maps

represented in different ways with labels being frequently used among all the representations. There were a few participants who represented it as a combination of footprint and open-ended shape.

University Hospital

The University Hospital is also a non-contiguous region wherein its boundary is hardly defined because of other dispersed buildings situated in another area. Figure 7 shows the street map with the location of all the buildings (in blue) that are collectively known as part of the University Hospital. In the northeast part of the map, there are some distant buildings that still belong to the university Hospital. Some participants drew the two multi-storeyed towers to represent the University Hospital but majority represented it as an area including some of its surroundings. Not a single participant included any of the distant buildings as part of the University Hospital. Thus, in analyzing the sketch maps, the University Hospital is considered either as a *neighborhood* or as a *site* depending on what aspect was represented.

In the categorical analysis of the vague places, regular shape is the dominant representation for both groups. The next frequent type of representation is simple label. No representation for hatch was drawn to represent a neighborhood.

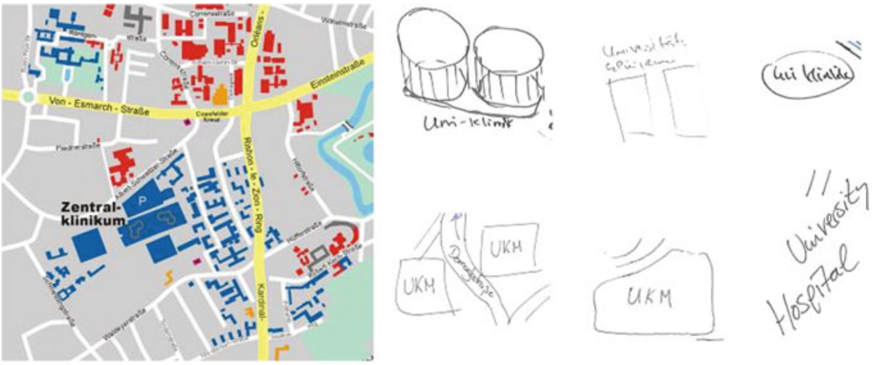


Fig. 7 Metric map (*left*) and participants' representation (*right*) of university hospital in sketch maps

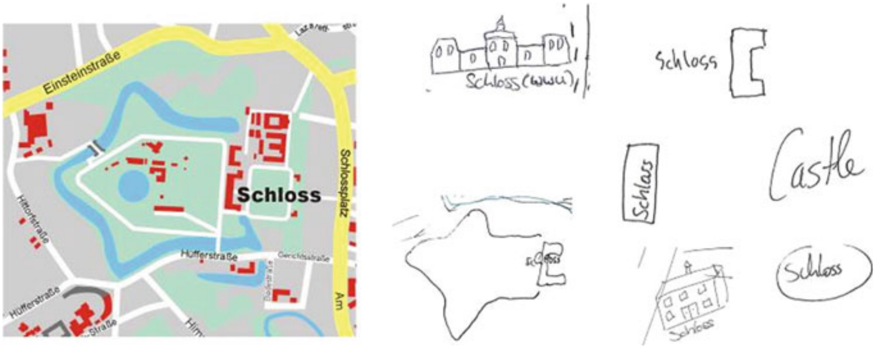


Fig. 8 Metric map (*left*) and participants' representation (*right*) of castle (Schloss) in the sketch maps

4.2.3 Representing Site

Castle

The Castle is a well-defined landmark but sometimes represented as a regional feature. For example, the castle which is shown in Fig. 8 has a distinct building footprint but some participants also drew its surroundings as part of it. In the current experiment, most participants represented the castle frequently with regular shape for both groups. The second most common representation was footprint.

University Hospital

As mentioned earlier, the University Hospital was considered either as an example of a site or a neighborhood. For site, the University Hospital was represented as a footprint. Some participants tend to draw the two towers (see Fig. 6) referring to the

University Hospital. With neighborhood, it was represented mostly as a regular shape. This is followed by a footprint wherein most participants from Group 2 drew this type of representation. Representing the site as a symbol was more common for participants in Group 1. Compared to the other categories, simple label was not represented frequently. No participant represented a site as an irregular shape, open-ended shape, indecisive boundaries or hatch pattern.

5 Discussion

5.1 Types of Sketch Representations

Vague places often appear in everyday communication for instance when describing an environment or giving wayfinding instructions. This occurs in some route instructions wherein participants refer to vague places such as city center as an example of regional landmark for orientation (Schwering et al. 2013). One result of this study is that labels are used frequently even when their use as simple annotations is disregarded. This was not in line with the findings of Blaser (2001) who found a low percentage of use of labels as stand-alone objects. This might be because our study focused mainly on the representation of vague places in sketch maps.

Sketchiness of lines might be associated with vagueness of a place. This is what Boukhelifa et al. (2012) investigated when they look at different visual variables such as blur, dashed and solid lines to qualitatively analyze vague information. This type of representation occurred in our study wherein some participants drew dashed lines and wavy lines to represent regions such as city center and Natural Science Campus. This could be interpreted that participants were uncertain about its spatial extent when drawing such types of representation because these lines were used only for drawing vague places and not for well-defined places on the maps.

In general, the dominant form of representation in the sketch maps is regular shapes. This matches Blaser's (2001) findings wherein 78% of objects have non-complex shapes which is referred in this study as regular shapes. In our case, this appeared more frequently when representing the University Hospital, Castle, and the Natural Science Campus. The city center and Harbor, on the other hand, were represented often as simple labels. One reason might be that participants are not certain of their spatial extent unlike the other places where they refer to specific buildings.

5.2 Sketch Representation of the Category of Places

Participants used different ways of representing category of places. It appeared that the common types of representation for all the categories of places are simple labels and regular shapes. But, it was also shown that there are some representations that may only apply to a specific category of place. For instance, hatch was never used to represent a neighborhood, but was rather drawn to represent mostly districts and sites.

In representing *districts*, simple labels appeared to be the most common sketch representation. This shows that regardless of the route drawn (intra-city or inter-city), participants use labels to represent the city center. However, more often it is combined with a regular shape. District could also be represented with indecisive boundaries and irregular shape. This is similar to the study of Orleans (1973) wherein urban residents represent it mostly with both regular and irregular shape combined with labels. This presents a clear understanding that district, being a region, has to be drawn in an enclosed figure (which might be the reason why in this study it was never represented as an open-ended shape). Furthermore, district was not represented as a footprint.

When representing *neighborhood*, indecisive boundary and irregular shape were frequently used like the case of district. But, there could also be more than one representation for a specific place. This happens with the University Hospital which is represented either as a site or a neighborhood. Participants sometimes represented the hospital as either point or as a region. This is because some buildings of the University Hospital are dispersed and some participants refer to distinct buildings to represent the whole region (see Fig. 7). This confirms our expectation wherein different categories of places are represented using different types of representations.

In representing *site*, participants frequently use regular shapes. It was not represented as irregular shape, open-ended shape and indecisive boundaries. Perhaps this is because 'site' shows a less abstract representation. This also explains why such category is also represented as motif and footprint. But similar to the University Hospital, the castle, under the site category, is also represented either as point or regional feature (see Fig. 7). Participants not only refer to the building but tend to include its surrounding features referring to it as the entire castle.

The types of representations used to sketch *sites*, *neighborhoods* and *districts* were similarly distributed in both groups. However, Group 2 seems to have used 'simple labels' much less when naming neighborhoods. While labels can precisely identify individual areas, they do so through the semantic, and not visual, uniqueness. They require knowing what the place is, not how it looks like. It seems that the larger spatial extent of the task this group faced (drawing a route to another city), decreases the need for, or relevance of, this particular type of information.

The results provide ideas as to how people's representations of places may be interpreted into whether it refers to district, neighborhood or site. Bertin (1983) extensively investigated different graphic representations to better understand how to visualize data both quantitatively and qualitatively making sense of cartographic principles. This study, on the other hand, provided additional interpretations to some of these visualizations based on how humans represent different categories of places with unclear spatial extent on sketch maps. The results showed the relevance of such places to be represented in a map or any navigation system because people often use it in daily communication (Montello et al. 2003). This might enhance readability of maps as it will show only features of interest to avoid visual clutter.

For example, these representations may be applied in creating schematic maps since spatial relations of places are more important than its actual extent where in many cases, the extent is indeterminable. In designing maps, these types of representations can be used to visualize vague places. Furthermore, when drawing sketch maps for directions, this will help establish a common understanding on how people may interpret it based on the different category of place such as neighborhood, a district or a site. Since people are able to interpret sketches which are abstract, it may help them to understand and interpret similar type of places.

6 Conclusion and Future Work

Interpreting representation of vague places is a challenge since people have different ways of drawing them, i.e. in sketch maps where there are no consistent guidelines. One reason that influenced their sketch representations could be that they have acquired it through reading maps or by experience. Our observations from the experiment suggest that:

- *Vague places with regional extent are mostly represented using labels.* For example, with districts such as the city center, the dominant representation is stand-alone labels. The Harbor, which has an areal extent, is also represented frequently as simple labels.
- *Point features are sometimes represented as regions and vice versa.* Well-known buildings situated inside a region are oftentimes referred to in sketch maps. But sometimes surrounding features are also included in the representation even if the place refers only to a building.
- *Representation may depend on the category of vague place.* A place may be classified as either a site or neighborhood. In this study, it was the University Hospital which was considered either site or neighborhood.

The study provided empirical evidence of how places with unclear extent are represented on sketch maps. The category of place may serve as a basis in classifying related vague places in a more general way which may help build a common understanding in sketched route instructions between the receiver and the giver. The findings may be useful for researchers developing applications for location-based services to visualize places with vague extents on mobile devices. Such places are oftentimes used as reference point in giving wayfinding instructions and it will be interesting to find out how people represent them. This study could also benefit researchers dealing with pattern recognition to understand the semantics of what a sketched pattern represents. For future work, we plan to generate series of sketch maps with the different representations based on the results of this study to find out how people would interpret such an environment. Furthermore, it will also be interesting to generate visualizations that mimic sketches drawn by humans from a set of route instructions in natural language as an alternative to street maps and assess their usability.

Acknowledgements This study was supported by the German Research Foundation (DFG) with project number SCHW1371/15-1 and the ERC StRG Agreement No. 637645.

References

Anacta V, Humayun M, Schwering A (2013) Visualizing vagueness in sketch maps. In: Workshop paper at visually-supported reasoning with uncertainty. Conference on spatial information theory (COSIT)

Anacta VJA, Humayun M, Schwering A (2015) Map-off the city: how uncertain places are represented in sketch maps. Poster presented at 18th AGILE international conference on geographic information

Bennett B (2010) Methods for handling imperfect spatial information. Springer, Heidelberg

Bennett B, Agarwal P (2011) Exploring the place of vagueness in spatial information, COSIT Tutorial

Bertin J (1983) Semiology of graphics: diagrams, networks, maps (Berg W, Trans), The Univesity of Wisconsin Press and Esri Press

Blaser AD (2001) A study of people's sketching habits in GIS. Spat Cogn Comput 2(4):393–419

Boukhelifa N, Bezerianos A, Isenberg T, Fekete J-D (2012) Evaluating sketchy lines for the visualization of qualitative uncertainty. IEEE Trans Visual Comput Graphics 18(12): 2769–2778

Chipofya M, Wang J, Schwering A (2011) Towards cognitively plausible spatial representations for sketch map alignment, spatial information theory. In: Proceedings of 10th international conference on COSIT 2011, Belfast, ME, USA, 12–16 September 2011. Springer, Heidelberg

Fisher P, Comber A, Wadsworth R (2006) Approaches to uncertainty in spatial data. In: Devillers R, Jeansoulin R (eds) Fundamentals of spatial data quality, ISTE, pp 43–59

Humayun MI, Schwering A (2013) Selecting a representation for spatial vagueness: a decision making approach, geographic information science at the heart of Europe. Lecture notes in geoinformation and cartography. Springer International Publishing

Metz HM (1990) Sketch maps: helping students get the big picture. J Geogr 89(3):114–118

Montello DR, Goodchild MF, Gottsegen J, Fohl, P (2003) Where's downtown?: behavioral methods for determining referants of vague spatial queries. Spat Cogn Comput 3(2–3):185–204

Orleans P (1973) Differential cognition of urban residents: Effects of social scale on mapping. Aldine Publications, Chicago

Schwering A, Li R, Anacta VJA (2013) Orientation information in different forms of route instructions. Proceedings on the 16th AGILE international conference on geographic information science

Taylor HA, Tversky B (1992) Descriptions and depictions of environments. Mem Cogn 20(5): 483–496

Tversky B (1981) Distortions in memory for maps. Cogn Psychol 13(3):407–433

Voudouris V, Fisher PF, Wood J (2006) Progress in spatial data handling. Springer, Berlin Heidelberg

Wang J, Schwering A (2009) The accuracy of sketched spatial relations: how cognitive errors influence sketch representation, presenting spatial information: granularity, relevance, and integration. Workshop at COSIT 2009, Aber Wrac'h, France

Wise N, Kon JH (1990) Assessing geographic knowledge with sketch maps. J Geogr 89(3): 123–129

Mining Rainfall Spatio-Temporal Patterns in Twitter: A Temporal Approach

Sidgley Camargo de Andrade, Camilo Restrepo-Estrada, Alexandre C. B. Delbem, Eduardo Mario Mendiondo and João Porto de Albuquerque

Abstract Social networks are a valuable source of information to support the detection and monitoring of targeted events, such as rainfall episodes. Since the emergence of Web 2.0, several studies have explored the relationship between social network messages and authoritative data in the context of disaster management. However, these studies fail to address the problem of the temporal validity of social network data. This problem is important for establishing the correlation between social network activity and the different phases of rainfall events in real-time, which thus can be useful for detecting and monitoring extreme rainfall events. In light of this, this paper adopts a temporal approach for analyzing the cross-correlation between rainfall gauge data and rainfall-related Twitter messages by means of temporal units and their lag-time. This approach was evaluated by conducting a case study in the city of São Paulo, Brazil, using a dataset of rainfall data provided by the Brazilian National Disaster Monitoring and Early Warning Center. The results provided evidence that the rainfall gauge time-series and the rainfall-related tweets are

S.C. de Andrade (✉)
Federal University of Technology – Paraná, Curitiba, Toledo, Brazil
e-mail: sidgleyandrade@utfpr.edu.br; sidgleyandrade@usp.br

S.C. de Andrade
University of São Paulo, São Carlos, Brazil

C. Restrepo-Estrada
São Carlos School of Engineering, University of São Paulo, São Carlos, Brazil
e-mail: camilo.restrepo@udea.edu.co

A.C.B. Delbem
Institute of Mathematical and Computing Sciences, University of São Paulo, São Carlos, Brazil
e-mail: acbd@icmc.usp.br

E.M. Mendiondo
Brazilian National Center of Monitoring and Early Warning of Natural Disasters, São José dos Campos, Brazil
e-mail: emm@cemaden.gov.br

J.P. de Albuquerque
Centre for Interdisciplinary Methodologies, University of Warwick, Coventry, UK
e-mail: J.Porto@warwick.ac.uk

A. Bregt et al. (eds.), *Societal Geo-innovation*, Lecture Notes
in Geoinformation and Cartography, DOI 10.1007/978-3-319-56759-4_2

not synchronized, but they are linked to a lag-time that ranges from −10 to +10 min. Furthermore, our temporal approach is thus able to pave the way for detecting patterns of rainfall in real-time based on social network messages.

Keywords Social network · Twitter · Rainfall · Temporal analysis · Time-series correlation

1 Introduction

Social networks are playing an increasingly important role in the fields of information fusion and data mining because of the assistance they give in detecting and monitoring targeted events that attract the attention of users, such as rainfall episodes. In the last few years, there has been a growing interest in analyzing social network messages for disaster management (Steiger et al. 2015). In this context, the focus of recent research has been on analyses of social network messages to detect events such as earthquakes (Crooks et al. 2013; Earle et al. 2012; Sakaki et al. 2010), hurricanes (Kryvasheyeu et al. 2016), and forest fires (Spinsanti and Ostermann 2013). Another group of studies has adopted an approach for extracting and classifying relevant information about the target-event, e.g. situational updates about an ongoing event (Albuquerque et al. 2015; Herfort et al. 2014; Starbird et al. 2012; Imran et al. 2013).

However, the previous studies fail to address the problem of the temporal validity of the social network data in real-time. This validity is important for establishing the correlation between social network activity and the different phases of targeted events. For instance, in real situations a decision-maker might want to understand the relation between the social network messages and the phase of the targeted event. In view of this, there is a need to fill a gap in these studies since they only examine social network messages after the event and show how they spread from the moment the targeted event starts. In other words, previous studies concentrated on conducting a post-hoc analysis based on the whole dataset instead of treating the social network stream as a time-series.

Furthermore, another limitation is that some of studies used a large time-scale for exploring the spatio-temporal correlation. Others assigned a temporal resolution that is based on social network messages alone, rather than supporting them with authoritative data to improve credibility. For example, Albuquerque et al. (2015) explored the correlation between the significance and proximity of the flood-related tweets with water levels using a one-day time-scale. Kryvasheyeu et al. (2016) investigated the Twitter messages aggregate hourly to predict the location and severity of hurricane damage. Earle et al. (2012) and Sakaki et al. (2010) did not include the temporal resolution of the authoritative data for detecting earthquakes. For example, Sakaki and Matsuo (2012) “detect an earthquake when five positive tweets arrived in 5 min”. Assis et al. (2015) prioritized flood-related tweets only if a sensor reported a high water level within of catchment. However, rainfall detection in real-time using rainfall gauge data with social network messages in urban areas requires a flexible and a finer time-scale than other events such as earthquakes, hurricanes, fires and floods. It is because extreme rainfall episodes, such as convective rainfall, are events that occur in few hours or minutes, with a higher degree of intensity for rainfall durations of less than 1 h (Llasat 2001; Marchi et al. 2010).

The problem of temporal validity can be understood in terms of the existing timing differences between two or more time-series. It is closely linked to event detection in real-time or near real-time with the aid of social network messages. For example, we know that social network messages can occur at the beginning, middle or end of the targeted events, but *how can the appropriate time for using the social network messages in a real-time setting be detected?* This question can be answered through a temporal validity of data using a lag-time of the time-series. It allowed us to take note of the existing correlation between the data in different phases of the rainfall events, and thus could be useful for detecting, monitoring or disaster management in real-time. Hence, the time-series analysis must include the lag-time variable to ensure it is more effective. Thus, it can serve to support the extraction of key information and improve the use of social network messages in practical solutions, such as monitoring and early warning systems.

With regard to the lag-time and its potential use, the following research question can be raised:

- Is there a temporal relationship between the rainfall gauge time-series and the rainfall-related social network stream?

Our answer to this question is based on the assumption that the temporal relationship between authoritative data and event-related social network messages can be explained by duration, frequency and the lag-time of the time-series. For this reason, the cross-correlation between rainfall gauge data and rainfall-related tweets is analyzed by means of the temporal units and their lags, rather than only the duration and frequency of the time-series. In addition, an attempt is made to describe the lag-time as a means of showing how it can be applied (together with the existing approaches) to detect events and support the retrieval of appropriate information in rainfall events.

The aim of this paper is thus to adopt a temporal approach to assess the correlation between rainfall data and social network messages at different lag-times. This assessment is based on the spatio-temporal features of a single time-series obtained from the social network messages and rainfall measurements. In light of this, this work seeks to provide a novel approach by means of lag-times rather than only the duration and frequency of the time-series.

This paper is structured as follows. Section 2 introduces the problem statement and hypothesis of this research. Section 3 gives a detailed account of the case study used for evaluating the approach, while Sect. 4 examines the methodology employed. Following this, Sect. 5 shows the results. Finally, Sect. 6 discusses the results obtained, draws some conclusions and makes recommendations for future work.

2 Problem Statement and Hypothesis

Figures 1 and 2 depict, to a limited extent, the increase in the rainfall and frequency of rainfall-related tweets in two periods of January 2016, in São Paulo, Brazil, during the peak of the rainy season. As can be seen, there is a similarity (based on the peaks) between the rainfall time-series and rainfall-related tweets time-series. However, there is not an exact correspondence between both the time-series. In ordinary circumstances, social network users often report messages related to rainfall episodes. That is, they post texts, photos and videos, for example of gray clouds, drizzle, rainfall, and floods, and on rare occasions, give a forecast of rain. These posts appear at different times, i.e. the temporal scale and frequency of rainfall-related social network messages are not known. They may occur before or after the rain, whilst other occur during a period of rainfall (Table 1). Moreover, the social network users can forward older information instead of new information (Earle et al. 2012).

On the other hand, there is a well-known temporal scale for rainfall data. The information usually, comes from specialist equipment (e.g. rainfall gauges) installed in urban areas. In Brazil, for instance, a widely-used temporal scale in urban areas is 10 min (Figs. 1 and 2). This means that it is a challenge to explore the correlation between the rainfall time-series and rainfall-related social network messages.

On the basis of this problem statement, the hypothesis raised here is that *there is a lag-time between rainfall and rainfall-related social network messages*. In formal terms, our hypothesis can be expressed as follows:

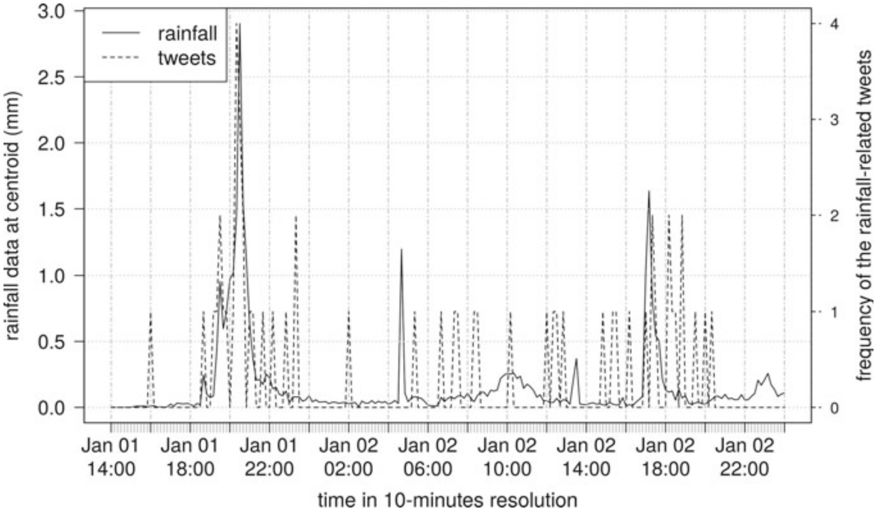


Fig. 1 Increasing rainfall and frequency of rainfall-related tweets from January 1st 14:00 BRST to 3rd 00:00 BRST, São Paulo, Brazil (with 10-min temporal resolution)

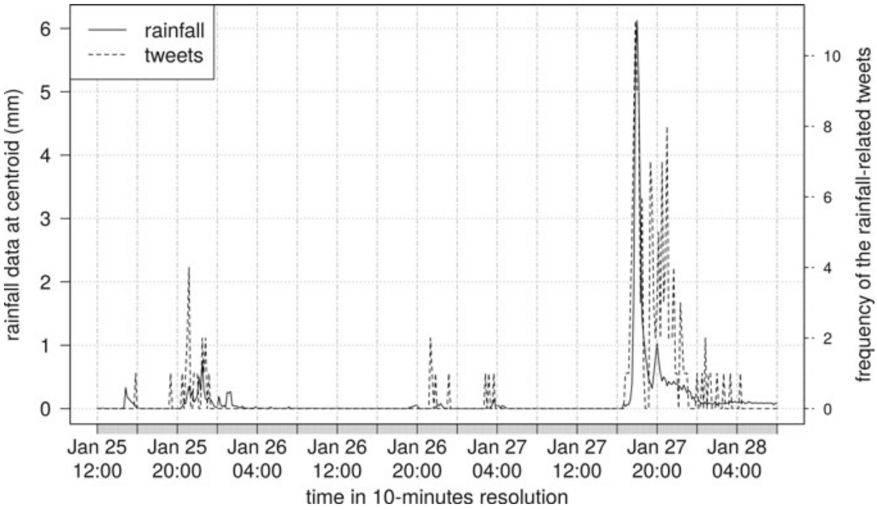


Fig. 2 Increasing rainfall and frequency of rainfall-related tweets from January 25th 12:00 BRST to 28th 08:00 BRST, São Paulo, Brazil (with 10-min temporal resolution)

Table 1 Examples of rainfall-related tweets classified per time (before, after or during the rainfall)

Date/Time	Rainfall-related tweets	Translation	Time
2016-01-12 06:11:39	"Escuro e quente, a tarde a chuva vem, não se enganem, é verão!!!! (…)"	"Dark and warm, in the afternoon there will be rain; make no mistake, i, it's summer !!!! (…)"	Before
2016-01-28 13:47:57	"A chuva de ontem @ Em MOEMA https://t.co/3kggiLckGZ"	"Yesterday's rain @ In MOEMA city https://t.co/3kggiLckGZ"	After
2016-01-27 17:37:39	"Chuva chuva e mais chuva… https://t.co/vzC9w01qQ8"	"Rain rain and more rain… https://t.co/vzC9w01qQ8"	During

Hypothesis. Let $Q_t = \{q_1, q_2, \ldots, q_i, \ldots, q_m\}$ and $P_t = \{p_1, p_2, \ldots, p_j, \ldots, p_n\}$ be defined by two time-series, the frequency of rainfall-related social network messages and rainfall data, respectively. The elements q_i and p_j are indexed in time T. For all $t \in T$ there is a constant $k \in Z$ that makes the relationship function ρ at time t other than zero.

$$\forall_{t \in T},\ \exists_{k \in Z}\ :\ \rho(P_t, Q_{t+k}) \neq 0 \tag{1}$$

where P_t and Q_{t+k} are observations (values) of the variables at time t and $t+k$, and k is the lagged k-periods. Here, the function ρ is a measure of correlation between both time series. We have checked ρ through the cross-correlation of time-series, as will be seen further on.

Furthermore, an attempt is made answer the two following questions: (Q1) *What happens when the k constant is other than zero?* (Q2) *How many lagged k-periods are needed to provide a suitable description of the social networks messages about rainfall gauge data?* These questions are necessary to show how our temporal approach can be applied, together with the existing approaches, to detect patterns and support the extraction of key information about rainfall events.

3 Case Study

3.1 Context of the Study—Rainfall in Brazil

The case study was carried out in São Paulo, Brazil, where heavy rainfall often affects the city's infrastructure and citizens. Figure 3 shows the scenario with the spatial distribution of active rainfall gauges and rainfall-related tweets in January 2016. It can be seen that the rainfall-related tweets are closer to the regions with rainfall gauges data than the regions without rainfall gauges data.

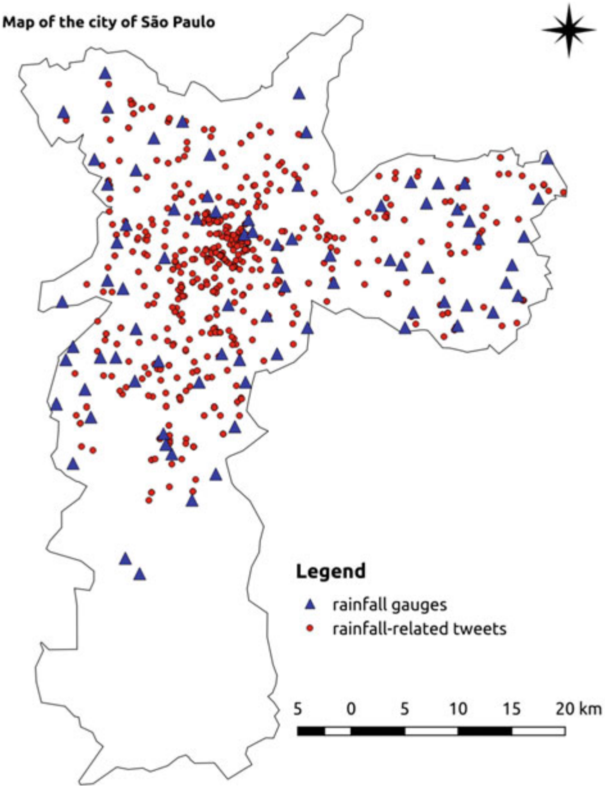


Fig. 3 Map of the distribution of active rainfall gauges and rainfall-related tweets in January 2016, in São Paulo, Brazil

3.2 Description of the Dataset and its Time-Series

Social network messages

Our social network dataset contains 243,333 georeferenced tweets retrieved by Twitter Streaming API from January 1st to January 30th 2016. All the tweets selected were geotagged within the administrative borders[1] of the city of São Paulo.

At first, the tweets collected via Twitter Streaming API were filtered with the aid of keywords and synonyms (including stem words and wildcard). The keywords in Brazilian-Portuguese were “chuv*” (chuva, chuvisco, chuvarada, etc.), “garoa*” (garoando, etc.), “temp*” (temporal, tempestade, tempo ruim, etc.), “alag*” (alagamento, alagado, etc.), and “inund*” (inundação, inundado, etc.). These keywords were chosen and extended from a list of previous studies (Assis et al. 2015). After this, we carried out a manual search for the subset of rainfall-related tweets to remove the false-positives (2,916 tweets were removed). In this context, the false-positive are tweets that contains one or more keywords, but they are not inside the context, i.e., rainfall-unrelated tweets. For example, there are several rainfall-unrelated tweets containing the keyword “garoa” because the city of São Paulo is known as the land of drizzle. Furthermore, we removed all retweets messages (228 retweets), i.e., messages concerning of another Twitter user and that contains the text “RT”. With regard to messages with Twitter Emoji (e.g. 🌧) and grammar mistakes (e.g. “Xuva” instead of “CHuva”), they were classified along with unrelated messages (239,569 tweets were classified as unrelated). Table 2 shows some examples of rainfall-related tweets. In our case study only 0.25% (620 tweets) were marked by us as rainfall-related tweets.

Table 2 Examples of rainfall-related tweets that have been classified manually

Date/Time	Rainfall-related tweets	Translation
2016-01-09 01:19:57	“CHUVAAAAAAAAA”	“heavy rain”
2016-01-09 15:20:40	“Que chuvinha mais gostosa”	“what wonderful rain”
2016-01-11 07:43:41	“ta garoando aqui”	“it’s drizzling here”
2016-01-27 15:58:40	“Preparando o barco, chove muito em sampa.”	“Get the boats ready, it’s pouring with rain in São Paulo”
2016-01-27 18:56:30	“Tempestade! @ Avenida Ipanema Sorocaba https://t.co/HkvrBlBgeW”	“A Storm! @ Ipanema Sorocaba Avenue https://t.co/HkvrBlBgeW”

[1] We took the geometry of the city from the Global Administrative Areas (GADM).

Rainfall data
The rainfall data were obtained from the National Center for Monitoring and Early Warning of Natural Disasters (CEMADEN)[2] through a REST API. There were 81 active rainfall gauges in São Paulo in the data collection period. The rainfall measurements were provided in linear depth (millimeters) and with two sizes of temporal window: 10 min when it was raining and 60 min otherwise.

We only used reliable rainfall gauges and removed those that had no data throughout the month and those which registered negative values. In total, we removed 21 rainfall gauges.

4 Methodology

As can be seen from Fig. 4, we designed a temporal approach for analyzing the cross-correlation between rainfall gauge data and rainfall-related rainfall-related Twitter messages. We evaluated this approach through the case study that was undertaken.

Our temporal approach comprised three steps, which are as follows:

Step 1 to gather sets of rainfall-related tweets and authoritative rainfall data;
Step 2 to generate a time-series of the rainfall data and rainfall-related tweets; and
Step 3 to analyse the cross-correlation between both the rainfall time-series and rainfall-related tweets time-series.

In Step 1, we selected a set of tweets that were published with the same spatio-temporal rainfall data obtained from the rainfall gauges and that contained content related to rain events. These filtered tweets are called rainfall-related tweets. After this, we generated two time-series, i.e., one for rainfall gauges data and another for rainfall-related tweets (Step 2). Both time-series were generated with time-scales of 10, 20 and 30 min. The rainfall-related tweets time-series corresponds the frequency of rainfall-related tweets during the period of analysis, whereas the rainfall time-series are rainfall gauges data interpolated at the centroid using the Inverse Distance Weighting method. Finally, we calculated the correlation between them (Step 3) for each time-scale (10, 20 and 30 min). Step 1 involves manual tasks, whereas Steps 2 and 3 are both automatic. The task types and process flow are drawn up in the Business Process Model and Notation (BPMN) to allow a visualization of the whole process used in our methodology.

[2] CEMADEN website is available at www.cemaden.gov.br.

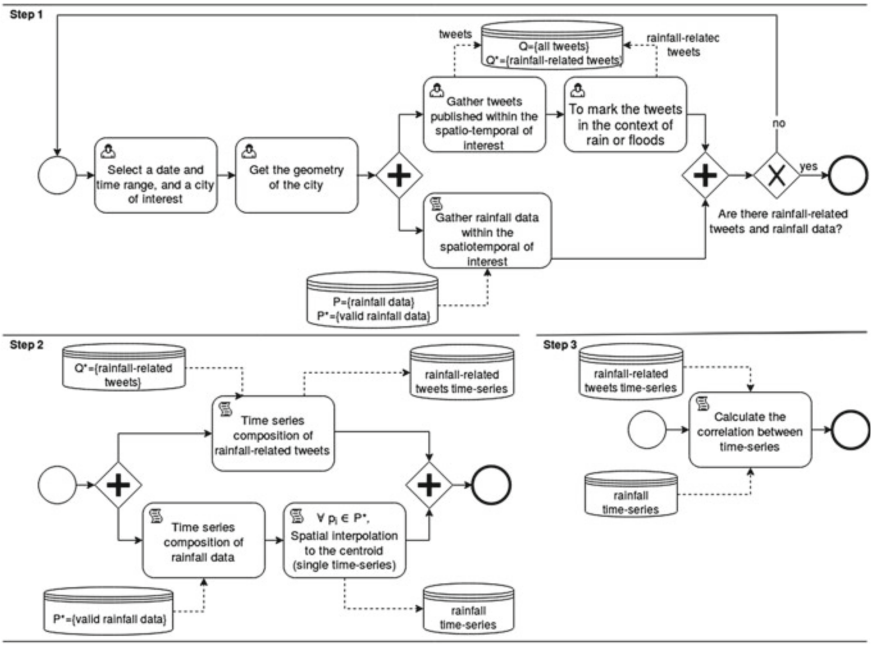

Step 1
tweets
Q={all tweets}
Q*={rainfall-related tweets}
rainfall-relatec tweets
Gather tweets published within the spatio-temporal of interest
To mark the tweets in the context of rain or floods
no
Select a date and time range, and a city of interest
Get the geometry of the city
yes
Are there rainfall-related tweets and rainfall data?
Gather rainfall data within the spatiotemporal of interest
P={rainfall data}
P*={valid rainfall data}
Step 2
Q*={rainfall-related tweets}
rainfall-related tweets time-series
Time series composition of rainfall-related tweets
Time series composition of rainfall data
∀ pi ∈ P*, Spatial interpolation to the centroid (single time-series)
P*={valid rainfall data}
rainfall time-series
Step 3
rainfall-related tweets time-series
Calculate the correlation between time-series
rainfall time-series

Fig. 4 Methodological approach for designing the temporal approach

4.1 Methods

The next sections describe the methods related to spatial interpolation of a rainfall time-series (Step 2), cross-correlation between rainfall and rainfall-related tweet time-series (Step 3) and how they are applied to our case study.

4.1.1 Inverse Distance Weighting

Inverse Distance Weighting (IDW) is a technique that is commonly used for estimating the values of rainfall measurements at the centroid of a catchment by calculating the weighted averages of the available rainfall gauges (Bartier and Keller 1996; Ahrens 2006; Yang et al. 2015; Mair and Fares 2011). The rainfall gauges nearest to the centroid will have a greater weight for calculating the average rainfall. The rainfall measurements at the centroid can be calculated as follows:

$$p_c^t = \frac{\sum_{i=1}^{N} (\frac{1}{d_i^r}\, p_c^t)}{\sum_{i=1}^{N} (\frac{1}{d_i^r})} \quad \textit{if } d_i \neq 0, \textit{ for all } i \tag{2}$$

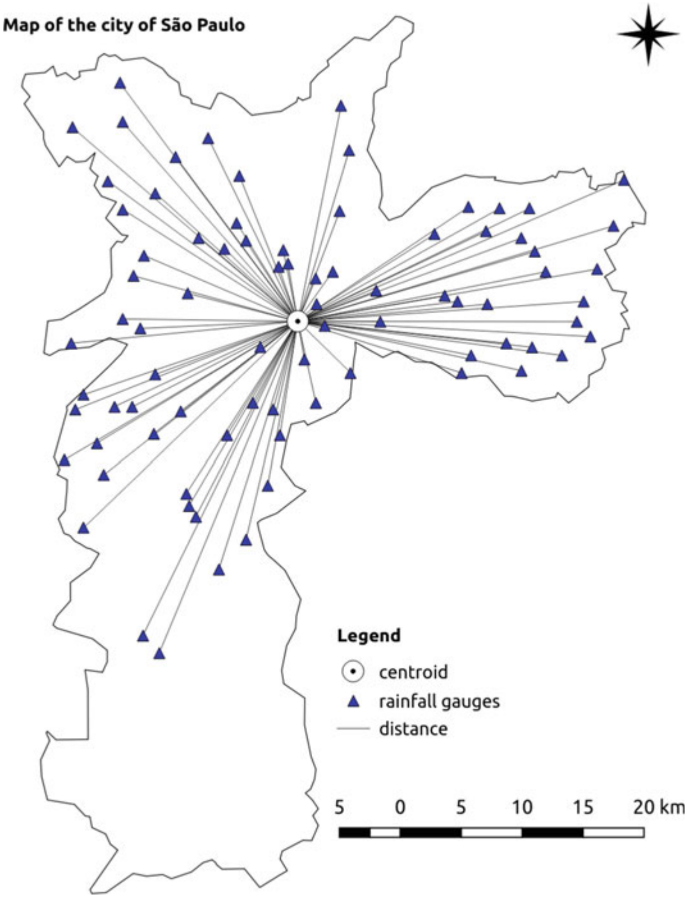


Fig. 5 Spatial interpolation calculated from rainfall for the time period 1st–30th January 2016

where p_c^t is a rainfall at the centroid, p_i^t is a rainfall in each rainfall gauge, d_i is the distance from the rainfall gauge i to the centroid and r is the power parameter. In our case, we considered, for the sake of simplification, that the urban area of the city of São Paulo forms a single catchment. Figure 5 depicts the spatial interpolation that is calculated from rainfall data using the parameter $r = 2$ for the entire period of the case study.

4.1.2 Cross-Correlation

Let q_t and p_t be two stationary processes. The correlation is any statistical relationship, whether causal or otherwise, between two random variables or two sets of data. The cross-correlation is already a function that estimates the correlation between q_t

and p_t at pairs of time points (Bacchi and Kottegoda 1995). The cross-correlation at lag k is given as follows:

$$\rho_{qp}(k) = \frac{\sum_{t=1}^{N} (q_{t+k} - \mu_q)(p_t - \mu_p)}{\sqrt{\sum_{t=1}^{N} (q_t - \mu_q)^2 \sum_{t=1}^{N} (p_t - \mu_p)^2}} \tag{3}$$

where μ_q and μ_p are the expected values of q and p, respectively, and k is a lag of the variable q. The cross-correlation is a method to check if the two series have randomness. This randomness is ascertained by computing autocorrelations for data values at varying lag-times. If they are random, these correlations should be near zero for any and all lag separations. Otherwise, one or more of the correlations will be significantly non-zero. With k negative, they are predictors of p_t, it is sometimes said that q leads to p. When one or more q_{t+k}, with k positive, are predictors of p_t, it is sometimes said that q lags p.

5 Results

5.1 10-min Temporal Resolution

As can be seen from Figs. 1 and 2 (see Sect. 1), the rainfall gauge time-series and the rainfall-related tweets are not synchronized, i.e., there is not an exact correspondence between both the time-series. This evidence appears throughout the case study. Another piece of evidence is that the peaks of both the time-series sometimes shift in time, i.e., the peak of the rainfall-related tweets time-series appears before the peak of the rainfall gauge time-series and vice versa. For example, from 20:00 BRST to 21:00 BRST on 1st January the frequency of rainfall-related tweets came before the rainfall (Fig. 1), but from 17:00 BRST to 18:00 BRST on 2nd January the frequency of rainfall-related tweets came after the rainfall (Fig. 2). A similar kind of behavior can be observed in other periods in January 2016.

Figure 6 depicts the cross-correlation for the period 20:00 BRST–21:00 BRST on 1st January 2016, whereas Table 3 summarizes the cross-correlation values for the same period. The blue dotted line (Fig. 6) represents the confidence interval of 95%. Although the significance correlation is in the range of −7 to 7 lag-times, the greatest correlation is −1 lag-time ($k = -1$). This means most rainfall-related tweets come 10 min before the rainfall phenomenon. The most likely explanation is that rainfall-related tweets can predict future rainfall or detect rainfall activity in real time or near real-time, i.e., they can occur before the rain (see Table 1). This means that this setting can be used for forecasting and monitoring rainfall and thus provide better support for decision-making.

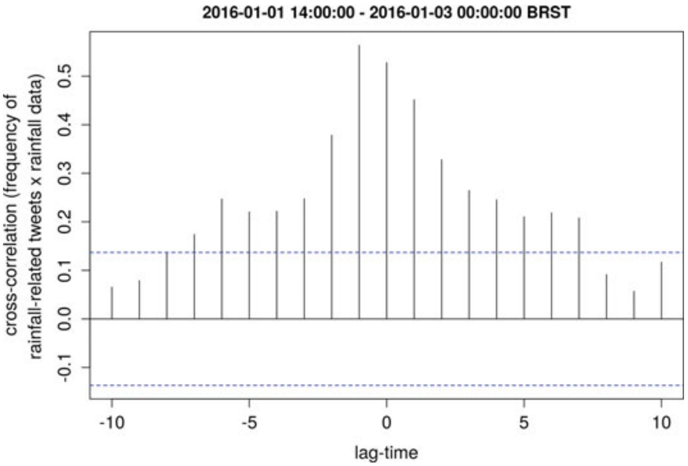


Fig. 6 Cross-correlation and visualization between rainfall data time-series and rainfall-related tweets with a 10-min time-scale from January 1st 14:00 BRST to January 3rd 00:00 BRST 2016

Table 3 Cross-correlation values between rainfall data time-series and rainfall-related tweets with a 10-min time-scale from January 1st 14:00 BRST to January 3rd 00:00 BRST 2016

Lag-time	Cross-correlation	Lag-time	Cross-correlation
−10	0.065	1	0.451
−9	0.078	2	0.328
−8	0.135	3	0.264
−7	0.174	4	0.245
−6	0.246	5	0.210
−5	0.220	6	0.218
−4	0.221	7	0.207
−3	0.247	8	0.091
−2	0.378	9	0.056
−1	0.563	10	0.116
0	**0.528**		

A similar result is achieved in the period 25th–28th January (Fig. 7), i.e., the greatest correlation is −1 lag-time ($k = -1$). In this case, the confidence interval (Fig. 7) and cross-correlation values (Table 4) are highest.

In fact, the results show that the highest correlation is negative, but the positive correlation of 1 lag-time ($k = 1$) could be of value for monitoring activities. For example, the decision-maker might be monitoring what the social network users are saying during the rain phenomenon, as well as assessing the consequences and effects of the rain (e.g. situational updates about an ongoing event). This setting is very useful for monitoring extreme events, such as heavy rainfall episodes. Hence, the lag-time can be used to define a threshold time for monitoring the affected area. For

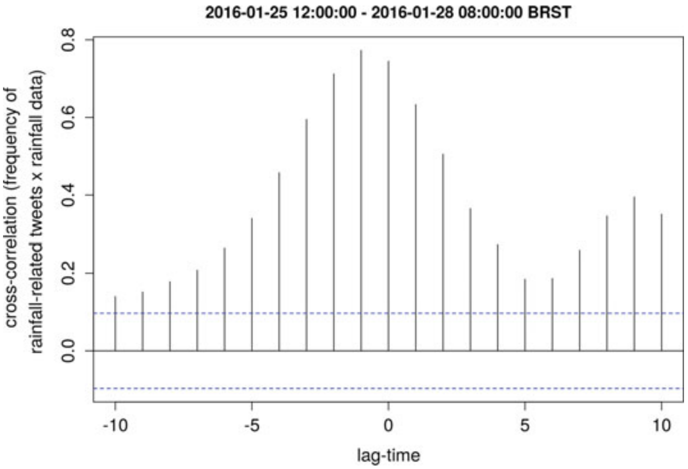


Fig. 7 Cross-correlation and visualization between rainfall data time-series and rainfall-related tweets with a 10-min time-scale resolution from January 25th 12:00 BRST to January 28th 08:00 BRST 2016

Table 4 Cross-correlation values between rainfall data time-series and rainfall-related tweets with a 10-min time-scale from January 25th 12:00 BRST to January 28th 08:00 BRST 2016

Lag-time	Cross-correlation	Lag-time	Cross-correlation
−10	0.140	1	0.633
−9	0.151	2	0.505
−8	0.178	3	0.365
−7	0.207	4	0.273
−6	0.264	5	0.184
−5	0.340	6	0.186
−4	0.457	7	0.259
−3	0.594	8	0.345
−2	0.712	9	0.394
−1	0.772	10	0.350
0	**0.744**		

example, a maximum size of a temporal window of 20 min can be defined for getting information after the event has been detected, i.e. $0 < k < 2$.

Another possible setting is when the lag-time is zero ($k = 0$). In this case, rainfall data and rainfall-related tweets could be analyzed for monitoring activities, either together or separately. The first alternative is ideal to improve the analysis in real-time. The second alternative could be applied when the rainfall data are insufficient. In this case, it is necessary to assess the credibility of the social network messages.

Figure 8 shows the results for the entire period (from January 1st to 30th January 2016). It can be seen that the cross-correlation is greater than 0.5 in the lag-time from

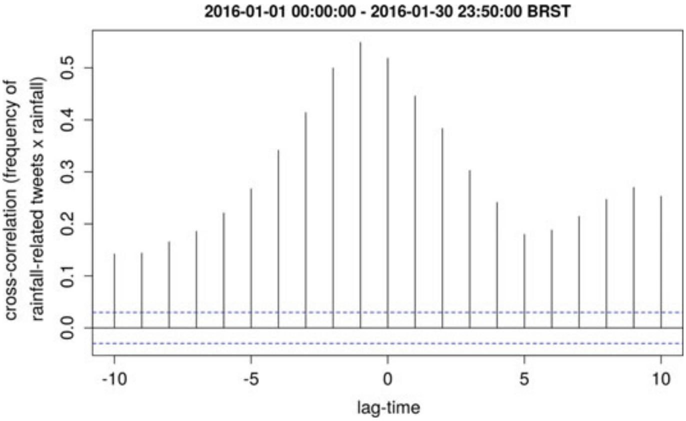


Fig. 8 Cross-correlation and visualization between rainfall data time-series and rainfall-related tweets with a 10-min time-scale in January 2016

Table 5 Cross-correlation values between rainfall data time-series and rainfall-related tweets with a 10-min temporal resolution in January 2016

Lag-time	Cross-correlation	Lag-time	Cross-correlation
−10	0.141	1	0.445
−9	0.143	2	0.383
−8	0.165	3	0.302
−7	0.185	4	0.241
−6	0.220	5	0.179
−5	0.267	6	0.188
−4	0.341	7	0.214
−3	0.413	8	0.247
−2	0.499	9	0.270
−1	0.548	10	0.253
0	**0.518**		

−1 to 0 (Table 5). In brief, the −1 lag-time setting (−10 min) is enough to be used in forecasting rain, whereas the 0 lag-time (0 min) can be used in a real-time setting.

5.2 20 and 30-min Temporal Resolution

Similar results were obtained for the entire period of the case study with 20 and 30-min time-scales (Figs. 9 and 10). Nevertheless these results reveals additional findings. They show clearly the temporal validity problem for large time-scales. As can be seen from Tables 5, 6 and 7, the cross-correlation in the 0 lag-time increases as the time-scale increases too. In this way, the results indicate that a large time-scale affect the temporal cross-correlation, by making the rainfall data time-series converge with the rainfall-related time-series in an “exact correspondence”—i.e., it shows a false view of the scenario for decision-making. Figure 11 compares the cross-correlation values for the time-scales of 10, 20 and 30 min. In additional, it shows that large time-scales are not suitable for detecting and monitoring rainfall-events in real-time.

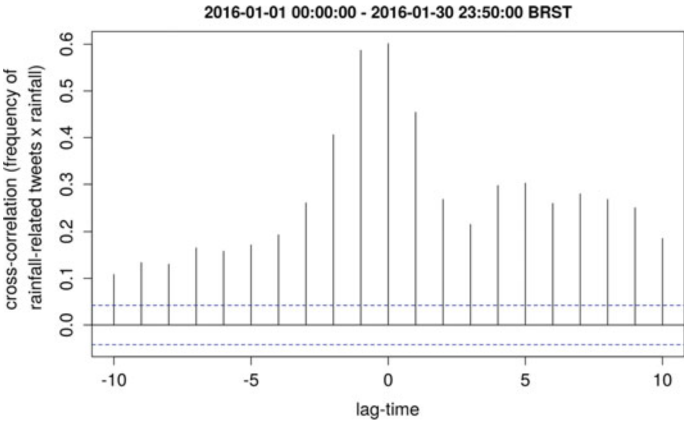


Fig. 9 Cross-correlation and visualization between rainfall data time-series and rainfall-related tweets with a 20-min time-scale in January 2016

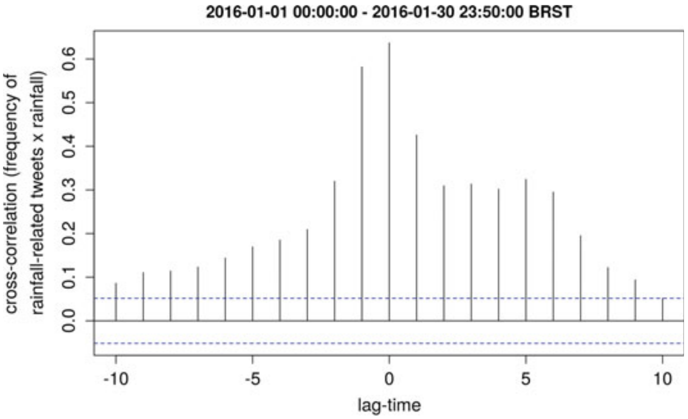


Fig. 10 Cross-correlation and visualization between rainfall data time-series and rainfall-related tweets with a 30-min time-scale in January 2016

Table 6 Cross-correlation values between rainfall data time-series and rainfall-related tweets with a 20-min time-scale in January 2016

Lag-time	Cross-correlation	Lag-time	Cross-correlation
−10	0.108	1	0.454
−9	0.133	2	0.267
−8	0.130	3	0.215
−7	0.165	4	0.297
−6	0.158	5	0.302
−5	0.171	6	0.259
−4	0.193	7	0.279
−3	0.260	8	0.267
−2	0.406	9	0.249
−1	0.586	10	0.185
0	**0.601**		

Table 7 Cross-correlation values between rainfall data time-series and rainfall-related tweets with a 30-min time-scale in January 2016

Lag-time	Cross-correlation	Lag-time	Cross-correlation
−10	0.085	1	0.425
−9	0.110	2	0.309
−8	0.114	3	0.313
−7	0.123	4	0.301
−6	0.143	5	0.324
−5	0.169	6	0.294
−4	0.184	7	0.195
−3	0.209	8	0.122
−2	0.319	9	0.093
−1	0.581	10	0.051
0	**0.637**		

6 Discussion and Conclusion

This paper outlines a temporal approach to explore the cross-correlation between social network messages and rainfall data from a meteorological source. A case study was undertaken in São Paulo, Brazil, and we identified lag-time between a time-series containing data from rainfall and a time-series of rainfall-related tweets. The case study shows clearly the temporal validity problem, as well as the need for a cross-correlation analysis to investigate the use of social network messages in practical solutions, such as monitoring and early warning systems.

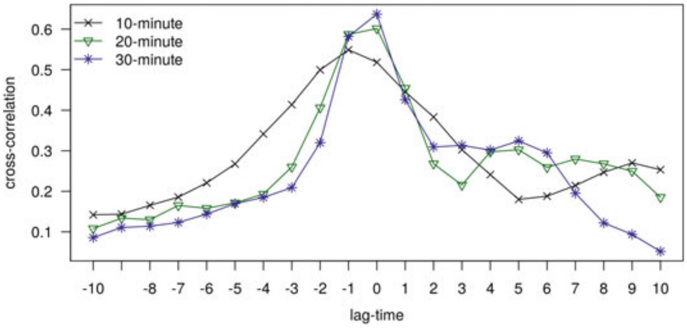


Fig. 11 Comparison of cross-correlation between rainfall data time-series and rainfall-related tweets with 10, 20 and 30-min time-scales in January 2016

The results provide evidence that the rainfall gauge time-series and the rainfall-related tweets are associated with different lag-times (Figs. 6, 7 and 8), but they are highly associated with a lag-time that ranges from −10 to +10 min (Tables 3, 4 and 5). Statistically, the lag-time can vary from negative to positive (see Tables 3, 4 and 5). This temporal characteristic can be hard to detect with the human eye without a temporal correlation. For example, when the analyst chooses a small time-scale, the peaks of rainfall data and rainfall-related social network activity are relatively close. Thus, it is not clear to an decision-maker whether the social network activity is or is not useful for decision-making. However, our results show that the rainfall time-series patterns can be approximated by social network messages when proper account is taken of the duration, frequency, and lag-time. In view of this, our approach can be applied to detect patterns of rainfall events in real-time using authoritative data and social network messages. In addition, the findings suggest that this approach can be useful to fit or approximate the best temporal scale between two time-series in any setting (e.g. real-time, near real-time and post-hoc analysis) since the individual data sources can be represented from a single time-series. This is very useful for monitoring activities, where the scale and frequency of the data can change over time.

Thus, the main value of this work is that we have put forward a novel approach to describe the temporal relationship between authoritative data and rainfall-related social network messages using lag-times rather than only relying on the duration and frequency of the time-series. This opens up a new way of exploring and improving the temporal models and approaches with the aim of creating functional relations to enhance hydrological modelling for monitoring rainfall in real-time. Moreover, it is possible to improve existing detection systems, such as PrioritizeSN (Assis et al. 2015), Toretter (Sakaki and Matsuo 2012), GeoCONAVI (Spinsanti and Ostermann 2013), to name just a few.

Future work should further extend this approach by incorporating other social network platforms (e.g. Instagram and Flickr) and case study scenarios (e.g. other cities and countries) to be able to obtain a generalization. Furthermore, the centroid of the rainfall-related tweets might be explored to understand the extent to which they can be correlated with a rainfall time-series, although a spatial analysis is needed for this. Finally, methods for handling factors of uncertainty can be employed to correct the rainfall gauge data instead of removing them for spatial interpolation.

7 Data Access Statement

All data created during this research are openly available from the University of Warwick data archive at http://wrap.warwick.ac.uk/87173.

Acknowledgements This research was partially funded by the Engineering and Physical Sciences Research Council (EPSRC) through the Global Challenges Research Fund. The authors would like to express their thanks for the financial support provided by the Coordination for higher Education Staff Development (CAPES, Grant No. 88887.091744/2014- 01). S. C. Andrade would like to thank the Araucária Foundation of Supports Scientific and Technological Development in the State of Paraná (FAPPR) and State Secretariat of Science, Technology and Higher Education of Paraná (SETI) for their financial support. C. Restrepo-Estrada is grateful for the financial support from CAPES-PROEX.

References

Ahrens B (2006) Distance in spatial interpolation of daily rain gauge data. Hydrol Earth Syst Sci 10(2):197–208

de Albuquerque JP, Herfort B, Brenning A, Zipf A (2015) A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. Int J Geogr Inf Sci 29(4):667–689

Assis LFG, Herfort B, Steiger E, Horita FEA, de Albuquerque JP (2015) Geographical prioritization of social network messages in near real-time using sensor data streams: an application to floods. In: Proceedings of the XVI Brazilian symposium on geoinformatics, pp 26–37

Bacchi B, Kottegoda NT (1995) Identification and calibration of spatial correlation patterns of rainfall. J Hydrol 165(1):311–348

Bartier PM, Keller C (1996) Multivariate interpolation to incorporate thematic surface data using inverse distance weighting (idw). Comput Geosci 22(7):795–799

Crooks A, Croitoru A, Stefanidis A, Radzikowski J (2013) #earthquake: twitter as a distributed sensor system. Trans GIS 17(1):124–147

Earle P, Bowden D, Guy M (2012) Twitter earthquake detection: earthquake monitoring in a social world. Ann Geophys 54(6):708–715

Herfort B, de Albuquerque JP, Schelhorn SJ, Zipf A (2014) Exploring the geographical relations between social media and flood phenomena to improve situational awareness. Springer International Publishing, Cham, pp 55–71. doi:10.1007/978-3-319-03611-3_4

Imran M, Elbassuoni S, Castillo C, Diaz F, Meier P (2013) Extracting information nuggets from disaster-related messages in social media. In: Proceedings of the ISCRAM 2013—10th international conference on information systems for crisis response and management, pp 791–801

Kryvasheyeu Y, Chen H, Obradovich N, Moro E, Van Hentenryck P, Fowler J, Cebrian M (2016) Rapid assessment of disaster damage using social media activity. Sci Adv 2(3). doi:10.1126/sciadv.1500779

Llasat MC (2001) An objective classification of rainfall events on the basis of their convective features: application to rainfall intensity in the northeast of spain. Int J Climatol 21(11):1385–1400

Mair A, Fares A (2011) Comparison of rainfall interpolation methods in a mountainous region of a tropical island. J Hydrol Eng 16(4):371–383

Marchi L, Borga M, Preciso E, Gaume E (2010) Characterisation of selected extreme flash floods in europe and implications for flood risk management. J Hydrol 394(1–2):118–133 (flash floods: observations and analysis of hydrometeorological controls)

Sakaki T, Matsuo Y (2012) Earthquake observation by social sensors. InTech. doi:10.5772/29629

Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on world wide web, pp 851–860

Spinsanti L, Ostermann F (2013) Automated geographic context analysis for volunteered information. Appl Geogr 43:36–44

Starbird K, Muzny G, Palen L (2012) Learning from the crowd: Collaborative filtering techniques for identifying on-the-ground twitterers during mass disruptions. In: Proceedings of the ISCRAM 2012—9th international conference on information systems for crisis response and management, BC, Simon Fraser University

Steiger E, de Albuquerque JP, Zipf A (2015) An advanced systematic literature review on spatiotemporal analyses of twitter data. Trans GIS 19(6):809–834

Yang X, Xie X, Liu DL, Ji F, Wang L (2015) Spatial interpolation of daily rainfall data for local climate impact assessment over greater sydney region. Adv Meteorol 2015:1–12

Residential Choices as a Driving Force to Vertical Segregation in Whitechapel

Shlomit Flint

Abstract This study examines the impact of habitat choices and householder migration on Interbuilding Vertical segregation in Whitechapel, a diverse inner-city neighbourhood in London. For migrants living in this absorption area, the need for a sense of belonging and continuity leads to the development of micro mechanism that improve the individuals' ability to cope with the urban challenges. Based on residential records at the resolution of single families and flats that cover a period of 17 years, the study reveal and analyse powerful mechanism of residential segregation at the vertical dimension of buildings, which the dwellers are recognise, adjust to and obey. Taken together, this mechanism is a candidate for explaining the dynamics of residential segregation in Whitechapel during 1995–2012.

Keywords Spatiotemporal data acquisition • Modelling • Analysis • Volunteered geographic information • Community observatories

1 Introduction: Non-economic Vertical Segregation—What Do We Know About It?

Ever since cities became large and complex, they developed cultural and urban mechanisms—technological, organisational, legal, and social—to tackle their own pressing problems of demographic and economic growth (Hall 1998). One of these age-old mechanisms, characterizing both ancient cities and current urban centres is Segregation, a well-known residential pattern which is the outcome of householder migration and habitat choices (Boal 1978, 1996; Chivallon 2001). Despite extensive research of the causes and patterns of 17 different forms of neighbourhood segregation—among them the rare types of segregation between the front and the back of residential buildings and overcrowding in similar buildings—the vertical dynamics of ethno-religious enclave and the role of individuals' preferences and

S. Flint (✉)
Centre for Advanced Spatial Analysis, University College London, London, UK
e-mail: shlomit.flint@ucl.ac.uk

A. Bregt et al. (eds.), *Societal Geo-innovation*, Lecture Notes
in Geoinformation and Cartography, DOI 10.1007/978-3-319-56759-4_3

social relations in shaping minority groups' spaces still await deeper insight. This article, however, examines how individuals' identity and local residential preferences play a central role in everyday life and are reflected in the vertical segregation of Whitechapel neighbourhood.

White (1984) described vertical differentiation as the most widespread type of segregation, allowing people from different classes to share and interact in the same living space. This phenomenon characterises mainly compact built heterogeneous cities that developed gradually through individual-level residential decisions (White 1984; Maloutas and Karadimitriou 2001). A description of this mechanism, operated in eighteenth century Paris, is provided by Roche (1987) who outlines it as "inequality began in relation to space". In addition to the horizontal Segregation—the northwest quadrant of Paris was generally inhabited by the wealthy, and the northeast by the proletariat—there was also a Vertical Segregation. As people of different classes often lived together in the same building, there were clear advantages to being wealthy: "The lower storeys were reserved for owners, middle-class master-craftsman, shopkeepers or the principal tenants. Poorer families lived on higher floors, with many more people in each room" (Roche 1987). Nowadays, with the exception of some slum areas and modern housing districts, the middle and working classes live together in vertically stratified apartment blocks: the working class and service labourers live in lower floors while the wealthier on top floors and in penthouses (Allum 1973; Leontidou 1990).

Current research, however, relates vertical segregation to gentrification processes, and claims that the integration of various population groups in the same buildings solves the emerging problems caused by the city's own growth, while simultaneously preventing the formation of slums (Coing 1966; Smith and Williams 1986; Glass 1989). In this process of social redistribution and re-appropriation of residential space in mature urban settings, where change no longer refers primarily to rapid urbanization but to internal reforming, changes of scale and form of segregation are often involved through the diversification of patterns and mechanisms for the social allocation of urban space. Hamnett (2004) illustrates this diversification for East London, where gentrifies gradually replaced the shrinking working class from the 1970s and created more variegated segregation patterns at the micro-scale in the place of former broad socio-spatial divisions. A gentrified area becomes less segregated, in terms of segregation indices, at least for some time, an outcome which in principle should be accounted as positive (Slater 2006). However, since lower levels of segregation could be temporary—as a result from the loss of working class population in a working class area in crisis, the increased internal social mobility within a working class area under regeneration or even the loss of upper and middle class population in areas of filtering down—but actually lead towards more segregated situations, segregation indices are not unequivocally socially negative or positive, and there is a need to recognise the specific context of this social dynamic (Maloutas et al. 2012).

White (1984) makes a distinction between segregation and differentiation. For White, where economic factors are the primary drivers of vertical social separation within a building, this should be considered differentiation; where people are making a conscious choice to separate themselves from other ethnic or religious groups, this should be termed segregation. Since vertical differentiation is expected to be found in societies where class relations (but not ethno-racial differences) are the primary differentiating element in urban space, and where less discrimination and more egalitarian approaches underlie their regulation, White (1984) claims that it would be misleading to call this 'segregation', because it constitutes its antithesis. This research consensus, however, is that vertical segregation is particularly relevant for migrants living in diverse absorption areas, such as the Whitechapel neighbourhood—the case study of this research. Similar to other diverse neighbourhoods in UK and around the world, Whitechapel developed gradually, its character was influenced primarily by its central location and individual-level decision making. As the ethno-religious population groups in this study voluntary segregate themselves, the research thus refers to this phenomenon as "vertical segregation" and examines its cultural-economic characteristics. Whitechapel's residential dynamics offer an example of residential relations between population groups similar in their residential preferences, while different in their economic abilities. These circumstances are reflected in vertical residential patterns, enabling maintenance of unique cultural identity.

2 Theoretical Remarks

Much of the geographic literature deals with segregation resulting from either inner forces encouraging people to congregate as a means to preserve the group's culture, language and customs (Macedo 1995; Boal 1996; Wahlstrom 2005) or from external forces through the spatial exclusion of unwanted groups from the majority group's space (Lee 1977; Boal 1978; Knox 1982). The main factors for analysis of householders' residential behaviour—the socio-economic characteristics of the individual and the household on the one hand, and the socio-economic status of the housing and neighbourhood on the other—identified by Speare only in the mid 1970s (Kasarda 1972, 1978; Speare 1974; Speare et al. 1975). Further inquiry into these factors refined these variables, and now it is common to distinguish between revealed preferences, the actual individual behaviour; and stated preferences, individuals' declared attitudes (Giffinger 1998; Iceland 2004; Ihlanfeldt and Scafidi 2002).

Studies of urban dynamics and residential choice explain segregation by referring solely to economic factors or by looking at a mixture of economic and non-economic factors (Borjas 1998; Clark and Withers 1999). Non-economic factors of segregation, such as family lifestyle, ethnic relationships or life-cycle characteristics, are usually merged with the economic factors (Johnston et al. 2007), and thus blurring the impact of non-economic factors (Flint 2016).

A basic approach to non-economic segregation between householders who belong to one of two groups was offered by Sakoda and Schelling (Sakoda 1971; Schelling 1971, 1974). According to this approach, each householder considers the surrounding population to consist of 'friends'—householders belonging to the same group; and 'strangers'—householders belonging to other groups. Sakoda and Schelling further reduced the non-economic factors influencing the householder's decision to stay or to move to the fraction of 'friends' within the householder's neighbourhood. According to this model, householders aim at residing in a neighbourhood where the fraction of friends, F, is above a certain threshold. In the abstract versions of a model, which consider a square grid of cells, each populated by one householder only, a threshold value of F varies lies within the interval 1/4–1/3, depending on the other model parameters. This means that the tendency of a householder to reside within the neighbourhood where the fraction of friends is above one third eventually results in complete residential segregation. Despite the essential advance in studying Schelling model in its abstract 2D and 3D forms, examples of the real-world dynamics that can be described by the Schelling-like rules are very few (Flache and Hegselmann 2001; Benenson et al. 2002; Bruch and Mare 2006; Fossett 2006a, b).

Research shows that residential choices are determined by socio-cultural-economic interactions (Mobius and Rosenblat 2002; Schnell and Benjamini 2005) and that different levels of social organization play an important role in shaping segregated residential spaces (Knox and Pinch 2000; Christensen and Hogen-Esch 2006; McNair 2006). In this respect, ethno-religious minorities who require spatial congregation for maintaining their meaningful social contacts and lifestyle are usually tend to combine spatial and social segregation. In what follows this study considers Whitechapel's residential pattern as driven by the interactions between householders of different groups and investigates whether the tendency to reside among people of their own groups can explain non-economic vertical residential patterns there.

3 The Case Study of Whitechapel

London's East End (Fig. 1) developed gradually from medieval times, growth quickening in the extra-mural district in the late 16th century. In the late 17th century, Huguenot refugees inhabited a new weaving suburb in Spitalfields, followed in the 18th century by many Irish Catholics, in the late 19th century by Jews and in the late 20th by Bangladeshis (Lupton and Power 2004). Many Jewish and Bengali immigrants worked in the clothing industry with low wages and poor conditions, and from around 1890 the area became associated with poverty, overcrowding, disease and criminality. Official attempts to address under-investment in housing stock through the public sector began in the 1890s under the auspices of the London County Council. World War II devastated much of the East End, leading to dispersal of the population. During the 1950s, the area reflected the structural and social changes of slum clearance and war-time destruction. New public housing was built and a high proportion of immigrants and their descendants eventually found places in council accommodation (Lee and Murie 1997). The closure of the last of the East London docks in 1980 led to attempts at regeneration

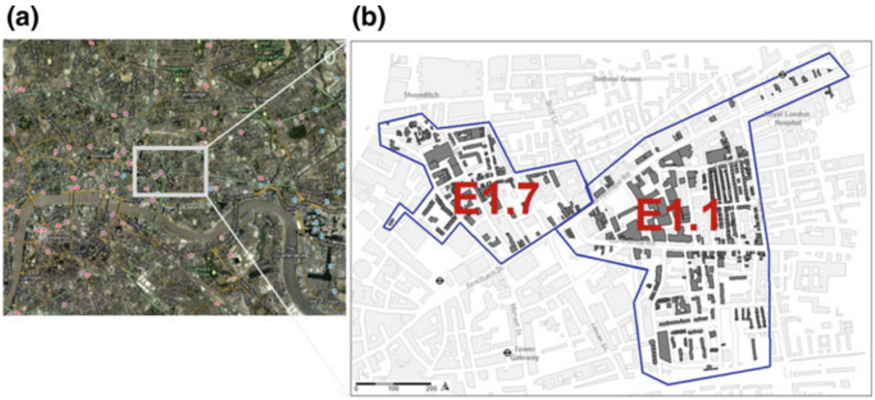


Fig. 1 **a** The research area within its surroundings **b** Postcode areas E1.1 and E1.7

to the south and east of Whitechapel. Subsequently, with its close proximity to the financial centre of London and the strong presence of economic regeneration together with social policy activity, has led to much new development in Whitechapel (Hamnett 2004). However, despite renewal and a massive gentrification process, some parts of the East End have continued to suffer considerable social and economic disadvantage, containing some of the most deprived areas in Britain. These areas are largely populated by the UK's youngest and fastest growing minorities, who are encouraged to preserve tradition based on family ties in compact areas (Kintrea et al. 2008; Dustmann and Theodoropoulos 2010).

Today, the large number of 61 religious institutions in the area reflects the diversity of Whitechapel's population. The area is populated by Muslims, Hindus and Christians of African, British, South-Asian, East-Asian and European origin. Each of these groups has brought its own distinctive customs and traditions to enrich the area's life and culture. The largest sub-group living in the Borough of Tower Hamlets is Bangladeshi-Muslim (30%), representing one of the largest concentrations of Muslim ethnicity in Europe (Carey and Shukur 1985; Dench et al. 2006; Stillwell and Duke 2005). The other groups in the area are relatively small, divided in this research by their place of origin.

3.1 *Construction of Whitechapel's Spatio-Temporal Population GIS*

To investigate residential relationships in the research area among Whitechapel's population groups, a detailed spatio-temporal database that contains exact geo-referenced data on family religious affiliation was constructed. The field research was conducted during 2011–12 at the level of individual families and flats. Together with a local interviewer, a young male from the Bangladeshi community (who has requested anonymity), the author conducted a door-to-door survey and interviewed 4656 Families living in 3186 Flats. As the interviewer was already familiar with the Bangladeshi civil community, and the author speaks Bengali, they were able to gather rich and sometimes controversial data by this means. The households were asked to identify themselves as well as the flat's former dwellers, going back to at least 1995. Several researchers stress that the identity of previous residents is important for traditional families (Waterman and Kosmin 1987, 1988), a conclusion confirmed by this research. Identification of past residents allowed us to understand which group's members had occupied each flat for the past 20 years. All other questions asked related to the present occupants in order to ascertain their socio-spatial behaviour.

Householders were also questioned about motives for choosing the flat, and asked to rank the relative importance of the flat's price, their neighbours' identity and institutional (e.g., Churches and schools) proximity (stated preferences). This field survey also collected data about the flat cost, the location of institutes and services that the families attend (revealed preferences), ownership versus rental of the flat, and the source of information about flats prior to buying or renting. Despite early apprehensions regarding cooperation, the response rate reached 83%. A high level of cooperation with the survey enables a comparison between stated and revealed preferences and recognizes similar preferences amongst the groups.

In order to complete the fundamental part of the research, 172 interviews had conducted with key figures such as community leaders, municipal planners and real estate agents. Interviewees were chosen on the basis that they offer a range of different types of knowledge and perspectives on their community. Among the interviewees also reside the last of the veteran inhabitants who could provide explanations and describe the processes taking place in the neighbourhood from their point of view. The cross-referenced data produce information on the population exchange and express the dynamic processes. The interviews will also assist in identification of further key contacts.

Construction of the Whitechapel GIS was based on layers updated to 2011 and provided by the Ordnance Survey.[1]

The characteristics of all the research area's flats and households were organized as GIS layers, in which every record in the table is related to the corresponding building. The layer was then included in the area's high-resolution GIS. A quality control process ensured consistency. This involved piloting, whereby the interviewer was required to carry out three pilot interviews to refine approaches and questions where necessary, and ongoing basis review, whereby the interviewer's field notes were reviewed weekly to ensure consistency across the project and that relevant data was picked up.

[1] http://ordnancesurvey.co.uk/opendatadownload/products.html.

Whitechapel's GIS contains additional layers pertaining to topography, roads, land parcel, and buildings, the latter characterized by use and number of floors. There are 1,149 families in 47 communal buildings and 3,507 families in 1,615 privately owned flats in 241 buildings.

Taken as a whole, the survey's spatio-temporal GIS enables evaluation of residential patterns at the resolution of flats, buildings, and neighbourhood; it thus makes investigation of the residential micro-dynamics in this limited environment empirically possible.

3.2 Estimation of Residential Segregation

The level of spatial segregation at the resolution of buildings was estimated with the Moran I index (Zhang and Linb 2007) of spatial autocorrelation. The Moran I index was applied for estimating the correlation between the fraction of a given group D in building i and the fraction of D over the buildings U(i) that are adjacent to i:

$$\frac{N\sum_i \sum_{j\in \bigcup(i)} w_{ij}(D_i-\bar{D})(D_j-\bar{D})}{\left(\sum_i \sum_{j\in \bigcup(i)} w_{ij}\right)\sum_i (D_i-\bar{D})^2}$$

where N is the number of buildings and $\bar{D}$ the average fraction of a group D in Whitechapel. The influence w_{ij} of the neighbouring buildings $j \in \cup(i)$ on i is calculated as $w_{ij} = 1/N\cup(i)$, where $N\cup(i)$ is the number of buildings in $\cup(i)$. The

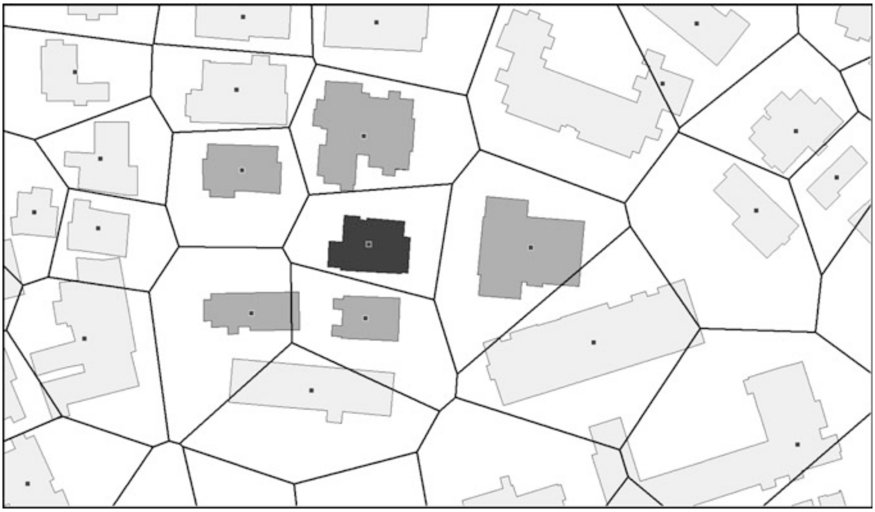


Fig. 2 Whitechapel buildings and the coverage of Voronoi polygons constructed based of buildings' centroids. Voronoi-neighbours of the selected building (*Black*) are shown in *Gray*

proximity of buildings is defined by a Voronoi partition constructed on the basis of the buildings' central points, as proposed by Benenson et al. (2002). According to this definition, two buildings are adjacent if the central points of their foundations are directly visible by the other (Fig. 2):

4 Residential Segregation in Whitechapel

4.1 Whitechapel Population Dynamics

Whitechapel's population grew until the 2000s, in tandem with the East-European migration and the construction of new apartment buildings. In 1995 the area was populated mainly by people from South Asian origin, (most of them were from India and Bangladesh, with negligible Pakistani and Sri Lanka groups), and the general British population, with Bangladeshis steadily substituting the other South Asian groups (Fig. 3). During the 2000s, new population groups entered the area, and nowadays we can recognize 30 sub-groups.

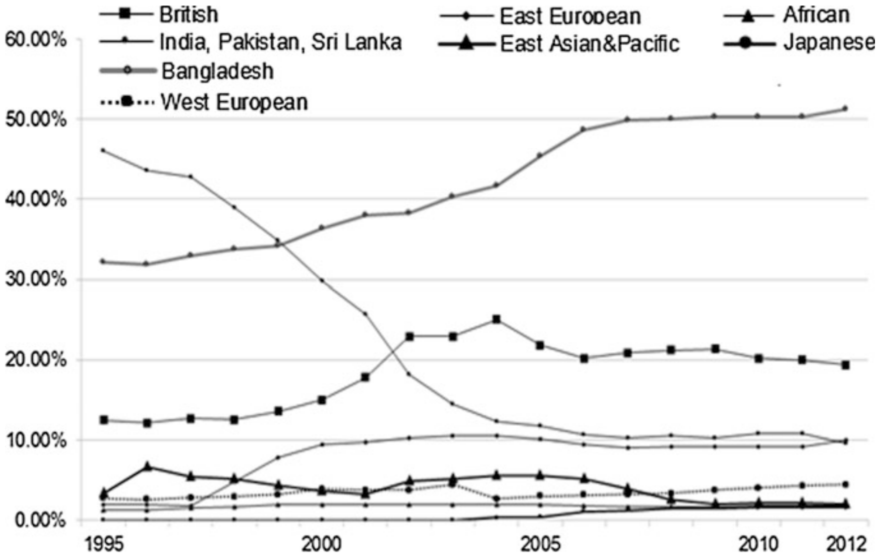


Fig. 3 Population dynamics in the research area of Whitechapel (percentages)

Table 1 Importance of flat cost, neighbours' identity and proximity to institutions in flat choice by population group, Whitechapel (2015)

Group factor	British	West European	East-Asian and Pacific	Japanese	East-European	African	India, Pakistan, Sri Lanka	Bangladesh
Price	**76%**	**55%**	**76%**	**76%**	4%	14%	9%	16%
Institutions	6%	13%	0%	4%	**81%**	26%	5%	25%
Neighbours	18%	32%	24%	20%	15%	**60%**	**86%**	**59%**
N	879	212	83	63	453	129	431	2416

4.2 Stated Residential Preferences of Whitechapel Householders

Table 1 shows that several population groups in Whitechapel share similar concerns (chi-square test, $p \sim 0.5$). Only the East-European dwellers chose the location of institutions as their main concern. Whitechapel's location close to the city centre ensures the proximity of such institutions. Contrary to economic theory, only one third of the population indicated that price was a critical issue for them. Most important rather is the fact that, despite the neighbourhood's reputation as a migrant's neighbourhood, the majority of Whitechapel dwellers reported that the identity of their immediate neighbours is their principal concern. As this stated preference appears to be in the first or second place and shared by members of all groups, the research can assume that the Schelling-like mechanism of actively distinguishing between "friends" and "others" remains relevant in Whitechapel. Apparently, most of the neighbourhoods' dwellers feel the need for at least a few "friends" in order to feel at home in their apartment building. What are the spatial consequences of the above stated preferences? Are they also expressed in the vertical dimension of Whitechapel residential pattern? The study thus turned to investigate the impact of these declared preferences on the revealed preferences of Whitechapel's dwellers based on the data of 1995–2012, when Whitechapel infrastructure remained almost steady yet residential patterns changed.

4.3 Whitechapel Residential Pattern at the Neighbourhood's Level

Based on the survey records, Whitechapel's residential patterns had re-constructed from 1995 until 2012 (Fig. 4). In cases of strong tendency to reside in a friendly environment, Schelling's model results in complete spatial segregation.

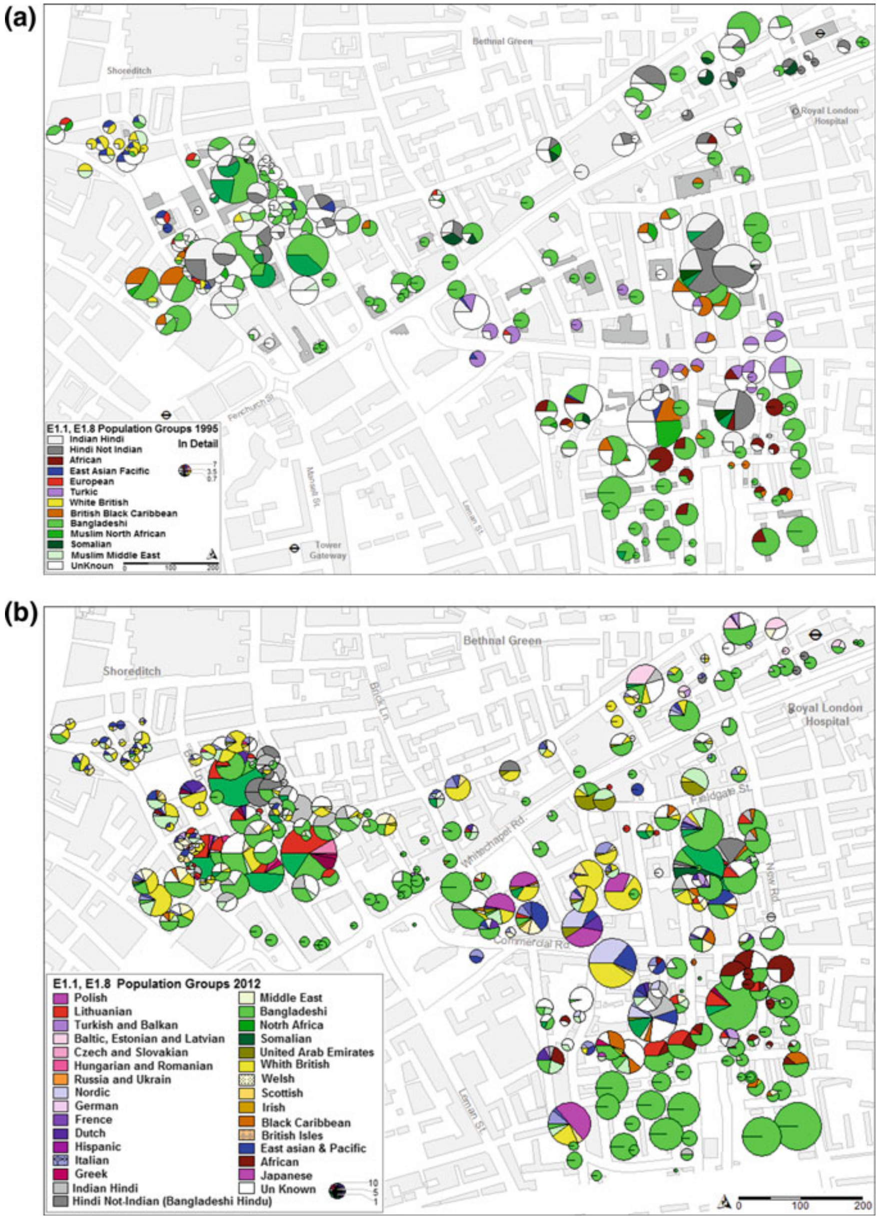


Fig. 4 Spatial distribution of various population groups in apartment buildings in Whitechapel, 1995 (**a**) and 2012 (**b**)

Despite the clear tendency to segregate, though, the maps in Fig. 4 indicate the spatial integration of Whitechapel's residents, with members of several groups living in close proximity to each other. Quantitative estimation of the level of segregation is thus necessary. Moran I index (Fig. 5), Indicates a significant level of segregation exists throughout the entire period for Bangladesh, and that they are the most highly segregated groups in Whitechapel, although the residential segregation of the other religious groups has been steadily growing over the years. In 1995 Moran's I index appears high for the South Asian groups from India, Pakistan and Sri Lanka, that start then start to decline, with most of these groups having left the area by 2004.

Unlike the Schelling's model assumption, the capacity of Whitechapel's spatial units (buildings) is essentially higher than one family. Let us investigate the segregation processes in Whitechapel at the vertical level of residential buildings.

5 Whitechapel Vertical Segregation

5.1 Vertical Segregation in the Individual Building Level

Analysing the revealed preferences of Whitechapel's resident according to their faith (Table 2), shows that 80.5% of the flats occupied by families that identify themselves as Muslims are located in the upper quarter of the building. Since people that identify themselves as Christians or have no religion demonstrated a non-segregated pattern in the bottom quarter, the Muslims' tendency to live in the upper parts of the buildings demands further inquiry.

Note that although Fig. 5 shows that the African group is segregated, they live mainly in council housing, and the estimation of their group's segregation in the areas' private-ownership's buildings (0.4%) is insufficient. Therefore, the major

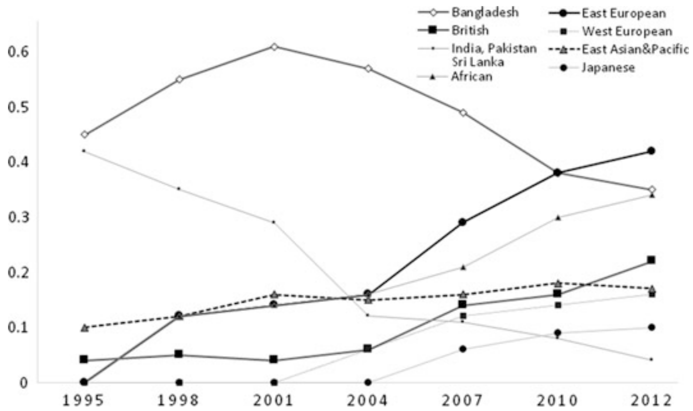


Fig. 5 The dynamics of groups' residential segregation in Whitechapel, according to Moran I index, during the period of 1995–2012

Table 2 Distribution of families in buildings by faith

	Living in the…			
	Upper quarter		Bottom quarter	
	N	of the flats	N	of the flats
Muslims	**1629**	**80.5%**	84	3.5%
Christians (Inc. CofE, Catholic, Protestant and other denominations)	108	10.2%	**91**	**24.8%**
No Religion	28	2.5%	**86**	**22.8%**
Hindus	24	2.2%	178	48.7%
Sikhs	51	4.1%	0	0

Table 3 Distribution of families in buildings by Muslim groups

	Living in the …					
	Upper quarter			Bottom quarter		
	N	Of the flats	Of the total pop. group	N	Of the flats	Of the total pop. group
Bangladeshi	**1268**	**67.6%**	**52.5%**	**79**	**3.2%**	**3.2%**
Middle East	**22**	**5.5%**	**75.8%**	**5**	**0.3%**	**17.1%**
North African	19	1.8%	55.8%	0	0	0
United Arab Emirates	**59**	**5.6%**	**89.3%**	0	0	0

Muslim groups living in the area are from Bangladesh, the Middle East, North Africa and United Arab Emirates origin.

Table 3 shows that Bangladeshi, Middle East and the United Arab Emirates groups reveal similar residential choices to live in the upper parts of the buildings.

There are no lifts in four- to six-story buildings, nor loft storage nor concierge; moreover, the price difference between the bottom and upper quarters is within normal levels for London. The study therefore went on to examine the average income of the groups, as stated by the residence, according to the location of the flat in the buildings (Fig. 6).

Figure 6 shows that families from various groups are also different in terms of their economic capabilities. While the stated average income of Families from United Arab Emirates living in Whitechapel is the highest, that of the Bangladeshi group is much lower, up to 60% of the average London salary. Nevertheless, the differences between the stated average incomes of families living in the Upper quarter and the rest of the building are marginal for all the groups. This indicates

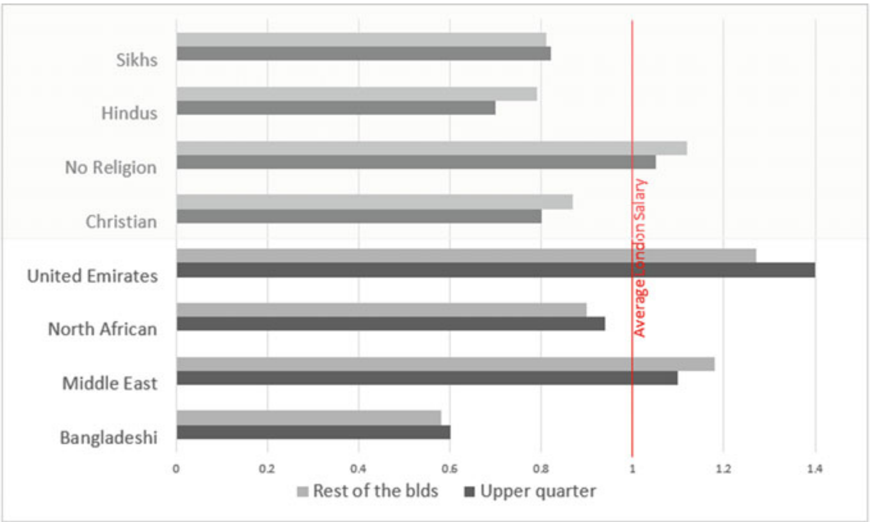


Fig. 6 Stated average income (including family support) by group and location of the flat in the buildings 2012. *Top* by faith, *Bottom* by Muslims sub-groups. Source for the Average London Salary: Labour Force Survey, ONS 2012

that although better economic ability provides more opportunities in a free housing market, the revealed preference to live in the upper floors of Whitechapel's buildings is not necessarily related to economic status. The study thus turns to examine the segregation as an outcome of the relationships between the groups.

5.2 *Inter-building Relationships*

To estimate the relations of a group D with the members of the rest of groups, the study examined the distribution of the number of D-families mD in Whitechapel buildings in 2012. Let the whole fraction of the D-families in Whitechapel be d. If D-families are neutral to the other groups, then the distribution of mD in Whitechapel's buildings with n flats will be binomial, $mD \sim B(d,\ n)$. The comparison between the actual distribution of mD and $B(d,\ n)$ enables recognizing the particular group that is not neutral to the rest of the groups. To combine the results for different n, the study thus transforms mD into $\xi = (mD\ -\ nd)/v(nd(1-d))$: if n is large enough and m is binomial, then ξ is distributed according to the normal distribution N(0; 1) (Von Collani and Dräger 2001). To compare the distribution of ξ to the N(0; 1) the Kolmogorov-Smirnov test was employed (Corder and Foreman

Table 4 Group segregation in the buildings, 2012; note that total population percentage of four population groups in Whitechapel is 91%

	Bangladesh	East-European	India, Pakistan, Sri Lanka	British
Population percentage of a group	51.8%	10.5%	10%	19%
Significance of the K-S criterion	p=0.012	p=0.000	p=0.002	p=0.78
Percentage of buildings without D, based on binomial distribution	**0.4%**	**31.4%**	**32.1%**	14.1%
Percentage of buildings without D, real	**7.8%**	**44.6%**	**39.7%**	15.9%
D-percentage in populated buildings, based on binomial distribution	25.3%	23.6%	21.4%	21.1%
D-percentage in populated buildings, real	32.4%	31.1%	23.2%	21.8%

Table 5 Averaged over 1995–2012 probability to replace a family of an own group

Period	Bangladesh	East-European	India, Pakistan, Sri Lanka	British
1995–2012	0.99	0.49	0.46	0.74

2009). Table 4 demonstrates that the families of each group, besides British, tend to segregate from the others. The fraction of buildings where the group is not found and the average fraction of the group in the buildings where it is found, are essentially higher than should be expected in case of the binomial distribution.

Members of each group tend to reside in flats vacated by householders of their own group, which can be considered as an expression of their stated preferences. For the group *D* these probabilities are calculated as *DReplacing_D/DLeft*, where *DReplacing_D* denotes the number of families of a group *D* that replaced the families of *D* during the year, and *DLeft* the overall number of the families of *D* that left during the year (Table 5).

To conclude, when averaged over the period of 17 years, East-European, Bangladeshi and other South Asian groups in Whitechapel are segregated within the buildings (Table 4) two first of them are segregated within the neighbourhoods (Fig. 5). The replacement of the tenant of the same group (Table 5) is a strong candidate mechanism for supporting this segregation in time.

5.3 Inter-floor Relationships

Schelling's lesson is that people prefer to enter a flat in buildings where the residence rate of their group is significantly higher than the percentage of their group in the population. But how could they know about the building's composition? And how come that families, sometimes with young children, prefer to live in the upper floors of a building without a lift? Could it be that the "identity" of a building is affected also by vertical segregation?

In this study, 141 interviews revealed that the visibility of ethnicity in the buildings affects the identity of the building and the entry of other populations into it. The interviewees mentioned mainly the mixing of public and private usages in the building level such as seating and talking outside internal doors, children's games in the stairwell, hanging laundry in common areas and smell of cooking "push" or "pull" them to a specific building. Apparently visibility of identity hint about the composition of the buildings. In this regard, 58 interviewees who live in the Upper quarter of buildings explain their benefits in living there. One of the interviewee is Afia (age 32) from Bangladesh who explains: *"When you live upstairs your contact with the building increase. In this weather the children play mainly in the stairwell, people meet there and actually this is the living room of the building. When other of us seeking for a flat they immediately recognise that we live here and join. Having other around make the area more secure and the kids happy to have friends, don't need to go outside"*. Khalid (early 40s) from United Arab Emirates adds: *"We are all migrant here and whoever live upstairs designated the building with his (group). I live here almost two years and I learned that it is better like that when I was looking for a flat"*.

Another explanation for the preferences to live in the upper quarter of buildings supplied by Rimi (early 50s) from Bangladesh, who explains: *"in the upper level you can attach part of the hallways to your flat and increase it, so you can live with your siblings' families"*. Examining of this statement about the semi-private/public space reveal that although one may expected to find about 937 Muslims families—which are 80.5% of the upper quarter of Whitechapel families—to live in the upper quarter, the data indicates that 1629 families actually lives there. Table 6 shows that level of density is varied between groups. Comparing this data with the stated average income (Fig. 6), one can safely assume that it is a non-economic tendency.

Table 6 Level of density in the Upper quarter of buildings for certain population groups

Upper quarter of buildings	Bangladeshi	Middle East	North African	United Arab Emirates
Population percentage of a group	27.2%	0.5%	0.4%	1.2%
Level of density (families per flat) upper	4.6	0.9	2.6	2.6

6 Conclusions: Local Residential Choices as a Driving Force to Non-economic Vertical Segregation in Whitechapel

This study examines the impact of ethno-religious identity on Inter-building Vertical segregation, focusing on Whitechapel, an inner-city diverse neighbourhood in London. The literature describes the global cities of their time, Renaissance Florence between 1400 and 1450; Shakespearean London, Vienna in the 18th and 19th centuries; and Paris between 1870 and 1910, assumes that economic forces, namely, economic classes and employment status are the main driving forces of vertical differentiation (White 1984; Hall 1998). This study, however, providing an extraordinary opportunity to explain this phenomenon as a result of habitat choices and householder migration, and thus recognise it as Vertical segregation.

Although it was expected to find that better economic ability provides more opportunities and better chances to bridge the gap between the stated and revealed residential preferences in a free housing market, the study surprisingly indicates that the revealed preference to live in the upper floors of Whitechapel's buildings is not necessarily related to economic status. Despite different economic abilities shown by Whitechapel's dwellers stated average income and location of the flat in the buildings (Fig. 6), Table 2 shows that 80.5% of the flats occupied by families that identify themselves as Muslims are located in the upper quarter of the buildings. Most of this dwellers are Bangladeshi group whose stated average income is up to 0.6% of the average London salary. Although this is much lower than the 1.4% stated average income of United Emirates, both groups reveal similar preferences to live in the upper quarter of the buildings. The high probability to the replacement of the tenant of the same group as expose in Table 5 is a strong candidate mechanism for supporting this segregation in time.

Previous reserach on Sanhedria (Flint et al. 2012) coined the term 'Micro-segregation'. For migrants living in diverse absorption areas, such as the Whitechapel neighbourhood, the human need for a sense of belonging and continuity may lead to the development of micro mechanisms to improve the individuals' ability to cope with the challenges of urban life. The unique information collected via a comprehensive census, reveals powerful mechanisms that govern this segregation. Relatively high density of families per flat as shown in Table 6 together with small number of families from Middle East, North African and United Arab Emirates origin blurs this study's ability to apply Kolmogorov-Smirnov test on the vertical dimension. However, examining the distribution of Whitechapel's resident according to their faith in the vertical dimension and applying the K-S test on the inter-buildings level produces information on the vertical dynamic processes, while analysing the in-depth interviews provides a glimpse into why they might choose to live in the upper quarter of the buildings. We can see that similar to the age-old economic vertical differentiation, ethno-religious Vertical segregation allows people from different population groups to share and interact in the same living space. Unlike the vertical differentiation, as people of different classes lived together in the same building, there is no clear advantages to being wealthy.

The Whitechapel study has revealed that although the urban fabric may look as a patchwork of economic activity, class relations and cultural-ethno-racial mosaic, there is a clear inner-order, identity-based, which both current residents and newcomers recognise, adjust to and obey. Could it be that there are more other latent inner-orders operating in this area? To what extent does this mechanism affect the area as a whole? Going beyond Whitechapel—can we find non-economic Vertical segregation elsewhere? Is this Inter-building Vertical mechanism another form of segregation, additional to the 17 already recognised by White (1984), or is it a latent micro order which could integrated with other forms? The answers to these questions demand the development and studying of a 3D Schelling-like model that accounts for the buildings of varying capacity and neighbourhoods of a varying shape and size. The results of this study will be presented in the next paper.

References

Allum PA (1973) Politics and society in post-war Naples. Cambridge University Press, Cambridge

Benenson I, Omer I, Hatna E (2002) Entity-based modeling of urban residential dynamics—the case of Yaffo, Tel-Aviv. Environ Plann B 29:491–512

Boal FW (1978) Ethnic residential segregation. In: Herbert DT, Johnston RJ (eds) Social areas in cities. Wiley, London, pp 57–95

Boal FW (1996) Integration and division: sharing and segregating in Belfast. Plann Pract Res 11 (2):151–158

Borjas G (1998) To ghetto or not to ghetto: ethnicity and residential segregation. NBER Working Papers 6176, Natl Bur Econ Res 44(2):228–253

Bruch EE, Mare RD (2006) Neighborhood choice and neighborhood change. Am J Sociol 112 (3):667–709

Carey S, Shukur A (1985) A profile of the Bangladeshi community in East London. J Ethn Migr Stud 12(3):405–417

Chivallon C (2001) Religion as space for the expression of Caribbean identity in the United Kingdom. Environ Plann D: Soc Space 19:461–483

Christensen T, Hogen-Esch T (2006) Local politics a practical guide to governing at the grassroots. M E Sharpe Inc, United States

Clark WAV, Withers S (1999) Changing jobs and changing houses: mobility outcomes of employment transitions. J Reg Sci 39(4):653–673

Coing H (1966) Rénovation urbaine et changement social. Les Éditions Ouvrières, Paris

Corder G, Foreman D (2009) Nonparametric statistics for non-statisticians: a step-by-step approach. Wiley, NY

Dench G, Gavron K, Young M (2006) The new east end: kinship, race and conflict. Profile Books, London

Dustmann C, Theodoropoulos N (2010) Ethnic minority immigrants and their children in Britain. Oxford Econ Pap 62(2):209–233

Flache A, Hegselmann R (2001) Do irregular grids make a difference? Relaxing the spatial regularity assumption in cellular models of social dynamics. J Artif Soc Soc Simul 4(4). http://www.soc.surrey.ac.uk/JASSS/4/4/6.html

Fossett MA (2006a) Ethnic preferences, social distance dynamics, and residential segregation: theoretical explorations using simulation analysis. Math Sociol 30(3–4):185–273

Fossett MA (2006b) Including preference and social distance dynamics in multi-factor theories of segregation. Math Sociol 30(3–4):289–298

Flint S, Alfasi N, Benenson I (2012) Between friends and strangers: micro-segregation in a Haredi Neighbourhood. City and Community 11(2):171–197

Flint S (2016) A decision not to decide: a new challenge for planning. Eur Plann Stud. doi:10.1080/09654313.2017.1302411

Giffinger R (1998) Segregation in Vienna: impacts of market barriers and rent regulations. Urban Stud 35(10):1791–1812

Glass R (1989) Cliches of urban doom. Blackwell, Oxford

Hall P (1998) Cities in civilization: culture, technology, and urban order. Weidenfeld & Nicolson, London, New York

Hamnett C (2004) Economic and social change and inequality in global cities: the case of London. Greek Rev Soc Res 113A:63–80

Iceland J (2004) Beyond black and with metropolitan residential segregation in multi ethnic America. Soc Sci Res 33:248–271. Pantheon Books

Ihlanfeldt K, Scafidi B (2002) Black self segregation as a cause of housing segregation: evidence from the multi-city study of urban inequality. J Urban Econ 51:366–390

Johnston RJ, Poulsen M, Forrest J (2007) The geography of ethnic residential segregation: a comparative study of five countries. Ann Assoc Am Geogr 97(4):713–738

Kasarda JD (1972) The theory of ecological expansion: an empirical test. Soc Forces 51:165

Kasarda JD (1978) Urbanization, community, and the metropolitan problem. In Street D (ed) Handbook of contemporary urban life. Jossey-Bass, San Francisco, pp 27–35

Kintrea K, Bannister J, Pickering J, Reid M, Suzuki N (2008) Young people and territoriality in British cities. Joseph Rowntree Foundation, York

Knox P (1982) Urban social geography: an introduction. Longman, London

Knox P, Pinch S (2000) Social interaction in urban environments. Urban Soc Geogr 7:219–223

Lee TR (1977) Race and residence. Oxford University Press, Oxford

Lee P, Murie A (1997) Poverty, housing tenure and social exclusion. The Policy Press in association with the Joseph Rowntree Foundation, UK

Leontidou L (1990) The Mediterranean city in transition. Cambridge University Press, Cambridge

Lupton R, Power A (2004) Minority ethnic groups in Britain, CASE Brooking Census Briefs No. 2

Macedo S (1995) Liberal civic education and religious fundamentalism: the case of God John Rawls? Ethics 105:468–496

Maloutas T, Karadimitriou N (2001) Vertical social differentiation in Athens: alternative or complement to community segregation? Int J Urban Reg Res 25(4):699–716

Maloutas T, Arapoglou VP, Kandylis G, Sayas J (2012) Social polarization and de-segregation in Athens. In: Maloutas T, Fujita K (eds) Residential segregation in comparative perspective. Ashgate, Farnham, pp 257–283

McNair D (2006) Social and spatial segregation: ethno-national separation and mixing in Belfast. Unpublished PhD dissertation, School of Geography, Archaeology and Paleoecology, Queen's University, Belfast

Mobius M, Rosenblat T (2002) The process of ghetto formation: evidence from Chicago. Manuscript, Harvard University

Roche D (1987) The people of Paris: an essay in popular culture in the 18th century, University of California Press

Sakoda JM (1971) The checkerboard model of social interaction. J Math Sociol 1(1):119–132

Schelling T (1971) Dynamic models of segregation. J Math Sociol 1(1):143–186

Schelling T (1974) On the ecology of micro-motives. In: Marris R (ed) The corporate society. Macmillan, London

Schnell I, Benjamini Y (2005) Globalisation and the structure of urban social space: the lesson from Tel Aviv. Urban Stud 42(13):2489–2510

Slater T (2006) The eviction of critical perspectives from gentrification discourse. Int J Urban Reg Res 30(4):737–757

Smith N, Williams P (1986) Gentrification of the city. Allen & Unwin, London

Speare Alden (1974) Residential satisfaction as an intervening variable in residential mobility. Demography 11(2):173–188

Speare A, Goldstein S, Frey WH (1975) Residential mobility, migration, and metropolitan change. Ballinger, Cambridge, Mass

Stillwell J, Duke-Williams O (2005) Ethnic population distribution, immigration and internal migration in Britain. What evidence of linkage at the district scale. In: BSPS (ed) British society for population studies. University of Kent at Canterbury, Canterbury

Von Collani E, Dräger K (2001) Binomial distribution handbook for scientists and engineers. Birkhauser

Wahlstrom AK (2005) Liberal democracies and encompassing religious communities: a defense of autonomy and accommodation. J Soc Philos 36(1):31–48

Waterman S, Kosmin BA (1987) Residential change in a middle-class suburban ethnic population: a comment. Trans Inst Br Geogr (N.S.) 12(1):107–112

Waterman S, Kosmin B (1988) Residential patterns and processes: a study of Jews in three London boroughs. Trans Inst Br Geogr 13:79–95

White P (1984) The West-European city. A social geography. Longman, London

Zhang T, Linb G (2007) A decomposition of Moran’s I for clustering detection. Comput Stat Data Anal 51:6123–6137

Reference Resolution for Pedestrian Wayfinding Systems

Jana Götze and Johan Boye

Abstract References to objects in our physical environment are common especially in language about wayfinding. Advanced wayfinding systems that interact with the pedestrian by means of (spoken) natural language therefore need to be able to *resolve* references given by pedestrians (i.e. understand what entity the pedestrian is referring to). The contribution of this paper is a probabilistic approach to reference resolution in a large-scale, real city environment, where the context changes constantly as the pedestrians are moving. The geographic situation, including information about objects' location and type, is represented using OpenStreetMap data.

Keywords Pedestrian navigation · Wayfinding · Data-driven methods · Reference resolution · Natural language processing · Openstreetmap · Probabilistic approach

1 Introduction

When humans give wayfinding instructions to each other, they are extensively using referring expressions, phrases that are referring to objects and actions about which they want to convey information. The hearer needs to link the words to representations of these entities, making several choices along the way, and taking different sources of information into account: Is the speaker talking about a landmark in the immediate vicinity? Is he referring to something that has recently been mentioned? Which of the objects match his descriptions and which one is most likely to be the correct target?

If the conversation involves solving a task such as finding the way in an unknown area, it is not enough to understand the meaning of the word "bakery" in an instruction like "Turn left at the bakery with the blue sign" in a general way. Not only does the hearer need to know what a bakery is in a general sense, he also needs to identify

J. Götze (✉) · J. Boye
KTH Royal Institute of Technology, Stockholm, Sweden
e-mail: jagoetze@kth.se

J. Boye
e-mail: jboye@kth.se

A. Bregt et al. (eds.), *Societal Geo-innovation*, Lecture Notes
in Geoinformation and Cartography, DOI 10.1007/978-3-319-56759-4_4

the *target* object, in this case the particular bakery in his environment in order to carry out the task successfully. That means he needs to ground the meaning of the word 'bakery' in the real world (or some representation of it). Methods that automate the understanding and grounding of referring expressions in the physical environment are required for a number of applications in which robots need to carry out actions of various kinds, such as grasping particular objects (Matuszek et al. 2014) or following route directions (MacMahon et al. 2006).

In this work, we focus on reference resolution for pedestrian wayfinding. Wayfinding instructions typically involve many references to landmarks (Denis 1997), i.e. to objects in the environment of the pedestrian. At each point along a route in a city environment, there are many geographical objects of different types, such as buildings and streets, that a pedestrian can refer to. Automatically understanding exactly which objects someone is referring to is an important part of interactive wayfinding systems. The pedestrian might ask clarification questions such as "Do you mean the red building to the right?" or signal problems of understanding such as "I cannot see any church, but I can see a shop straight ahead." Situations in which the system refers to a landmark that the pedestrian cannot identify are unavoidable, and lengthy subdialogs where the system "tries" the next-best landmarks can instead be replaced by letting the user choose a landmark, e.g. by asking an open-ended question like "What can you see?".

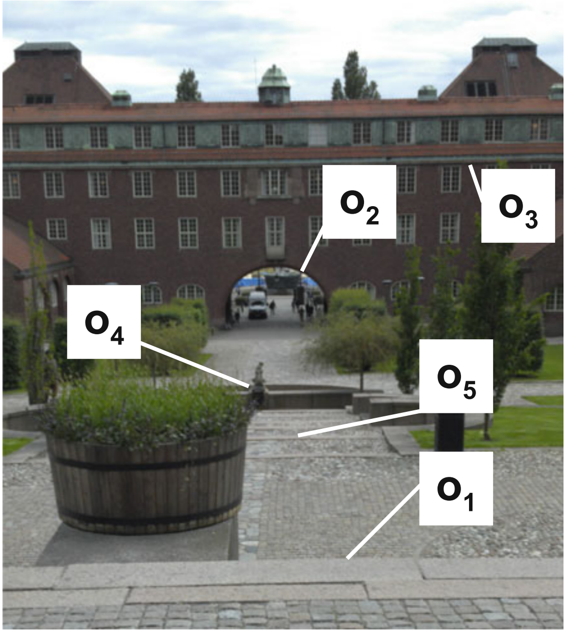
The contribution of this paper is to show how a probabilistic approach to reference resolution can successfully be applied to difficult real-life situations. We base our research on a corpus of route instructions, that are given by pedestrians while they are walking along a path (Götze and Boye 2016b). In this setting, the environment is rich in geographical objects of various kinds that a pedestrian could possibly refer to, and it changes continuously as they are moving. We show how the words-as-classifiers method applied by Kennington and Schlangen (2015) to a toy domain can be applied to our data on the basis of the OpenStreetMap representation of the pedestrians' environment. We then explain how we extend the original method to deal with frequently occurring phenomena in this context: that the referring expression has no or several target objects.

2 Reference Resolution Method

Whenever the pedestrian uses a referring expression (RE), we want to identify the *target object(s)*, the object or objects that the pedestrian intended to refer to when saying the words in the RE. In a probabilistic framework, we want to find the o that maximizes $P(o|r)$ for some set of objects, i.e. the object o that most likely was referred to by the set of words r in the RE.

In the domain we consider, the pedestrians are walking along a route and are primarily referring to objects in their immediate environment. As they are moving along the path, objects are appearing and disappearing from their view: the set of objects that are possible referents for their descriptions—*the candidate set*—changes constantly. This means that in addition to the words, we need to consider the pedestrian's position p. Furthermore, we assume that dialog context information c about what the pedestrian has previously referred to during his walk also plays a role. Therefore, what we really want to model are the probabilities $P(o|r,p,c)$. Technically, however, we will encode the position p and the dialog context c as part of the object properties (see Sect. 3.3). Thus, we want to estimate $P(o|r)$, where o is a geographical object as seen from position p, and referred to in a context c.

Figure 1 shows an example utterance u from the data we use. The pedestrian uses two REs to refer to three objects. RE_1 ("down the stairs") refers to two objects, RE_2 ("towards the arch at the bottom") refers to one object. When applying the classifiers to a new RE, each word determines whether the expression can refer to an object in the new candidate set. Usually, objects are described by noun phrases. However, we expect that more information than just the noun phrase will contribute to the correct resolution of an RE. For example, the classifier for the preposition *along* will learn to associate itself with objects of type `street` or `building`, but not with



"I continue in a southwesterly direction *down the steps* [RE_1] *towards the arch at the bottom* [RE_2]"

u_1 : 'I continue in a southwesterly direction down the steps towards the arch at the bottom'
RE_1 : "down the steps", $o_{t_1} = \{o_1, o_5\}$
RE_2 : "towards the arch at the bottom", $o_{t_2} = \{o_2\}$
candidate set $cs_1 : \{o_1 \ldots o_k\}$

Fig. 1 Example utterance containing 2 REs

type shop. Therefore, we define an RE rather loosely as any substring from the utterance that contains information about an object. Specifically, we included spatial prepositions like *along* and *through* and transitive motion verbs like *cross*. Relevant REs are annotated manually. In the particular situation in Fig. 1, there are n objects o the pedestrian could refer to. Every object o_i is represented as a vector of features, encoding information about what kind of object it is, how it is positioned with respect to the pedestrian, and whether it has been mentioned before.

The task is then, given each of the REs and the set of candidate objects, to find the target set of objects o that the words in r are most likely referring to. We approach this task following Kennington and Schlangen (2015), who addressed the problem of reference resolution in a small-scale puzzle piece scenario.

The objects are represented as vectors of numerical features that encode, for example, their type (see Sect. 3.3). We train individual word classifiers c that, when applied to the vector representation of a geographical entity $o_i = (x_1, \ldots, x_n)$, compute the probability that the word r_j refers to o_i. That is, $c_{r_j}(o_i) = P(o_i|r_j)$, where c_{r_j} is the classifier for the word r_j. Each c_{r_j} is a logistic regression classifier.

In general, for a referring expression r consisting of several words $r_1 \ldots r_m$, we compute the probability that r refers to each of the objects in the candidate set as a function of the probabilities for each word:

$$P(o_i|r = r_{1\ldots m}) = f(c_{r_1}(o_i) \ldots c_{r_m}(o_i)) \tag{1}$$

Following Kennington and Schlangenwe let f be the arithmetic average of all $c_{r_j}(o_i)$. Then, objects with a higher probability value are more likely to be the intended targets of the RE.

Using the data we describe in Sect. 3, we train these logistic regression classifiers that compute an object's suitability as a referent based on the object's features. For every RE, the target object o is a positive example for each word in the RE. As negative examples for each of the words, we randomly choose another object from the candidate set. If an RE has more than one target in a candidate set, one positive example (and one negative example) is added for each target. During training, the information about how many targets an RE refers to is not represented explicitly.

Intuitively, the classifier for the word 'building' will learn to associate high probabilities to objects that represent buildings (because they appeared as positive examples), and lower probabilities to other objects, such as streets. The classifier for the word 'the' on the other hand is likely to associate equal probabilities to buildings and streets, because speakers use it with both kinds of objects.

Table 1 shows a small example of how the word classifiers are used. In this scene, the candidate set contains 4 objects. The pedestrian utters the words "towards the arch at the bottom". The method then takes each word r of the utterance, computes the probability $P(o|r)$ for each object o in the candidate set (Step a), and then computes a final probability score for each object o by averaging over all word probabilities for that object (Step b), as given by Eq. 1. If no classifier is available for a word (because it has not appeared in the training data), the word is ignored. In Step c, the

Table 1 Example application of word classifiers. If no classifier is available, the word is ignored

(a) Word-to-object classification $P(o_i\|r_j)$				
Words $r_{1\ldots m}$	Candidate objects $o_{1\ldots n}$			
	o_1	o_2	o_3	o_4
$r_1 = towards$	0.11	0.99	0.89	0.26
$r_2 = the$	0.59	0.90	0.76	0.76
$r_3 = arch$	0.19	0.88	0.95	0.19
$r_4 = at$	0.29	0.89	0.87	0.90
$r_5 = the$	0.59	0.90	0.76	0.76
$r_6 = bottom \rightarrow$	–	–	–	–
(b) Composition	↓	↓	↓	↓
$\frac{\sum_{j=1}^{m} P(o_i\|r_j)}{m}$	0.35	0.91	0.85	0.57
(c) Selecting the target		↓		
argmax $P(o\|r)$		o_2		

method picks the object with the highest probability as answer, i.e. it assumes that this object is the target object. In the example, it returns object o_2.

In practice, not every RE refers to exactly one object. It is possible that an RE refers to two or more objects (as in the example in Fig. 1), or that it refers to no object in the candidate set. As mentioned in Sect. 1, we assume that an RE refers to an object in the pedestrian's environment. It is however possible that the target object is not part of the candidate set. The target object may not exist in the database, or it was not considered as a candidate in the given situation, e.g. because it was not considered visible given the pedestrian's position. Furthermore (as will be described in more detail in Sect. 3), number information in the RE itself, i.e. whether it is a plural or a singular RE, does not necessarily correspond to the size of the expected target object set.

We therefore extend the method presented above to incorporate these additional cases. In the following section, we explain the data and features we use. We then first look at the case of simple references in which one RE corresponds to one target object and show how our data achieves results comparable to Kennington and Schlangen (2015). We then suggest an extension of the method, capable of dealing with the cases when there are 0, 1, or more target objects for a given RE, and present results from experiments on our data.

3 Data

3.1 The SPACEREF *Data*

The data that we are studying and that is described in (Götze and Boye 2015, 2016b) contains transcriptions of pedestrians describing their environment while walking along a given path. Referring expressions are annotated with the identifier(s) of their target referent as represented in the map (see Sect. 3.2). Positional information in the form of GPS coordinates was automatically logged.

The corpus contains a total of $1,303$ referring expressions that are annotated with one or more target referents or tagged as having no referent object in the map. 559 (42.9%) REs have exactly one target object, 218 (16.7%) have more than one target (3 on average), and 526 (40.4%) of the REs have no target referent in their respective candidate set. The candidate set contains on average 33 objects.

3.2 The Geographic Representation

To represent the city environment in which the pedestrians are moving, we choose OpenStreetMap (Haklay and Weber 2008). OpenStreetMap represents objects such as buildings, streets, and shops in a way that is suitable for this task: all objects have information about their position associated with them in the form of GPS latitude/longitude coordinates and the map covers about 96% of the objects mentioned by the pedestrians. This makes it possible to automatically compute a candidate set on the basis of the pedestrian's position.

As described in (Götze and Boye 2015), the way that OpenStreetMap segments space into objects does not always correspond to how a pedestrian views his environment. In OpenStreetMap, streets are cut up into many small segments, each with their own specification of speed limits or access restrictions. Likewise, plurals do not necessarily have more than one target object, e.g. a block of buildings can be represented as one object, but perceived and referred to as "the buildings". We address how to make the choice of how many objects to return as referents in Sect. 5. We modify the selection step in a way that it returns all the relevant objects as correct target referents.

3.3 Features

In order to train the word classifiers, we need to represent the objects in each candidate set using suitable features that capture a part of the word's meaning. Most REs contain descriptions of the objects' type. Therefore, an object's features should contain a notion of their type, i.e. whether the object is a street, a building, or a bench. As mentioned earlier, we use OpenStreetMap to compute the candidate set on the basis of the pedestrian's position: all objects that are in view at the time of using each word are computed as described by (Boye et al. 2014). In addition to positional information, OpenStreetMap provides semantic tags for each object, specifying information like names, types, opening hours or other access information.

Table 2 The features that describe each candidate object. The first five CONTEXT features correspond to L1–5 in (Iida et al. 2011)

Feature	Values
OSM	
type	0/1 The object is of that type (427 features)
POS	
dist	2-log distance from the pedestrian's position to the object
angle	Angle between the walking direction and the object direction, measured from the pedestrian's position
CONTEXT	
mrRE	0/1 The object is referred to by the most recent RE
m10	0/1 The time distance to the last mention of this object is $\leq$10 s
m20	0/1 The time distance to the last mention of this object is $\leq$20 s and $>$10 s
m20+	0/1 The time distance to the last mention of this object is $>$20 s
never	0/1 The object has never been referred to
t50	0/1 The distance to the last mention of this object is $\leq$50 m
t100	0/1 The distance to the last mention of this object is $\leq$100 m and $>$50 m
t100+	0/1 The distance to the last mention of this object is $>$100 m

OpenStreetMap is a crowd-sourced database. It defines the usage of many tags and their values,[1] but contributors are in no way restricted in what tags they assign to an entity. As in (Götze and Boye 2016a), we derive 427 binary type features from the OpenStreetMap annotation. If an entity is of a certain type, it has value 1 for this feature, otherwise 0. We base the derivation of features only on such tags that are defined in OpenStreetMap's wiki. That means that other, user-defined tags that also carry non-relevant information, are not excluded from the feature set and introduce a fair amount of noise to our object representations. We will show in Sect. 5 that these features, that we obtain with only little processing of the original data, perform well when computing word meanings.

In addition to the type features (called OSM in Table 2), we derive positional information (POS) for each object: the distance and angle relate each object to the pedestrian's position. The feature set CONTEXT contains context information on whether an object has been mentioned before and how recent this mention was in terms of time or traveled distance. This context feature set is an extension of features used in (Iida et al. 2011) and is intended to capture and incorporate the meaning of function words, such as the determiners 'a' and 'the', 'that', 'this' etc. For example, referring expressions of the form 'a *x*' are likely to refer to a new object while mentioning 'this *x*' is an indication of a previously mentioned object. Table 2 shows the full list of features.

[1] http://wiki.openstreetmap.org/map_features.

Table 3 Evaluation for one-to-one references

	FHS	MRR
OSM	54.64	0.66
OSM + POS	58.09	0.70
OSM + POS + CONTEXT	59.17	0.72

4 Experiments

4.1 *One-to-One References*

In this setting, we select from the data only those instances with target set size 1, i.e. where each RE corresponds to exactly one object identifier in the map. This is the case in 559 instances. We are training the word classifiers on different combinations of features, in the way as described in Sect. 2. Testing is done using 10-fold cross-validation. Since negative examples are chosen randomly at training time, this process is repeated 10 times and we report averages in Table 3. We report the First Hit Success Rate (FHS), i.e. in how many cases the target object was correctly ranked highest (and thus selected) by the method, and the Mean Reciprocal Rank (MRR), indicating how high the correct object was ranked on average.[2] For comparison, in the puzzle piece setting in (Kennington and Schlangen 2015), the classifiers found 42% (FHS) of the targets and reached a MRR of 0.61. Table 3 shows that already when using only the information contained in the OpenStreetMap tags, the FHS Rate reaches 55% with a MRR of 0.66. Including positional and context information, we obtain a FHS in 59% of the cases and a MRR of 0.72.

4.2 *One-to-Many References*

The assumption that there is exactly one object that is the correct target referent does not hold for more than half of the referring expressions in our data. In 40% the correct target is not among the candidate referents (cf. Sect. 3), and in another 17% the target corresponds to a set of more than one object in the database.

Using the original method and choosing the most likely object will result in a wrong answer when there is no correct target, and an insufficient answer when there is more than one. Instead, when there are several targets, a reference resolution method should preferably return all these targets, and when there is no target, the method should return the empty set.

[2] The Reciprocal Rank measure calculates the reciprocal of the rank. It is 1 if the correct object is ranked highest, 0.5 if the correct object is ranked second, etc. The Mean Reciprocal Rank (MRR) is the average across many such calculations.

A possible solution is to define a threshold value t, where only the object or objects that have a probability of at least t will be considered as referents. If the highest ranked object is below the threshold, no object will be returned.

In the next step, we split our data into a training set of 80% (1, 025 instances) and a development and test set of 10% each (132 and 146 instances, respectively[3]). The training set is used for training the word classifiers as described previously, whereas the development set is used for determining a suitable threshold value. The test set is used for evaluation. Unless otherwise stated, all training and testing is carried out in 10 iterations, and we report averages (negative examples for the word classifiers are chosen at random and differ for each training run).

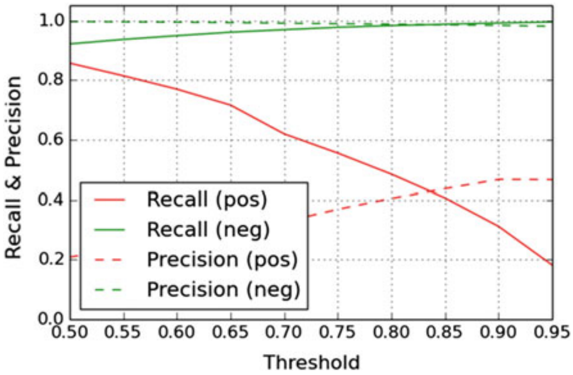
For evaluation, we now look at how many objects were (in)correctly classified for each RE. Recall that for each RE, there is an average of 33 candidate objects. Each object is assigned a probability that it is the correct referent of the referring expression in question. In finding the threshold value, we use all three feature sets, and vary the threshold over a range of [.5;.95] in steps of 0.05. All objects that obtain a probability of at least the threshold value will be returned by the method.

We computed accuracy, precision, recall, and F-measure for each threshold value. Every target referent is a positive, all other objects are negatives. For an RE without a target referent in the candidate set, all objects should be classified as negative. For each RE, there are many more objects that should be classified as negative, i.e. as not being the correct referent. When classifying all objects as negative, i.e. never returning a referent, we would obtain an accuracy of 0.97. With the original method of choosing the object(s) with highest probability, the accuracy on the development set is also 0.97. Starting at a threshold value of 0.8 the accuracy improves over the original setting.

Looking at the F-measure, a threshold of around 0.80 is best in terms of both positives and negatives ($F = 0.46$). For the positive class (objects chosen as referents), this threshold means a recall of 0.52 and a precision of 0.41. For the negative class (objects rejected as referents), both precision and recall are close to 1.0 (cf. Fig. 2). In a particular application, it may be desirable to prefer higher precision over higher recall (being sure that what was found is a correct referent), or vice versa (finding as many targets as possible at the expense of including false positives). Here, we are not making such a choice and set the threshold value at 0.80. At this threshold, the method also works well in terms of how many targets it finds for the different conditions: It predicts on average 1 object for the case where there is only 1 or no target referent and slightly over 2 in the case where there are more.

[3] The data is split on the utterance level, where each utterance contains one or more referring expressions.

Fig. 2 Evaluation for varying thresholds



4.3 *Testing on the Held-Out Test Set*

Table 4 shows the results for applying the learned models on the remaining 10% of the data (146 instances) with a threshold of 0.80 for selecting referent objects. The results on this test set are similar to the results on the development set.

4.3.1 Evaluation per Referring Expression

The evaluation measures in Table 4 show what happens within each candidate set. Table 5 shows how many of the referring expressions the method resolves correctly. In the strictest setting (in which the method returns all targets and no false positives), the method resolves 44.3% of the referring expressions correctly.

When there is no target referent, it answers correctly with the empty set in more than half of the cases. When there is one referent, it answers correctly in one third of the cases, when there is more than one referent, in one fourth of the cases. Allowing

Table 4 Evaluation results for the held-out test data when selecting objects that have a probability of at least 0.8

Measure	Test set		Dev set	
	Pos	Neg	Pos	Neg
Accuracy	0.97		0.97	
Precision	0.40	0.98	0.40	0.99
Recall	0.45	0.98	0.48	0.98
F-measure	0.42	0.98	0.44	0.98

Table 5 Evaluation per RE on the test set (threshold = 0.8). TP: True Pos., FP: False Pos

Target set size s			MRR
$s = 0$	Correct (TP $= 0$, FP $= 0$)	59.0%	0.59
$s = 1$	Correct (TP $= 1$, FP $= 0$)	37.0%	0.71
	Partly correct (TP $= 1$)	48.4%	
$s > 1$	Correct (TP $= s$, FP $= 0$)	25.9%	0.73
	Partly correct (TP $= s$)	26.8%	
	Partly correct (TP ≥ 2)	51.4%	
	Partly correct (TP ≥ 1)	67.3%	
Total	TP $= s$, FP $= 0$	44.3%	0.66
	TP $= s$	49.4%	

also false positives in the answer set, it answers correctly about half of the time. For all cases, the target set of objects obtains a rank of 1.5 (i.e. a MRR of 0.66) on average. When there are several targets, all of them are ranked high, with an average MRR of 0.73, i.e. about rank 1.4.

4.4 Results

The results in Sect. 4.1 show that the basic approach of training word classifiers and applying them to features derived from OpenStreetMap representations of objects works well. Choosing the most likely object resolves simple one-to-one references in almost 60% of the REs. In assessing the success of the method recall that this reference resolution problem is a difficult one—the candidate set contains 33 candidate referents on average.

For the general case—where there might be 0, 1, or more correct referents—the extended method using thresholds resolves 44.3% of the referring expressions correctly, meaning that it selected exactly the right set of referents, so this is an even more difficult problem than the one-to-one case. Not surprisingly, the result is not as good as the one-to-one case, but still higher than the 42% for the one-to-one references in Kennington and Schlangen's puzzle piece setting.

When there is one target referent, the extended method produces a completely correct answer in 33.3% of the REs, and a partly correct answer in 49.2%. The basic method of choosing the most likely object was correct in 59%. However, we can now also resolve the other cases without explicitly representing information about the target set size.

5 Discussion

Given the sparseness of the language data and the crowd-sourced nature of the geographical data, we consider the results a good step towards incorporating spatial reference resolution into a real-time system.

Since OpenStreetMap tags most often are plain English words, an obvious alternative idea is to simply look for those words in the input (i.e. if the user mentions a "building", this would translate to OpenStreetMap entities having the tag `building`). This is in fact what Götze and Boye (2015) have tried before. However, that straightforward approach has drawbacks: It is language-dependent, it requires manual intervention and translation-rule writing (since some words like "street" have OpenStreetMap tag counterparts that no user would ever say: `primary`, `secondary` etc.), and it presupposes that every reference refers to exactly one entity. The probabilistic approach presented in this paper has none of these drawbacks.

Recall that the geographical representation is imperfect in two ways. First, we cannot be sure that all information in OpenStreetMap is complete and correct. Second, the GPS signal of the pedestrian's position is only an approximation of his real position. This situation is however realistic in this domain (Modsching et al. 2006) and we have therefore not manipulated neither the map representation nor the GPS signal. The features that represent positional information are noisy, and a closer look at the classifiers of the words *left* and *right* reveals that they have not learned any association with these features, their associated weights are close to 0, i.e. they do not influence the object rank. We expect that a more accurate GPS signal will improve the results (cf. Misu et al., 2014).

On the other hand, semantic information about an object's type or appearance correlates well with type features that we would expect. For example, the classifier for the noun *building* associates the highest weight with the feature `building_yes` and one of the lowest weights to the feature `highway`.

Table 6 shows the highest and lowest weighted features for the nouns *road* and *building*.

Table 6 Semantic (OSM) features correlate well with types: extract of the word classifiers for *road* and *building*

c_{road}		$c_{building}$	
Feature	Weight	Feature	Weight
highway	1.8894	building = yes	1.3778
name	1.6214	website	1.1827
secondary	0.8873	t50	0.5660
bus_stop	−0.5856	waste_basket	−0.2552
distal	−0.7588	highway	−0.9854
never	−0.8692	distal	−1.6728

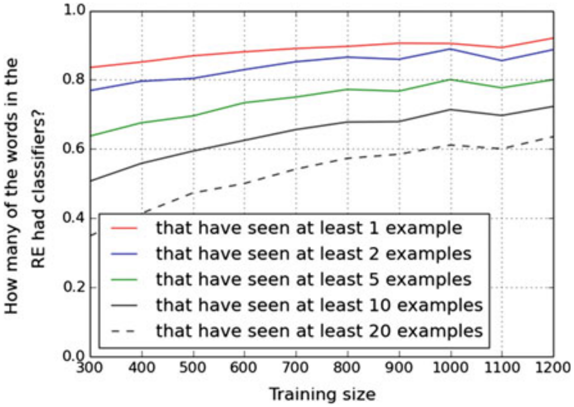
Table 7 Context features in function words: extract of the word classifiers for *this* and *same*. (cf. Table 2)

c_{this}		c_{same}	
Feature	Weight	Feature	Weight
highway	1.0784	t100	1.1097
mrRE	0.7123	m10	0.9109
secondary	0.6251	secondary	0.6382
track	−0.2252	never	−0.1744
angle	−0.3001	distal	−0.5068
distal	−1.5258	mrRE	−0.7048

Context information about whether an object has been mentioned before, within a certain time span, or a certain distance traveled, also shows the expected correlations. For the modifier *same*, the corresponding classifier learns high weights for features indicating that the object has been recently mentioned and the lowest weights to features that indicate that an object has not recently been mentioned as shown for the words *same* and *this* in Table 7.

The size of the vocabulary that the pedestrians use to describe their environment is relatively small. From our training set, we obtain about 280 word classifiers, most of them have seen only few examples. In the complete data set of (313 distinct tokens), the average number of examples per classifier is 63.3, the median is 31.5. Only 73 words occur at least 10 times, 45 occur at least 20 times. With the training set that we have used, at least 90% of the words in an RE had classifiers and 60% have classifiers that were trained on at least 20 examples. Figure 3 shows the classifier coverage for different training set sizes.

Fig. 3 Words per RE that have classifiers



6 Related Work

How the meaning of words can be grounded in perceptive information has recently become an active area of research (Roy 2005; Mooney 2008). When dealing with the problem of Reference Resolution, mainly visual information is considered to model the meaning of words and phrases (Kruijff et al. 2006; Matuszek et al. 2014; Kennington and Schlangen 2015).

In our domain of pedestrian wayfinding, direct visual input is hard to obtain. Some studies rely on photographs (Baltaretu et al. 2015), but in a working real-time wayfinding system, photographs will be insufficient as pedestrians are not restricted in their movement and can quickly turn around to face another direction.

Works that do not rely in visual information, but on a direct representation of the physical environment typically work on a domain that is considerably smaller than ours, with 10 candidate objects or fewer and where objects are of only few distinctive types (Gorniak and Roy 2005; Schütte et al. 2010; Funakoshi et al. 2012). In (Kennington and Schlangen 2015), the set of objects is comparable in size to ours. They process the set of puzzle pieces using computer vision methods. However, all objects are clearly distinct from each other and all are of the same type. Instead of direct perceptual input, we rely on a crowd-sourced map representation of the environment that covers the study area well and use features from the semantic tags associated to the objects and that are also crowd-sourced. The Pursuit corpus (Blaylock 2011), in which car drivers are also describing while they are moving along a path, is in principle suitable for this task as well, but the area is not very well covered in OpenStreetMap at this point and the object annotation contains references to several different databases, and no information on other candidate objects.

Misu et al. (2014) have attempted to resolve spatial references with good success in a similar setting. Instead of descriptions, car drivers pose queries about Points of Interest (POI). Like in the data we use, they use information about the speaker position and the POI positions and types. Additionally, they have access to the drivers' head pose when speaking and an analysis of the data showed that directional information (left/right) aligned well with the speakers' mentions of directions, even though they also report errors in the GPS information. In this work however, knowledge about the referent candidates is assembled manually. They also explicitly exclude context information such as dialog history.

How to segment the context into objects is an active area of research. Context segmentation is typically done independently for each modality and the information then fused. Kruijff et al. (2006) have proposed a framework to incorporate this step into a rule-based reference resolution algorithm, and Bruni et al. (2014) fuse information from the linguistic and visual context to obtain an integrated representation of meaning. For context representations based on visual input, computer vision algorithms are applied with good results for small domains and where the objects are clearly distinct from each other (Matuszek et al. 2012; Kennington and Schlangen 2015). Krishnamurthy and Kollar (2013) and Malinowski and Fritz (2014) perform this segmentation on photographs that depict rather everyday scenes and Malinowski and Fritz (2014) account for uncertainty in the image segmentation by utilizing the associated confidence scores. All of these approaches do however assume that every object (or segment) corresponds to a referent (unless the RE is a plural).

In OpenStreetMap, objects do have clear boundaries, but as we have described in Sect. 3.2, this segmentation does not align with the objects that the pedestrians in the data refer to. We handle this discrepancy by resolving REs to sets of objects based on the probability distribution returned by the word classifiers. An alternative approach is to structure the context representation beforehand, i.e. decide which sets of entities are available for reference and modify the candidate set accordingly. Funakoshi et al. (2012) use Reference Domain Theory (Salmon-Alt and Romary 2009), grouping tangram pieces based on proximity to determine which reference domains, i.e. sets of objects, can be referred to. There is good evidence for how humans perceptually group objects (Thórisson 1994), e.g. based on proximity. However, in our domain and with the geographic representation at hand, it remains unclear how to represent a set of objects based on the features of the individual objects. This is a known issue in research using OpenStreetMap (Ballatore et al. 2013) and we leave this as a question for future research.

Finally, in an interactive wayfinding system, references to landmarks are also an essential part of the generation process. There are at least two steps involved in this process. The first one decides which landmark to choose, usually on the basis of the current routing situation and some calculation of which objects are most salient (e.g. Raubal and Winter, 2002; Götze and Boye, 2016a). The second step decides how to translate the object representation into a suitable referring expression (e.g. Garoufi and Koller, 2011; Paraboni and van Deemter, 2014).

7 Conclusion

We have presented a method for situated reference resolution in a large-scale environment where the context changes with the speaker's movement. Using an existing, crowd-sourced geographic database that represents objects at different granularities than the speakers refer to them. We have shown a way to extend current methods to allow for cases where the correct set of target objects is empty or contains more than one object.

References

Ballatore A, Bertolotto M, Wilson DC (2013) Geographic knowledge extraction and semantic similarity in openstreetmap. Knowl Info Syst 37(1):61–81

Baltaretu A, Krahmer E, Maes A (2015) Improving route directions: the role of intersection type and visual clutter for spatial reference. Appl Cognitive Psychol 29(5):647–660

Blaylock N (2011) Semantic annotation of street-level geospatial entities. Proceedings of the IEEE ICSC workshop on semantic annotation for computational linguistic resources

Boye J, Fredriksson M, Götze J, Gustafson J, Königsmann J (2014) Walk this way: spatial grounding for city exploration. Natural interaction with robots, knowbots and smartphones, pp 59–67

Bruni E, Tran N-K, Baroni M (2014) Multimodal distributional semantics. J Artif Intell Res (JAIR) 49:1–47

Denis M (1997) The description of routes: a cognitive approach to the production of spatial discourse. Curr Psychol Cogn 16(4):409–458

Funakoshi K, Nakano M, Tokunaga T, Iida R (2012) A unified probabilistic approach to referring expressions. Proceedings of the 13th annual meeting of the special interest group on discourse and dialogue, pp 237–246

Garoufi K, Koller A (2011) The potsdam NLG systems at the GIVE-2.5 challenge. Proceedings of the 13th European workshop on natural language generation (ENLG), pp 307–311

Gorniak P, Roy D (2005) Probabilistic grounding of situated speech using plan recognition and reference resolution. Proceedings of the 7th international conference on multimodal interfaces, pp 138–143

Götze J, Boye J (2015) Resolving spatial references using crowdsourced geographical data. Proceedings of the 20th Nordic conference of computational linguistics, NODALIDA, pp 61–68

Götze J, Boye J (2016a) Learning landmark salience models from users' route instructions. J Locat Based Serv 10(1):47–63

Götze J, Boye J (2016b) SpaceRef: a corpus of street-level geographic descriptions. Proceedings of LREC

Haklay M, Weber P (2008) Openstreetmap: user-generated street maps. Pervasive Comput, IEEE 7(4):12–18

Iida R, Yasuhara M, Tokunaga T (2011) Multi-modal reference resolution in situated dialogue by integrating linguistic and extra-linguistic clues. In: The 5th international joint conference on natural language processing, pp. 84–92

Kennington C, Schlangen D (2015) Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing, pp. 292–301

Krishnamurthy J, Kollar T (2013) Jointly learning to parse and perceive: connecting natural language to the physical world. TACL 1:193–206

Kruijff G-J, Kelleher J, Hawes N (2006) Information fusion for visual reference resolution in dynamic situated dialogue. Percept Interact Technol 4021:117–128

MacMahon M, Stankiewicz B, Kuipers B (2006) Walk the talk: connecting language, knowledge, and action in route instructions. Proceedings of the 21st national conference on artificial intelligence, pp 1475–1482

Malinowski M, Fritz M (2014) A multi-world approach to question answering about real-world scenes based on uncertain input. NIPS, pp 1682–1690

Matuszek C, Bo L, Zettlemoyer L, Fox D (2014) Learning from unscripted deictic gesture and language for human-robot interactions. Proceedings of AAAI, pp 2556–2563

Matuszek C, FitzGerald N, Zettlemoyer LS, Bo L, Fox D (2012) A joint model of language and perception for grounded attribute learning. ICML

Misu T, Raux A, Gupta R, Lane I (2014) Situated language understanding at 25 miles per hour. Proceedings of the 15th SIGdial workshop on discourse and dialogue

Modsching M, Kramer R, ten Hagen K (2006) Field trial on gps accuracy in a medium size city: the influence of built-up. In: 3rd workshop on positioning, navigation and communication, pp 209–218

Mooney RJ (2008) Learning to connect language and perception. Proceedings of AAAI, pp 1598–1601

Paraboni I, van Deemter K (2014) Reference and the facilitation of search in spatial domains. Lang, Cogn Neurosci 29(8):1002–1017

Raubal M, Winter S (2002) Enriching wayfinding instructions with local landmarks. In: Geographic information science. Lecture notes in computer science, vol 2478, pp 243–259

Roy D (2005) Grounding words in perception and action: computational insights. Trends in cognitive sciences 9(8):389–396

Salmon-Alt S, Romary L (2009) Reference resolution within the framework of cognitive grammar. Int Colloquium Cognitive Sci, pp 284–299

Schütte N, Kelleher J, Mac Namee B (2010) Visual salience and reference resolution in situated dialogues: a corpus-based evaluation. AAAI symposium on dialog with robots, pp 109–114

Thórisson, K. R. (1994), Simulated perceptual grouping: an application to human computer interaction. Proceedings of the sixteenth annual conference of the cognitive science society CSS94, pp 876–881

3D Building Maps for Everyone —Mapping Buildings Using VGI

Laura Knoth, Manfred Mittlboeck and Bernhard Vockner

Abstract With regard to upcoming or emerging location-based technologies such as indoor positioning and augmented reality, the demand for 3D indoor building models increases. Yet, buildings are often only represented as points or extruded polygons for visualization purposes that are isolated from their surrounding environment. Currently, the whole building modeling workflow is done by modeling experts ("expert2expert") or specific companies. In order to be independent of specific companies for indoor mapping, we propose VGI as a way to capture building information. Local experts or the "experts of the building", thus people working or living in a certain building might contribute to this process, although they are not experts in building modeling. Using the tools provided by OSM gives the amateurs the tools at their hand to model what they need and in a simple manner to give them the possibility to provide their "expert knowledge" ("amateur2expert").

Keywords Volunteered geographic information · Smart building · Sensor data · Entrepreneurs

1 Introduction

With services such as Google maps, OpenStreetMap (OSM) and local services like the Austrian "Basemap.at", we live in a world which is measured and mapped to a higher extent and in more detail than ever.

> The year is 2016 AC. The world is entirely mapped by humankind. Well, not entirely… Buildings still hold out against the mapping invaders (modified after the first sentences of every "Asterix"-comic)

But is it really? For the outdoors, this might be the case, but the buildings, which make up 2.55 million km^2, are only represented as boxes (Knoth 2015).

L. Knoth (✉) · M. Mittlboeck · B. Vockner
Research Studios Austria iSPACE, Salzburg, Austria
e-mail: laura.knoth@researchstudio.at

A. Bregt et al. (eds.), *Societal Geo-innovation*, Lecture Notes
in Geoinformation and Cartography, DOI 10.1007/978-3-319-56759-4_5

The reason why the outdoor environment is better mapped than buildings is that mapping agencies as well as the communities of "Volunteered Geographic Information" (VGI) have mainly been focusing on the collection of objects that can be observed by methods of remote sensing (e.g. digitizing of ortho- and satellite imagery; acquisition of LIDAR-data). Up to now, indoor representations of buildings (floors, rooms, etc.) were out of scope within cartographic maps.

Besides the work of mapping agencies, groups of volunteers collect data and share it with the community. Thus, they contribute to mapping the entire (outside) world in detail. This is called "VGI approach". In contrast to the outdoors, the indoor world is not being mapped in such a widespread manner, especially not in VGI communities (see Sect. 2.3).

While with the increase of Wi-Fi and Bluetooth Beacons for indoor positioning, the demand for 3D indoor models has been rising (Schmitz et al. 2015), buildings still are often only represented very poorly as points or extruded polygons in 3D-maps for visualization purposes. When being represented in real 3D including the indoors, buildings regularly remain in encapsulated visualizations such as images or interactive 3D-representations that do not go beyond visualization. Geographic 3D map representations of buildings with real-world coordinates and the surrounding environment rarely exist.

To overcome this situation, this paper provides a method based on VGI to make it easier to capture the indoor information and integrate it into a well-organized structure in 2D and 3D. The paper starts with a short introduction on current ways to capture building information (Sect. 2) in the common practice, further ways to capture it and current approaches to using VGI for building modeling. Section 3 describes the use case scenario that was used to test our approach. Section 4 describes the implementation and the used tools and programs. In Sect. 5, we describe the results, while Sects. 6 and 7 discuss the results and provide an outlook.

2 State of the Art

2.1 The Process of Building Modeling ("expert2expert")

Today, building information (outdoors and/or indoors) is mainly created at two stages during the building life cycle: Either before its construction (e.g. CAD plan), or after construction (e.g. digitized from the CAD plan or captured by laser scanners) and is limited to professionals.

The phase before the construction of a building might also be referred to as the "design phase" (van Oosterom et al. 2006), where experts of the AEC-domain (Architecture, Engineering and Construction) use their methods and tools to design a building and all of its functionalities (electricity, plumbing, HVAC, etc.).

After the construction, when the building is taken into use, a disruption of the information flow usually happens: In many projects, we encountered the case that as soon as the work of the AEC-users is done, the whole detailed building information gets "lost" or just stays in its encapsulated data silo and often without further use in 2D or 3D (online) maps due to missing transformation possibilities (van Oosterom et al. 2006).

Nevertheless, users of the GIS-domain would need this information by the time the building becomes part of the real-world environment. The information from the former modeling could be used in order to realize, for example, 3D maps for indoor positioning. As of today, GI professionals mainly start their duty after the physical construction of the building in the "as-built phase", where they use tools and models to capture the "as-built"-building using techniques e.g. from remote sensing to capture and/or reconstruct the building for a use in a GIS (van Oosterom et al. 2006).

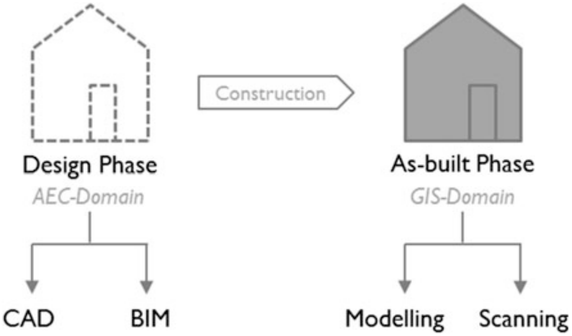
In doing so, the information gets disrupted between the two phases: In the "design phase", the AEC-users model the building with CAD and BIM, while after construction in the "as-built phase", users from the GIS-domain model the building *again*, mostly using their own modeling approaches or via scanning techniques (see Fig. 1).

Since the modeling in the "design" and the "as-built phase" are performed by different users, but capture the same object, one could assume that the models are the same. In fact, during construction, sometimes changes occur that are not represented in the plan created in the "design phase". For example, a wall may be installed in different places than initially planned. This change should at best get updated in the plan, but this is not always the case. Thus, the result of the "design phase" usually is a model that shows what should have been built instead of showing the real "as-built" building.

To compare the models of the "design" and the "as-built phase" and detect disparities, Musliman et al. (2010) developed a model to perform an as-built survey since "normally, there will be some differences between architectural drawing and actual constructed building such as overlaps or unintended protrusions, because of changes during field construction".

To overcome this common practice of capturing the same information twice by the two domains and to open ways to transfer information between them,

Fig. 1 Building models generation



several approaches aim at a direct harmonization and integration of the AEC-models into the GIS-domain and vice versa. One example is the work of Tashakkori et al. (2015), who developed the IESM (Indoor Emergency Spatial Model) for emergency situations based on IFC (standard for BIM). Another example can be found in Isikdag et al. (2013), where the BO-IDM (BIM oriented indoor modeling methodology) was developed for indoor navigation and orientation tasks. It is based on a BIM model. El-Mekawy et al. (2012) used an IFC-model that was translated into the UBM (Unified Building Model) and then into CityGML.

Using the expert2expert approach (CAD/BIM to GIS), thus moving from a model that contains the "design phase" information of buildings to a model that represents the "as-built"-information, further challenges arise:

1. Since CAD plans are derived from technical drawings, they are based on lines rather than on objects. The lines are "not smart", i.e. they know their geometric and visual parameters in fact, but they don't know what they represent. Additionally, in CAD, attributes in plain text are often placed near the lines that represent a geometry. Within the GIS-domain, there is a concrete distinction between geometry and signatures. Thus, any annotated objects in CAD are a challenge to transform into a GIS. To make the plans and attributes useful, harmonization and integration are of utmost importance.
2. As mentioned, changes in the design plan during the construction phase might not be updated in the drawing (CAD or BIM). Thus the design plan might differ from the real as-built building.
3. Even the best updated BIM model, which is object-based and very detailed, cannot be directly used in a GIS. BIM models often contain elements as detailed as single screws and bolts. This is far too complex for a use in a GIS. Thus, some generalization is necessary (Geiger et al. 2015).

2.2 Further Possibilities of Professional Building Modeling

If it is not transformed from CAD plans ("design phase"), building information is mainly captured by professional companies that use laser scanners to capture building information in the "as-built phase". A laser scanner can capture the surface of the walls and the interior of the building very well, but cannot "see" inside the walls or capture the attributes. Additionally, a laser scan does not generate objects. Therefore, this has to be done either using algorithms for object detection or manually by experts. Some tools, such as the Flexijet 4REVIT-System (Flexijet GmbH 2016), measure objects by telling the software with user inputs what is going to be measured in advance, i.e. a wall, a window, etc. Using this tool will result in a BIM as-built plan, but it is rather expensive and needs a lot of knowledge.

Hence, some companies started to capture indoor maps on their own, but with no common model. Google for example developed a workflow that involves sending an image of the indoor space model to the company (Google 2016). Google itself then converts the image to fit it into their data model. This workflow makes the users dependent on Google as they cannot decide which features are of interest, what they want to map and how they want it to be represented. Additionally, sending the data to Google might infringe the copyright of the data model or the privacy.

ESRI, on the other hand, also uses its own building data model, called ESRI BISDM (Building Interior Space Data Model) 3.0. A newer version, ESRI FISDM (Facilities Information Spatial Data Model) is in development as cooperation of ESRI, PenBay Solutions, Vertex3 and the University of Washington (FISDM 2014). BISDM models can be published online using the Campus Place Locator app. This enables users to view the map with different levels as a web-based 2D map and to find rooms and people in a campus via search fields.

2.3 VGI and Indoor Mapping (“amateur2expert”)

What all the models presented above have in common is that they are complex, often expensive and focused on a specific task. Furthermore, their use is often limited to 2D or if usable in 3D, then most often via Desktop-GIS. This is applicable for expert users, but restricts the access from further use by the public.

In order not to be dependent on specific companies for mapping, we propose VGI as a way to capture building information. This approach (“amateur2expert”) enables the local building experts (i.e. people who live or work in a certain building) to provide their “expert knowledge”, even though they might not be familiar with professional building modeling tools and approaches.

This follows the statement of Michael F. Goodchild as stated in Helft (2009), where he said that mapping should be done by the local people who know an area well, rather than limit it to mapping companies and agencies. This statement holds especially true for buildings, which cannot be captured by remote technologies.

The most popular VGI community is OpenStreetMap (OSM) (Goetz and Zipf 2012). The principle of OSM is similar to Wikipedia: everyone can contribute to its content. OSM is “totally open and no observation on the contributions is applied” (Arsanjani et al. 2013). However, this statement has to be handled with care, because there is no *automated* observation on quality, but users are well aware of changes and most of the time, changes are crosschecked by another user of the community. On the other hand, as stated by Rosser et al. (2012), “it is acknowledged that in some circumstances users may provide incomplete or inaccurate information and this may be accidental or even deliberate”.

However, it can be stated that even if VGI is done by non-experts, it provides the means to capture high-quality data. At the moment, VGI communities mainly focus on modeling the outdoors. Nevertheless, there are several approaches to create indoor data in OSM. The current ones are called “Simple Indoor Tagging”

(OSM Wiki 2016c) and “Simple 3D buildings” (OSM Wiki 2016b). Goetz and Zipf (2012) proposed another indoor ontology for OSM, called “IndoorOSM”, that is to be used to capture and tag 3D building information. This ontology was set “inactive” on March 5, 2015 by OSM users (OSM Wiki 2015) and recommended to not be used for new projects due to “technical problems (tag collisions, massive use of relations and direct use of osm element IDs)” (OSM Wiki 2015).

Since the publication of Goetz and Zipf in 2012, indoor tags in general have not been widely used in OSM. In 2012, the total use count of the “indoor”-tag was 642 (Goetz and Zipf 2012). Now, 4 years later, a total of 102.006 features (Taginfo 2016a) use the indoor tag worldwide, which is not even 1% of all OSM objects. From these values, 71% only use “indoor = yes” to tell the user that this feature is indoors (Taginfo 2016b). The tag “indoor = room” is used around 17.000 times worldwide. Other tags regarding indoor environments are “wall”, “no”, “corridor”, “area” and “door”. “Door” has been used 994 times. These numbers indicate that the indoor tags in OSM are not embraced by the OSM community. Besides the modeling issues, only some OSM viewers, such as “F4-Map” (F4 Map 2016) are even able to render 3D buildings and other elements such as trees, but there is no sophisticated renderer for 2D as well as 3D indoor information, yet. This additionally contributes to the modeling gap of building information that we are currently facing.

Nonetheless, the importance of building models for 3D online maps has already been noted by several authors. For instance, Rosser et al. (2012) realized the demand for building models in many different context-aware applications. Currently, this demand is not satisfied due to the fact that in contrast to the outdoor environment, nobody is directly responsible for the data acquisition of the indoor environment. Thus, Rosser et al. (2012) proposed a framework for capturing indoor information in a crowd-sourced manner using the camera of a smartphone. This approach allows the capture of buildings by its users and to keep it updated. However, using the camera to capture the inside of a building might cause problems in terms of privacy, both for people and as well for “valuable objects” that might be “identifiable from the imagery” (Rosser et al. 2012).

To avoid this special problem of privacy, our approach models the buildings without the need of a camera to shoot images. For modeling, we use the tools and the crowd-sourced approach of OSM to enable the creation of building models for everyone. To present the buildings after modeling, it provides a possibility to visualize the building, its attributes and context-information in a simple and appealing manner without the need of a Desktop-GIS.

3 Use Case: Energy Modeling in Schools Using JOSM

The use case of this paper was done as part of the transnational research project “THE4BEES” (Transnational Holistic Ecosystem 4 Better Energy Efficiency through Social innovation). The goal of THE4BEES is to change the behavior of users in public buildings in terms of energy use with different target groups, for example, pupils. With its specific requirements on energy modeling, THE4BEES provided the ideal prerequisites to test and validate two topics:

1. The applicability and adaptability of the building model that was proposed in Knoth (2015) and Knoth (2016) to the energy-topic of THE4BEES
2. The feasibility of the proposed "amateur2expert"-approach

The building model was developed for general use independent of a specific use case, thus it can be adapted and used for different purposes. This model provides the base for the model to be implemented. Once the model has been implemented by the users (amateurs), an automated workflow is utilized to transfer it to be used by the experts for further revision. The goal is to get a model in 2D as well as in 3D with its corresponding attributes that can be used beyond visualization. In the very end, the model should be available both for the experts as well as the amateurs so that both benefit from the amateur2expert-cooperation.

In THE4BEES, the amateur-users are pupils of different ages with the goal to model energy-relevant elements in their schools. Some of these pupils are technophile, while others are not used to mapping or to work on other technical drawings. Moreover, the building model should be integrated with real-time data from sensor stations to communicate the different values of energy consumption to the pupils.

Based on a thorough and extinctive analysis, the following elements have been identified for a sophisticated building model:

- Structural Elements: walls, stairs
- Opening Elements: doors, windows
- Spaces: rooms with their corresponding attributes such as the room number, use, etc.
- Energy-relevant elements: heaters/radiators
- Sensors: diverse sensors to capture room characteristics, such as temperature, humidity, CO_2-level, etc.
- Basic Furniture: tables, chalkboard

At this point, it is necessary to emphasize that the goal of this work was *not* to develop "yet another building model" for OSM. It is about giving the modeling tools in the hands of local "experts"/amateurs to model their environment using ready-to-use tools. Thus, the model and the workflow need to be simple enough to be used by these amateurs to create a sophisticated building model which then can be transformed by experts for further use ("amateur2expert").

4 amateur2expert Building Modeling

Within the context of THE4BEES, two key requirements need to be fulfilled in order for the target group of pupils to use the tools. The first one is that the tool needs to be easy to use in order that not only technophile pupils are able to use it. The second requirement is that it needs to be free of charge so that they can download and use it. Additionally, the results need to be transferrable from the acquisition tool into a GIS environment in 2D and 3D.

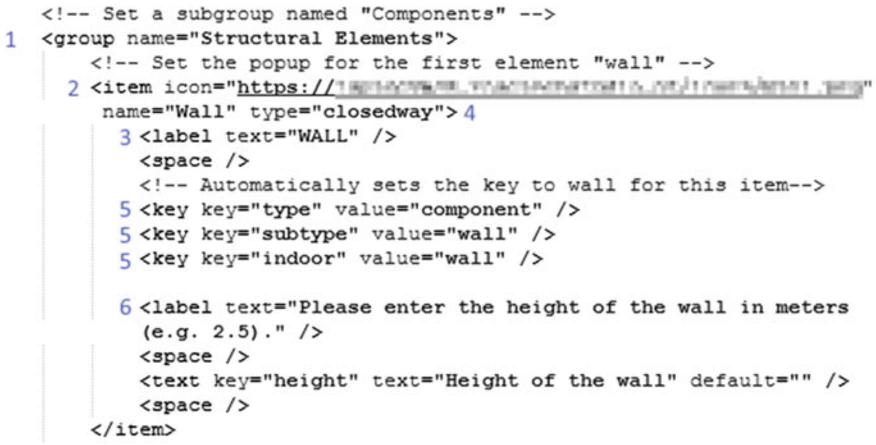
Due to their nature, the VGI-tools developed for OSM already are developed for amateurs. Thus, it suggested itself to look amongst them for an appropriate tool. In general, to edit OSM data, there are multiple possibilities: Online in the browser (e.g. with iD or with Potlatch 2), as Desktop-Version (JOSM or Merkaartor), directly in a GIS (QGIS, ArcGIS) or mobile with a smartphone (Vespucci, OsmAnd) (OSM Wiki 2016a).

In 2016, 70.8% of all edits in OSM have been performed using the Java OSM (JOSM) editor (N.A. 2017). JOSM has been developed since 2005 as java-based open-source editor for OSM. JOSM is free of charge, easy to use and can be customized using so called "presets" to adapt it to the use case and needs. Thus, it was the tool of choice to model the buildings.

The first step in using JOSM was to use the functionality of presets to implement the building model for a use by amateurs. At this point, it must be stated that the workflow is not a full "amateur2expert"-workflow: It does not directly start with the amateurs modeling the building. Letting the amateurs directly decide on what to model might produce chaos as everyone would model different elements and in a different way. Defining the elements in advance ensures that the required elements are modeled in a certain way and are not forgotten. On the other hand, the definition of elements should not "be set in stone" by the experts, dictating what has to be modeled and how it has to be modeled. There should always be the possibility of a dialogue between the experts and the amateurs about the elements, their attributes and their relationships.

Using JOSM, the elements to be modeled by the amateurs can be defined in a customized preset. A preset in JOSM is an XML-based structure that defines elements and their attributes for a use in OSM. In OSM, attributes are called "tags". They consist of two values, separated by an equals-sign. The first value is the key, while the other is the value. An example is "indoor = yes" with "indoor" as the key and "yes" as the value. In an attribute table, the key would be the column title, the values would be written in the columns.

In a preset, the definition is not restricted to the name of the element (e.g. wall) and one value (e.g. height). It is possible to assign more than one tag to an element. The wall-element could have a height, a color, a material and one could ask if the wall is insulated. Additionally, a preset defines a visual menu for the elements for the entry in the user interface. This makes the use by both, amateurs and experts, very simple. Figure 2 shows an example of the XML-structure for the element "wall". The parent group of "wall" is called "Structural Elements" (1). Below,



```
   <!-- Set a subgroup named "Components" -->
1  <group name="Structural Elements">
       <!-- Set the popup for the first element "wall" -->
     2 <item icon="https://[redacted]"
        name="Wall" type="closedway"> 4
         3 <label text="WALL" />
           <space />
           <!-- Automatically sets the key to wall for this item-->
         5 <key key="type" value="component" />
         5 <key key="subtype" value="wall" />
         5 <key key="indoor" value="wall" />

         6 <label text="Please enter the height of the wall in meters
           (e.g. 2.5)." />
           <space />
           <text key="height" text="Height of the wall" default="" />
           <space />
       </item>
```

Fig. 2 Exemplary XML-code for a wall-element

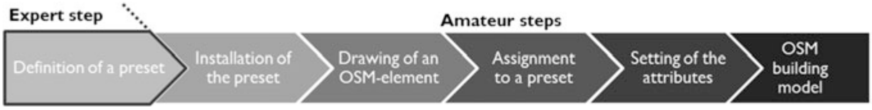


Fig. 3 Workflow with JOSM-presets

the name “wall” for the item and the corresponding icon was set for easy discovery (2, 3). The type shows that only “closedway” is allowed (4), thus, a wall can only be represented by a polygon. The next few lines automatically set three values, the user does not have to interact (5). The next lines define an input text-field for the height of the wall (6).

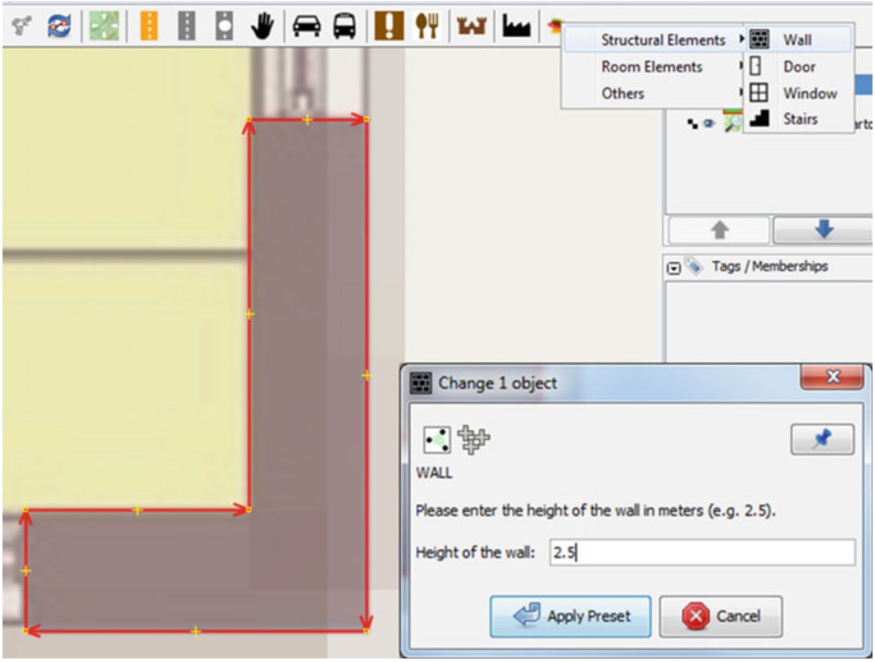
As stated, the creation of the preset should be done by an expert in consultation with people from the area of interest, thus, in this case, the pupils. Afterwards, the new preset can be used by the pupils to create the building model without the need for further intervention of an expert. Figure 3 shows the workflow with the expert step of the preset-definition and the amateur steps of the model creation that will be performed by the pupils.

1. Definition of a preset (expert). The structure of the XML preset-document is a fixed one that can be looked up in the OSM-wiki. This structure has to be followed to define the elements, their attributes and input fields or check boxes as well as the structure of the menu. Additionally, it is possible to define if the preset can be assigned every element or only to points, lines or polygons or a combination, such as only lines and polygons but not points. An example is a river, which can be drawn as a line or polygon, but won’t be represented as a point.
2. Installation of the preset. Before a preset can be used, it first has to be installed in JOSM. Installation means that the user has to tell JOSM about the new preset

and where to find it. Presets can be used from local sources or via server. We chose to provide the preset via server so that every user with the correct link can use the preset without any download.

3. Drawing of an OSM-element. Once the preset is installed, it cannot be used directly. In OSM, the element (point, line, polygon) has to be drawn, then be selected and then, the preset can be assigned via the menu.
4. Assignment to a preset. Once the element is drawn, the user selects it, then switches to the menu, chooses the element from the list and selects it.
5. Setting of the attributes. A popup appears, letting the user fill in additional values if necessary. Other elements might be filled in automatically, such as if the user wants to assign a wall to a polygon. In this case, the user does not have to choose “wall” in the popup, this is done automatically and not shown to the user. Elements in the pop-up window do not need to be mandatory, they also can be optional.
6. OSM building model. The outcome is the building as an OSM-file (OSM building model), which can then be passed on to an expert for postprocessing, i.e. transforming the model from OSM to a 2D and 3D geodatabase (GDB).

Figure 4 shows an example for the JOSM-preset and the pop-up as defined in Fig. 2 as code. The “L-shaped” element is the polygon-wall. In the case of a wall, only the height is needed as an attribute. Other attributes might be set automatically



Please enter the height of the wall in meters (e.g. 2.5).

Height of the wall: 2.5

Fig. 4 Example of a JOSM-preset

and are not necessarily presented to the user. In this case, as soon as the user chooses “wall” in the menu of the structural elements, it is a fact that the user wants the selected element to be a wall. Therefore, these attributes do not have to be asked, but can be set automatically. Only the height cannot be assumed and must be set by the user.

Once the whole building has been drawn with all elements, it can be transformed using an automated workflow implemented in the model builder. When defined properly, the workflow needs to be defined once and can be reused. This step needs to be done by an expert.

5 Results

The outcome of the workflow is a GDB containing all the defined and drawn elements in 2D as well as 3D. Figure 5 shows the automatically created result of a drawn class room with tables, windows, walls, a door, the chalkboard and a radiator. All elements can be queried and used for analysis because they contain all the attributes that the user provided in JOSM in advance.

As standard services were not able to visualize the building in 3D with different floors, we had to find another solution. After various tests, we concluded in using the ArcGIS JavaScript API 4.0 for a web-visualization of the data to make it available to the amateurs for visualization and analyses. The API can visualize 2D as well as 3D data and include real-time data such as sensor streams as well.

The direct comparison of the 2D and 3D-room shows the benefits of a visualization in 3D: While the position of the windows is obvious in the 3D-image, it might not be clear for an amateur user, what the corresponding elements represent in the 2D-image. Thus, a visualization in 3D might facilitate the communication especially for complex rooms to non-experts.

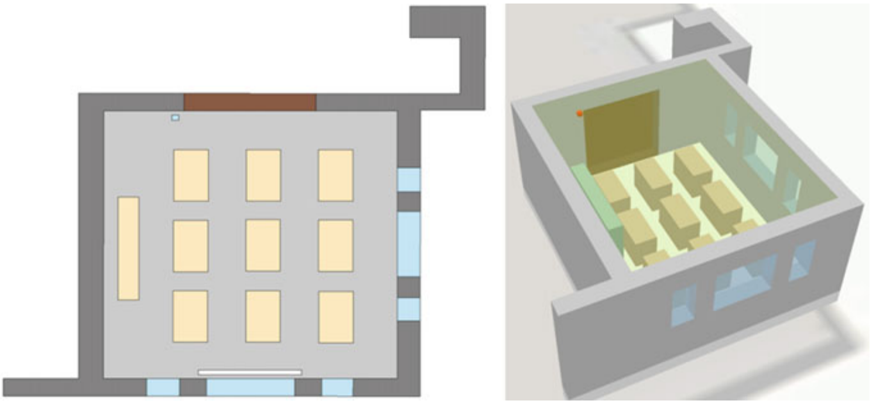


Fig. 5 Resulting building model in 2D and 3D

6 Discussion

In this paper, we proposed a way to use ready-to-use tools to amateurs to create sophisticated and sustainable VGI-building models in a very simple manner. However, there are some issues that have to be discussed.

First of all, while the post processing model works almost fully automated, it cannot be used free of charge as it is implemented using the ArcGIS ModelBuilder which is directly implemented in ESRI ArcGIS for Desktop. To be able to use it, the user needs to license this software and know how to use it. However, to make the post processing accessible to the amateurs, it would be desirable to make it available in e.g. ArcGIS Online as an online geoprocessing tool.

Secondly, using this workflow is not only a convenient way to transform the data, it could also guarantee a certain data quality through the implementation of various processes, data checks and filters. Thus, small mistakes by the users might be avoided. Moreover, we assume that building models are not static. Many of them are highly dynamic. The advantage of using this workflow that transforms the data from the JOSM input into a proprietary ESRI GDB is that it can be rerun automatically with one click. Thus, if changes occur, the users can update the model and the workflow automatically updates the output. Thus, the data can be kept up-to-date very easy.

A big issue in the development of the model is the visualization. Here, we used ESRI's ArcGIS JavaScript API 4.0 as it is right now the only online solution that is able to combine both 3D visualization and real-time data using a GIS-solution. Furthermore, using GIS for visualization provides the possibility to include additional information such as temperature layers, information on bus schedules per bus station etc. Additionally, GIS is not only a tool for visualization: you can use the information, perform analyses or search for specific objects. However, this powerful ability of GIS is often not used in current implementations. One exception is ESRI's Campus Place Locator, which provides the possibility to produce web-based searchable 2D indoor maps.

Another issue is the usability of the building maps. In Google, for example, the user has to click on the building to activate the floor selector, which might not be obvious to the user. However, the standard OSM viewer in its current version does not even have the ability to visualize different building floors at all. Based on our research, visualization, accompanied with the complexity of indoor spaces and their associated models to capture them, seem to be the main hindrances of a widespread capturing of building models.

Using presets in JOSM is not only possible with local files. Presets are also allowed to be on a server. That way, the file can be provided and the users do not have to download it, it can be integrated directly. This becomes handy, when changes of the preset are expected due to adaptations to user needs. Unfortunately, JOSM saves a local file and does not update the preset. We overcame this by writing a small batch script that can be used to reload the preset if necessary.

Another issue which needs to be considered and is not covered by sufficient literature yet, is the privacy of the building data. Using VGI and giving every user the tools to model any building, the user is provided with a powerful tool that should not be used without careful consideration and agreement of the building owner. This is an important topic, but a full discussion about it would go beyond the scope of this paper.

In addition, it remains open why today, users do not want to model the buildings. For example, Google Indoor Maps was launched in 2011, but does not contain many buildings. There are several possible explanations for this behavior. One might be that people don't want to have their building maps online because of privacy issues. Another one is that the available modeling possibilities are too complex, or that because of missing visualization possibilities, the users do either not know about the possibility and benefits of the models or they do not see the benefit of modeling them without the possibility to visualize them. This might be a "chicken-egg-problem": without sophisticated models, there might not be the possibility to visualize them and without visualization, the models won't be developed.

7 Conclusion and Outlook

This paper contributes to the topic of indoor modeling in terms of using VGI to enable building modeling for everyone. Today, the whole building modeling workflow is done by modeling experts, but not by local experts. For many buildings, documentation exists, but it cannot be directly transformed in an easy manner. Using the tools provided by OSM gives the amateurs the tools to model what they really need and in a simple manner that does not discourage them to do so.

Currently, the whole workflow from drawing a building to the 3D model was only tested by some test users, but not yet validated with the pupils of the use case of the project THE4BEES. Thus, the next step will be to go into the classes and let the pupils draw their school buildings. This will show if the workflow really is simple enough for amateurs.

Additionally, in a further step, it would be desirable to put the automatic post processing model online to be used by everyone from amateurs to experts.

This paper might be the first step in the direction to start filling up the missing 2.55 million km^2 of buildings, so that in the future, the buildings will not hold anymore against the mapping invaders. Instead, we can use the provided additional value of indoor building models in applications.

References

Arsanjani JJ, Barron C, Bakillah M, Helbich M (2013) Assessing the Quality of OpenStreetMap contributors together with their contributions. In: AGILE 2013, Leuven

El-Mekawy M, Östman A, Hijazi I (2012) A unified building model for 3D urban GIS. ISPRS Int J Geo-Inf 1(2):120

F4 Map (2016) F4 Map. http://demo.f4map.com/#lat=47.8155404&lon=13.0402142&zoom=18. Accessed 16 Nov 2007

FISDM (2014) FISDM. https://github.com/FISDM/FISDM. Accessed 16 Nov 2007

Flexijet GmbH (2016) Mobile BIM measuring system—the Flexijet 4REVIT. https://www.flexijet.info/en/products/flexijet-4revit/the-flexijet-4revit/. Accessed 16 Nov 2007

Geiger A, Benner J, Haefele KH (2015) Generalization of 3D IFC building models. In: Breunig M, Al-Doori M, Butwilowski E, Kuper PV, Benner J, Haefele KH (eds) 3D Geoinformation science: the selected papers of the 3D GeoInfo 2014. Springer International Publishing, Cham, pp 19–35. doi:10.1007/978-3-319-12181-9_2

Goetz M, Zipf A (2012) Extending OpenStreetMap to indoor environments: bringing volunteered geographic information to the next level. Paper presented at the urban and regional data management, UDMS annual 2011—proceedings of the urban data management society symposium 2011

Google (2016) Use indoor maps to view floor plans. https://support.google.com/maps/answer/2803784?p=gmm_guidelines&visit_id=1-636114317538680574-3652423955&rd=1. Accessed 16 Nov 2007

Helft M (2009) Online maps: everyman offers new directions. The New York Times. http://www.nytimes.com/2009/11/17/technology/internet/17maps.html. Accessed 16 Nov 2007

Isikdag U, Zlatanova S, Underwood J (2013) A BIM-Oriented Model for supporting indoor navigation requirements. Comput Environ Urban Syst 41:112–123. doi:10.1016/j.compenvurbsys.2013.05.001

Knoth L (2015) Spatial information infrastructures for indoor environments. Master's thesis, University of Salzburg

Knoth L (2016) Smarte 3D-Positionierung in Innenräumen. GISSCIENCE Die Zeitschrift für Geoinformatik 2(2016):70–73

Musliman IA, Abdul-Rahman A, Coors V (2010) Incorporating 3D spatial operator with building information models in construction management using Geo-DBMS. In: International archives of the photogrammetry, remote sensing and spatial information sciences—ISPRS archives, vol 38 (4 PART W15), pp 147–154

N.A. (2017) Editor usage stats by number of edits. https://wiki.openstreetmap.org/wiki/Editor_usage_stats#by_number_of_edits

OSM Wiki (2015) Difference between revisions of "Proposed features/IndoorOSM". http://wiki.openstreetmap.org/w/index.php?title=Proposed_features/IndoorOSM&diff=next&oldid=1109857. Accessed 16 Oct 2004

OSM Wiki (2016a) Comparison of editors. http://wiki.openstreetmap.org/wiki/Comparison_of_editors. Accessed 16 Nov 2007

OSM Wiki (2016b) Simple 3D buildings. http://wiki.openstreetmap.org/wiki/Simple_3D_buildings. Accessed 2016 Nov 2007

OSM Wiki (2016c) Simple indoor tagging. http://wiki.openstreetmap.org/wiki/Simple_Indoor_Tagging. Accessed 2016 Nov 2007

Rosser J, Morley J, Jackson M (2012) Crowdsourcing of building interior models. In: Gensel J, Josselin D, Vandenbroucke D (eds) AGILE'2012 international conference on geographic information science, Avignon, 24–27 Apr 2012. Multidisciplinary research on geographical information in Europe and beyond, pp 130–134

Schmitz L, Schroth G, Reinshagen F (2015) Mapping indoor spaces with an advanced trolley. https://www.gim-international.com/content/article/mapping-indoor-spaces. Accessed 16 Nov 2007

Taginfo (2016a) Indoor key uses with the simple indoor tagging scheme/overview. http://taginfo.openstreetmap.org/keys/indoor#overview. Accessed 16 Oct 2004

Taginfo (2016b) Indoor key uses with the simple indoor tagging scheme/values. http://taginfo.openstreetmap.org/keys/indoor#values. Accessed 16 Oct 2004

Tashakkori H, Rajabifard A, Kalantari M (2015) A new 3D indoor/outdoor spatial model for indoor emergency response facilitation. Build Environ 89:170–182. doi:10.1016/j.buildenv.2015.02.036

van Oosterom P, Stoter J, Jansen E (2006) Bridging the worlds of CAD and GIS. In: Zlatanova S, Prosperi D (eds) Large-scale 3D data integration: challenges and opportunities. CRC Press, Taylor & Francis Group, USA

Follow the Signs—Countering Disengagement from the Real World During City Exploration

Markus Konkol, Christian Kray and Morin Ostkamp

Abstract Navigating and exploring unfamiliar urban areas are common and often challenging tasks for tourists, newcomers and other groups alike. Increasingly, people use mobile apps to get support in completing these tasks. A potential side effect of this is disengagement from the real world, i.e. not perceiving the actual environment due to constantly looking at the screen of the mobile device. This in turn can interfere with the construction of a mental map, can reduce people's awareness of their environment, and can cause high mental workload due to frequent attention shifts. In order to counteract these issues, we propose an approach to support navigation and exploration by explicitly managing attention shifts between the virtual and the physical world. It is based on the results of a survey (n = 102) investigating touristic navigation and on the *Blended Spaces* framework by Benyon. We describe the underlying concepts, a prototypical implementation and an initial field study evaluating it. The results provide initial evidence that the approach can successfully support navigation while also facilitating the perception of the environment.

Keywords City exploration · Blended spaces · Pedestrian navigation · Tourism

M. Konkol (✉) · C. Kray · M. Ostkamp
Institute for Geoinformatics, Heisenbergstrasse 2, 48149 Münster, Germany
e-mail: m.konkol@uni-muenster.de

C. Kray
e-mail: c.kray@uni-muenster.de

M. Ostkamp
e-mail: morin.ostkamp@uni-muenster.de

A. Bregt et al. (eds.), *Societal Geo-innovation*, Lecture Notes
in Geoinformation and Cartography, DOI 10.1007/978-3-319-56759-4_6

1 Introduction

Tourists visiting new cities as well as newcomers exploring their new environment typically face the challenge of orientating themselves and finding the way to their destinations. Usually they employ some tools that support navigation, which can be roughly subdivided into two categories: real-world objects and digital means (Montello and Sas 2006). Real-world objects, e.g. signage or landmarks, are embedded in the environment but not always available or potentially ambiguous (Montello and Sas 2006). Paper maps are another example tool in this category, which are popular due to being cheap and portable. They can also be part of a guidebook, which includes further structured information (Brown and Chalmers 2003; Norrie and Signer 2005). Due to their static nature, users of these tools might have to match potentially outdated information with the environment, which can be confusing (Brown and Chalmers 2003). In order to support tourists in particular, many cities have also deployed signage pointing to cultural highlights (Fig. 4 right). This signage is usually embedded in the environment in an unobtrusive way and thus prone to be overlooked by visitors.

Tourists and newcomers nowadays frequently make use of digital means, specifically mobile guides and navigation support tools on smartphones. Such apps can enable efficient navigation as they provide turn-by-turn instructions and self-localization. On the negative side, they often require a high degree of attention (Müller et al. 2008) and can prevent their users from experiencing their environment (Mokey et al. 2013). This in turn may result in varying degrees of disengagement from the real world. In addition, switching attention between a smartphone and the surroundings results in a high cognitive load. Navigating with the help of either real-world objects or by using digital means thus both come with specific benefits and limitations.

In this paper, we therefore tackle the question of how to fluidly integrate the real and the virtual world in order to support navigation and exploration while minimizing disengagement. We report on three key contributions: (i) a novel approach for integrating the virtual and real world in order to support urban exploration and navigation. It is based on the *Blended Spaces* framework by Benyon (2012) and aims at providing a better engagement to the environment by managing attention switches between the physical and digital space. (ii) We demonstrate the practical feasibility of our approach via a prototypical implementation that also incorporates insights from a survey on touristic navigation, and (iii) report on an initial evaluation of our approach.

In the remainder of the paper, we first review related work. We then present key results of our survey on touristic navigation, and describe our approach, which is based on the results and the *Blended Spaces* framework. In the subsequent sections, we present the specific realization of our approach that we then implemented in a prototypical app. This app enabled us to evaluate our approach with users in the field, and we report on key results of this study. We conclude by summarizing our main contributions and by outlining future work directions.

2 Related Work

An extensive amount of research has been carried out revealing differences between everyday navigation and touristic navigation. According to Montello (2005), efficiency is an important aspect in navigation. Turn-by-turn instructions support people in navigating efficiently and are composed of directional instructions, distances, and descriptions of the environment (Lovelace et al. 1999). According to Lovelace et al. (1999), route directions are of high quality if they are complete, precise, and efficient giving information on most of the turns complemented by landmarks. However, efficiency may not be the top priority of people who want to explore a city (Brown and Chalmers 2003). Hence, common navigation tools aiming at increasing efficiency are not necessarily the first choice.

One possible approach is to combine maps with photos (Beeharee and Steed 2006) or panoramic images (Vaittinen et al. 2013; Wither et al. 2013). Both solutions were investigated in user studies revealing pictures are helpful for confirmation in unfamiliar environments but not essential for successful navigation. Also, direction determination is not faster (Vaittinen et al. 2013). Instead, they suggest aiming for simplicity by offering directions via compass-based tools (Vaittinen et al. 2013) which is a key aspect of the next approaches.

Szymczak et al. (2012) and Robinson et al. (2010) successfully used rough route instructions composed of tactile signals indicating the direction and the remaining distance to a target. Participants reached the target and had a better perception of the environment. Giannopoulos et al. (2015) presented *GazeNav*, a hands-free pedestrian guide based on eye-tracking. Users receive vibration signals from their smartphone at decision points if they look at the right direction. This approach was compared to map-based turn-by-turn instructions. The results show that users learn more about their environment without losing efficiency when using *GazeNav*. Schirmer et al. (2015) proposed a hands-free and eyes-free approach thus fostering exploration and reducing disengagement from the environment. They place wearables into the shoes of a user. Tactile cues indicate direction changes at each intersection.

The *Blended Spaces* framework by Benyon enables to identify linkages between the physical and virtual space for mobile apps thus enhancing user experience and a user's perception of the environment (Benyon 2012). He suggests implementing these linkages by using, for example, location-based event-triggering. This approach assists people in not only exploring the physical world but also the virtual space which becomes increasingly important. Several systems already exploit the notion of *Blended Spaces*. The *Global Village Exploration* by Mokey et al. (2013) tracks the user's location and sends notifications if information about a point of interest (POI) is available. The area comprising all POIs is surrounded by a digital border. A user who crosses this border receives instructions for the way back. Apart from that, users do not receive any guidance. O'Keefe et al. (2014) guide tourists in an open-air museum from one POI to the next. Navigation support is provided by a compass on a mobile device. The device notifies visitors if they approach a POI and then shows additional content. However, both approaches do not include real world navigation means during the navigation process.

To conclude, people exploring unfamiliar environments need particular considerations in the context of navigation support. Efficiency is not the top priority. Instead, experiencing new environments appears to be more important. Several approaches already focus on these aspects and propose hands-free and eyes-free navigation instructions. They refrain from providing detailed turn-by-turn instructions as they demand too much attention. However, they do require additional wearables which might be cost-intensive or not available. They do also lack of linkages between the real and virtual space, e.g. references to real-world objects such as signage. These linkages can be created by using the *Blended Spaces* framework. It was already applied to send location-based notifications but was not examined in the context of explorative navigation. Similar to the presented apps, the application presented in this work conveys rough route directions composed of the direction and the distance to a destination. The main difference lies in the usage of signage indicating cultural highlights as a linkage between the real and virtual space. Thus, a user's attention is engaged to the environment while navigating to the target.

3 Survey on Touristic Navigation

Prior to developing an approach, we conducted a survey on touristic navigation. We asked tourists for preferred navigation means when visiting new cities, their benefits and limitations, preferred navigation means when becoming lost, and disengagement caused by smartphones. In the following we present the design, procedure, and results of the survey. The results serve as a basis for the development of our approach which aims at smartly integrating the physical and virtual space in order to shift the attention from one world to another.

3.1 Design and Procedure

The survey started with a brief introduction into the purpose of the study. In the following, we asked participants for their familiarity with the city where the study took place and if they find it easy to explore new cities. We then provided participants with a list of navigation means and asked participants how often they make use of them when visiting new cities and which (dis-)advantages they have. Finally, we asked for the preferred navigation means when becoming lost and the disengagement caused by smartphones during orientation. Two hotels and one tourist information center were asked to distribute the survey to visitors. The completion of the survey took around 10 min and was available for 3 weeks. We also asked potential participants personally and handed the survey out if they informally affirmed they are tourists and not residents. Their answers were analyzed by using *LimeSurvey*,[1] an open source survey tool, and *R*,[2] an open source software for statistical computing. Answers given for the (dis-)advantages were generalized and summed up.

3.2 Survey Results

102 visitors (63 female, 39 male, age ≈40.53, sd ≈13.21) took part. For 35 of them it was the first visit. 54 people visited the city more than once, and 13 participants visit the city regularly. 65 participants disagreed or strongly disagreed with the statement "*I am familiar with the city*". 23 visitors neither disagreed nor agreed with this statement. 13 participants agreed or strongly agreed. While eight people disagreed or strongly disagreed with the statement "*I find it easy to explore new cities*", 34 visitors neither agreed nor disagreed. 60 tourists agreed or strongly agreed.

Figure 1 shows the frequency of use for each device listed. Signage is used most often. 86% orientate themselves with the help of signage often or always followed up by internet (73%), landmarks (67%), and analogue maps (53%). Maps behind billboards (38%), apps on a device (34%), and guidebooks (33%) are used less often but are still popular. Passersby (22%), kiosk displays (8%), and tourist guides (5%) are least popular. In contrast, 40% prefer to ask passersby after becoming lost followed up by mobile apps (31%), signage (25%), and analogue maps (25%). 16% make use of maps behind billboards whereas 15% use landmarks. Internet (9%), guidebooks (4%), tourist guides (1%), and public displays (1%) are used rarely.

Table 1 lists (dis-)advantages mentioned by the participants. The major advantage of mobile apps is availability. In turn, usage is disengaging and using them depends on internet access. Participants appreciate analogue maps for their availability, e.g. in tourism information centers. However, they can also be unhandy and self-localization can become a challenging task. Signage and landmarks provide orientation support but can be missing or hidden. Tourist guides and guidebooks are informative but expensive. Kiosk displays and maps behind billboards provide a good overview but are deployed sparsely. Participants have contrary opinions regarding passersby who are knowledgeable but might give wrong or inaccurate information.

We also investigated if participants feel disengaged when using a smartphone for orientation (Fig. 2). While 37% of the participants do not perceive their environment to a lesser extent, this is the case for 32%. In turn, 64% of the participants are not better able to concentrate on the environment whereas 16% do so. 5% of the participants regularly miss the target (12% do not directly find the target) whereas 68% do not regularly miss the target (62% directly find the target). 19 participants do not make use of a smartphone for orientation at all.

[1] https://www.limesurvey.org, access date 03-Feb-2017.

[2] https://www.r-project.org/, access date 03-Feb-2017.

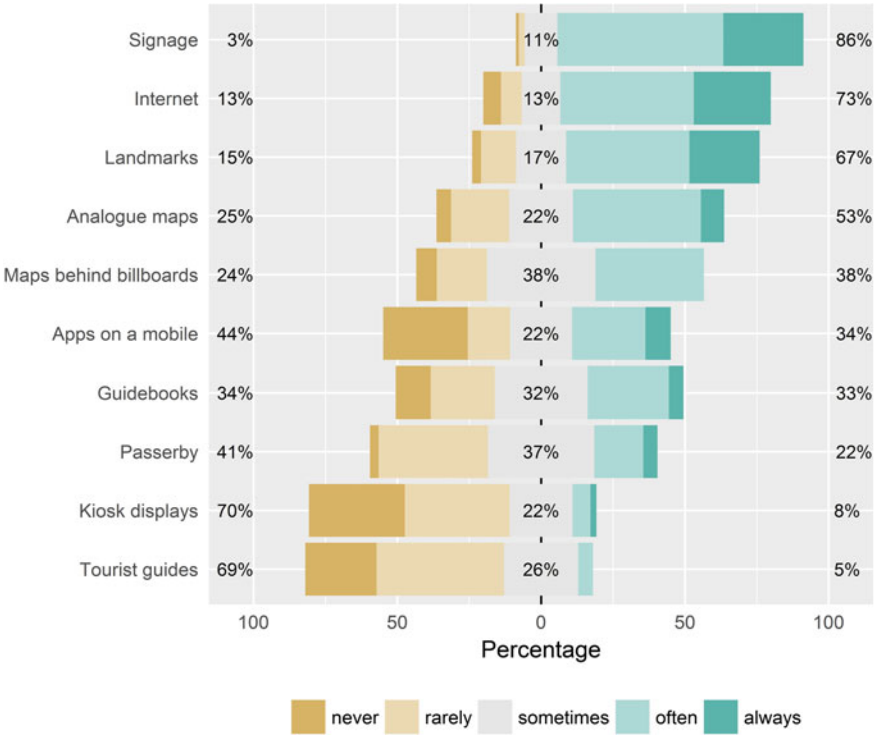


Fig. 1 Frequency of use for each navigation means listed

The *R* workspace including analysis script and dataset is available for reproduction.[3]

3.3 Analysis of Survey Results

The results show that tourists prefer to navigate with the help of real-world objects such as signage or landmarks. Those means are often available and enable fast orientation. Besides, they are embedded in the environment that tourists intend to explore. However, tourists do not refuse technology entirely and make use of virtual assets as well, e.g. in order to search for information. After becoming lost, receiving location information fast appears to be an important aspect. This may be a reason for why people prefer to ask passersby who are familiar with the environment and to use web mapping services based on Global Positioning Service (GPS). One part of the survey examined the disengagement caused by smartphones during the

[3] https://github.com/MarkusKonk/Survey-on-touristic-navigation, access date 03-Feb-2017.

Table 1 Main (dis-)advantages of navigation means. Numbers in the brackets indicate, how many of the participants mentioned them

Navigation means	Advantage	Disadvantage
Apps on a mobile	Availability (60 mentions), orientation support (15) rich in content (10)	Low battery (22), no connection (22), disengaging (10)
Kiosk displays	Overview (8), updated infos (7), aligned to location (7)	Rarely available (12), disengaging (6)
Analogue maps	Availability (25), overview (13), no battery (11)	Unhandy (25), localization challenging (13), outdated (11)
Maps behind billboards	Overview (34)	Sparse deployment (22), immobile (8)
Internet	Rich in content (18), fast (12), availability (16)	Lack of connection (35)
Signage	Orientation support (39)	Missing (16), ambiguous (15), hard to find (10)
Landmarks	Orientation support (22), unique (12)	Not always visible (12)
Passersby	Knowledgeable (46), personal contact (14)	Give wrong/inaccurate infos (46), unfriendly (7)
Tourist guides	Informative (40)	Expensive (26), not individual (11)
Guidebooks	Informative (45)	Outdated (12), expensive (10)

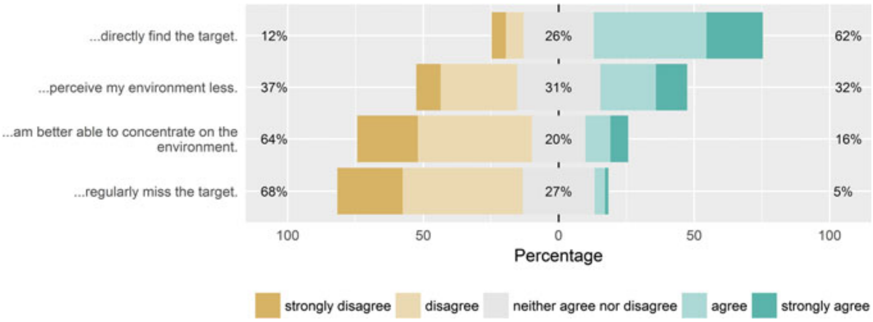


Fig. 2 Degree of agreement for the statement "*If I use a smartphone for orientation, I…*"

process of orientation. Results show that many participants feel at least partially disengaged and pay less attention to the environment while using a smartphone. Although a number of participants do not use smartphones at all or do not like to use smartphones while travelling, it is still a useful navigation means. In most cases, they find their target directly when using a smartphone.

The key findings relevant for this work are the following: Tourists want to perceive the environment and choose the navigation tool accordingly. Signage is popular but deployed sparsely and prone to be overlooked. Smartphones are particularly useful if tourists become lost but are limited as they disengage people from the environment. The opportunity arises to exploit benefits from smartphones and signage.

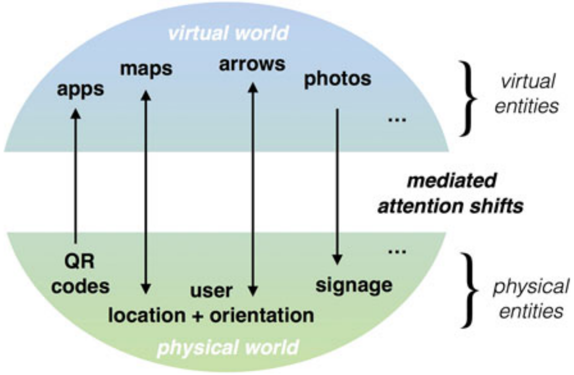


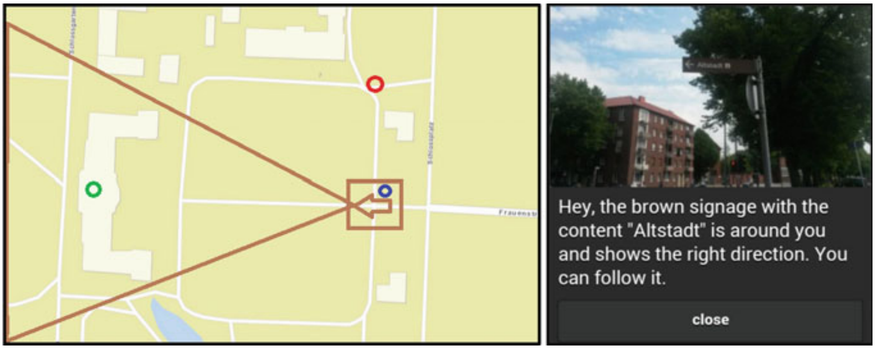
Fig. 3 Mediating attention shifts between the physical and the virtual world: potential transitions between real-world and virtual entities

4 Approach

Our approach is based on the notion of *Blended Spaces* which aims at identifying and specifying linkages between the real and virtual space. Figure 3 illustrates how we use this theoretical framework to explicitly mediate shifts in attention from one world to another in the domain of city exploration. As indicated by the illustration, the virtual and the physical world are separated by a gap. When paying attention to the virtual world, e.g. by using mobile guides, users become disengaged from the physical world. In contrast, real-world objects such as signage do not integrate virtual assets. However, the arrows in Fig. 3 show that both spaces contain suitable entities for referring from one world to the other.

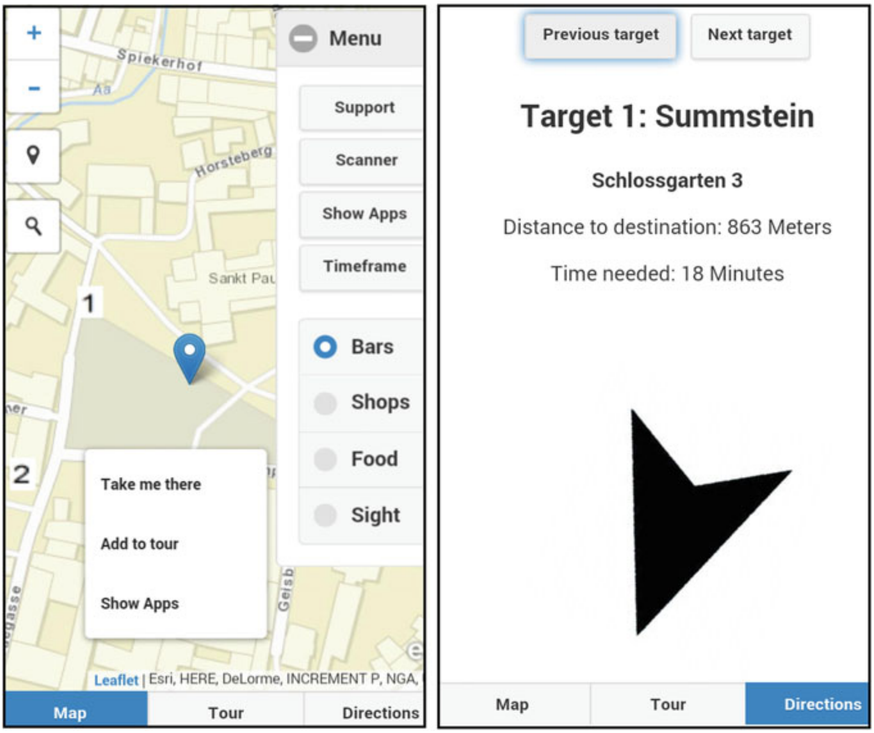
The proposed approach tries to smartly bridge the gap thus minimizing disengagement from the real world while also facilitating exploration of the virtual world that is relevant while exploring new cities (e.g. apps designed for a specific city). As Fig. 3 shows, entities from one space can be used to refer to entities in the other space thus fluidly mediating attention shifts. Unlike previous approaches, we propose to explicitly direct the user's attention from the virtual to the real world and vice versa, thereby avoiding the problem of people continuously looking at their smartphones. Figure 3 shows several entities that we identified in both worlds. We selected them based on *Related Work* and findings from the survey. QR Codes, for example, are physical entities that can be used to trigger a functionality in a mobile app or start an app relevant to the location where the QR Code is attached. Photos of real-world objects (see Fig. 4 right) can direct attention from the virtual to the real world, e.g. to signage and landmarks which are both popular navigation means but not always visible. Similar to the *Blended Spaces* framework, this approach considers location and orientation, too.

While maps (see Fig. 5) can direct the user's attention to a location, arrows can be used to referring to a certain orientation. In contrast, the location or orientation of the user can also trigger the display of a map or arrow. One possible scenario in which users benefit from the system is as follows: Users can fluidly direct their



Hey, the brown signage with the content "Altstadt" is around you and shows the right direction. You can follow it.

Fig. 4 *Left* User (*blue circle*) receives a notification if the target (*green circle*) is within the *brown triangle* but not, if the target (*red circle*) is outside. *Right* Notification about signage that points at the same direction a user has to follow



Target 1: Summstein

Schlossgarten 3

Distance to destination: 863 Meters

Time needed: 18 Minutes

Fig. 5 *Left* Map section of the app. *Right* Navigation support composed of the time needed to reach the target, distance, and an *arrow* indicating the direction

attention from real-world navigation means to their smartphone if signage is missing and back to the environment if the application points at existing signs.

We implemented several of these transitions in order to evaluate whether they can support tourists and newcomers in navigating and exploring the real and the virtual space. The following two subsections describe the selection of linkages (attention shifts) that we realized in our application, and how we implemented the prototype in order to carry out an initial user study.

4.1 Realization

As a first option, users are made aware of signage pointing at cultural highlights. After selecting a destination, users receive a notification about existing signage onto their smartphone if they approach a signage which roughly points at the same direction that users have to follow to reach their target. Therefore, several signs were photographed and their coordinates were stored in a database. In addition, two types of polygons were created by hand for each sign. The first polygon is a square with a distance of 20 m from the location of the sign (Fig. 4 left, brown square). Since a sign points to a certain direction, it can support people in reaching numerous destinations. These destinations are comprised by the second polygon (Fig. 4 left, brown triangle). Thus, each sign points to a certain area. The size and the shape of the area differ depending on the content of the sign. Figure 4 (left) illustrates a simplified scenario: A user (blue circle) approaches a sign and enters the square. The selected target (green circle) is within the polygon for which the signage is relevant. Consequently, the user receives a notification. A notification about relevant signage is composed of a vibration signal and a popup containing a photo of the sign and a message, e.g. "*Hey, there is a sign with the content 'Church' next to you. It shows into the right direction. Follow it*" (Fig. 4 right). The photo shows the sign and details of the scene thus enabling to detect the sign in the surrounding of a user. In contrast, the user does not obtain a notification if the target (red circle) is outside the polygon. In such cases, users do not benefit from the sign. As a result, users may be made aware of the signage and the depicted landmark. One of the main limitations of signs is their sparse deployment. Thus, following them can lack of confirmation on the way. We therefore created QR Codes as a second linkage between the real and virtual space. They are attached to signs and contain information on the address and coordinates of the destination displayed on the sign. Users scan the QR Code and then receive route directions to the corresponding target. Users are thus able to take the sign away and receive confirmation if they do not encounter further signage.

The system supports people in exploring the virtual space as well. People exploring a city may be interested in apps related to their current position and that may assist their activity. Apps can be relevant for the whole area of a city, e.g. apps for public transport, or smaller regions, e.g. an app for the zoo. Consequently, the system proposes only apps relevant for the current location of the user.

4.2 Implementation

We implemented the system using *Apache Cordova*,[4] a framework for developing mobile apps using web technologies and device features such as *Global Positioning Service* (GPS) and the compass. The app is composed of three parts: *Map*, *Tour*, and *Directions*. The *Map* section (Fig. 5 left) enables to localize the device and destinations. The menu allows to search for apps related to a user's location or a position on the map and to switch between heatmaps, each showing hotspots for a certain category. By clicking on the map, users can add locations to the *Tour* section which generates a tour based on the algorithm of the *Travelling Salesman Problem* (TSP). Approximating the TSP results in a shortest route that starts and finishes at the origin and visits each target once (Dantzig et al. 1954). Users gain navigation support in the section *Directions* (Fig. 5 right). It contains the destination address, distance, time needed to reach the target, and an arrow pointing at the target.

For proposing apps, we collected and stored those apps that are relevant for tourists. Each app entry is enriched by polygons defining the areas for which the particular app is relevant. Users receive a list of app proposals by triggering the functionality *Show Apps* in the *Map* section. It uses the current location of the device and checks if this position is within the area for which the app is relevant.

4.3 Data Preparation

Few steps are required to prepare the data for the application. First, signage needs to be photographed and gathered including information on location, orientation, and labelling. Until now, information on this type of signage is not publicly available, e.g. as mapped features in OpenStreetMap.[5] Then, QR codes including the information are attached to signs thus allowing users to take them away.

Finally, a list of apps is needed to realize the app proposals.

5 Initial Evaluation

In order to gain initial insights into the feasibility, effectiveness and usability of the approach, we conducted a field study using the prototypical implementation described above, which was followed by semi-structured interviews.

[4] https://cordova.apache.org/, access date 02-Feb-2016.

[5] https://www.openstreetmap.org, access date 03-Feb-2017.

Participants: We asked three employees (2f, 1m, mean age ≈51.67, sd ≈1.25), subsequently abbreviated with E, from a local tourism information to take part in the study. They have expertise in the field of tourism and thus can give useful insights. All of them were familiar with the environment and knew about the signs indicating cultural highlights.

Additional six participants (6f, mean age ≈25.17, sd ≈4.3), subsequently abbreviated with *P*, were recruited via social media, or by contacting them in the university. They were informally asked if they are familiar with the city and selected if this was not the case. The six participants, all of them newcomers, were composed of three Ph.D. students in the field of GIS and two women without a background in GIS who just moved to the city. One student with a background in Geoinformatics came from a nearby city. By asking newcomers, we aimed at receiving a broader cohort of participants thus complementing the tourists we asked for the survey.

Scenario: The scenario was aligned to what people who visit a city may do. Participants were tasked to create a tour from their current location via two sights finishing at the second sight. Both sights are located in the city center which is a popular place for tourists and thus suitable to simulate a situation in which people explore a city. Participants were asked to use the navigation support provided by the prototype. After arriving at the second target, participants were tasked to search for information related to the second sight. This might be a realistic scenario as tourists or newcomers may want to know more about the sight they visit. The route was slightly different for the experts and the remaining participants. The starting point and the first target was the same for every participant. The route for the experts was composed of one turn, included two references to brown signs, and finished at a church (distance: ≈350 m). As this route appeared to be too easy, the remaining participants had to turn three times, encountered three brown signs, and finished at another church (distance: ≈500 m). However, comparisons of the answers in the semi-structured interviews are based on similar experiences as both had to follow the arrow and received notifications onto their smartphone.

Procedure: First, participants received a short explanation and a questionnaire collecting demographic information, i.e. gender and age. After the introduction, participants received the smartphone with the installed system. The apparatus was a *Samsung Galaxy S4 Mini* running *Android 4.4.2* and having a display size of 4.3″. The participants were told that the smartphone might vibrate at some point of the tour. In this case, they should take a look at the display to see what happened. In addition, participants were told that they do not have to hurry. Finally, by telling them that they do not have to look at the display permanently, we aimed at avoiding participants to focus on the app more than they would do outside a study situation. They were allowed to hold the smartphone loosely in their hand and to look at it only if confirmation is needed. We asked participants to think aloud and recorded their statements with the *Zoom H2n Handy Recorder*. An observer followed them taking notes on their behavior. In the semi-structured interviews, participants were asked in which situations navigation support by an arrow, references to signage, and a combination of both might be helpful. They were also asked about situations in which tourists might (not) benefit from our app and if they can imagine other objects to which the system might refer. Signage and QR Codes are only two potential linkages between the real and virtual space. We finally asked about missing functionality. The procedure took around 40 min in total. Six interviews were conducted in German language, the remaining three in English. All interviews were recorded with the *Zoom H2n Handy Recorder*, anonymized, and then transcribed. We then screened the interviews and the think aloud recordings in order to collect key statements. Finally, the answers to the questions were categorized, i.e. similar comments were added together.

Results: All participants found the two targets by using the application but suffered from an inaccurate GPS localization. Thus, the notification for the first sign appeared too late in four out of nine cases. Though, all participants found the sign based on the photo. In two cases, the compass indicated wrong directions while approaching the first target and needed re-calibration. Apart from one person, all participants found the sign for the second trajectory though the message came too late in eight cases.

Five participants named simplicity as a benefit of the system. P6 stated:

> My parents would use this easier than Google Maps because there you have to enter addresses […] and to zoom in and out.

E1 liked that the arrow points directly to the target making identification easier if the look of the target is not known. Furthermore, E2 stated people are familiar with using arrows, e.g. by using in-car navigation systems. P1 can imagine people become more aware of buildings they would not see otherwise. Six participants mentioned the system improves the perception of the environment. P5 reported:

> It is really helpful to see the surroundings while you are walking. […] It forces you to look up and try to look for something that is not on your phone to get to the place.

Three participants mentioned the arrow and signage complement each other, e.g. if either the localization or the arrow is inaccurate. P5 struggled with a defective compass and stated:

> The arrow is like a compass, it sometimes goes everywhere. If you find signage that complements what the arrow is telling you, then you just can relax and say: It is over there. Then I go until I need the arrow.

The feature for proposing apps based on the user's location had mixed responses. While five participants would make use of it, the remaining four find it too time-consuming and as P5 stated:

> You need to find the app, to download the app, and then understand the other app, and trying to get further information for the place.

Participants named several situations in which tourists can benefit from our app. Two participants referred to people who arrive in a new city, e.g. at train stations. One participant named visitors who are in a hurry. Another four considered those who are unfamiliar with a city, e.g. people who just moved there. In contrast, two participants stated the system is not useful for those who already understand the city structure. Six participants suggested referring to potential sights such as buildings while navigating whereas P6 considered:

> It can be interesting to have a stop, look at that also, and then continue but not for helping in navigation because bringing too much information, it may bring confusion.

Asking about missing functionality, the three experts agreed to include recommendations for potential sights. The experts and three of the participants also wished to have at least a basic set of information on sights in the app itself making it possible to search for apps only if detailed information is needed.

6 Discussion

Overall, the approach presented in this work seems to successfully support people in navigating and exploring new cities without disengaging them from the environment. It addresses key limitations of existing navigation tools we identified from a survey such as hidden or missing signage and disengagement caused by mobile guides. The system fluidly integrates real-world objects (signage) and virtual assets (GPS-based event-triggering) based on the *Blended Spaces* framework. Participants liked that their attention is directed to their surroundings and that it is not necessary to keep track of the screen. Participants emphasized the simplicity of the arrow as a good feature. This might also result from being familiar with arrows, e.g. by using in-car navigation systems. Although the approach considers users who are not in a hurry, one participant stated the system may also suit to those who quickly want to reach a target. This can be an indication that the application supports efficient navigation as well. Further research is needed to assess this aspect in detail.

While five participants liked the app proposals based on the user's location, the remaining four rejected it. This may be caused by the list which provided too many options. People may find it exhausting and time-consuming to look through all apps. A proposed app should address their needs, e.g. by using filter options. However, the idea of proposing apps still appears to be useful.

The presented work is subject to a number of limitations. The evaluation was conducted with people who were newcomers and not actual tourists as we aimed at having a broader cohort of participants complementing the tourists we asked for the survey. Hence, their behavior and their answers may differ from what "real" tourists may do and say. However, it became clear that the approach might also be beneficial for inhabitants who are not familiar with the city. Nevertheless, it is still necessary to evaluate the approach with tourists.

Participants were asked to think aloud and were accompanied by an observer. These aspects might have led to a behavior which is different from everyday situations. One key limitation is the localization inaccuracy which resulted in notifications coming too late and users having to turn around to search for the sign shown on the photo. Ideally, people receive a message while approaching the sign from the front. The magnetometer did not work properly in two cases and showed wrong directions making reconfiguration necessary. Many people may not be aware of these problems and may follow a misbehaving compass and an inaccurate localization. However, in case of inaccuracies, signage and arrow complement each other as mentioned by the participants. Except for the mentioned inaccuracies, we successfully implemented signage as a linkage between the virtual and real world.

7 Conclusion

The key contributions of this work are the following:

(i) We proposed a novel approach for mediating attention shifts between the physical and the virtual world in the context of city exploration. The approach directly addresses one limitation of standard mobile guides, i.e. the high degree of attention required to follow turn-by-turn instructions. The proposed approach aims at engaging attention to the real space but also facilitates exploring the virtual world. We first scrutinized related work and identified rough route directions composed of the direction and the remaining distance to a target as a suitable component. We combined these directions with a fluid integration of virtual assets and real environments based on our survey on touristic navigation and the *Blended Spaces* framework.

(ii) We then implemented specific instances of (i) in a prototypical implementation. We therefore smartly integrated both spaces with the help of popular navigation means such as signage and digital maps. While being on the way, the application refers to signs indicating cultural highlights and point to the same direction a user has to follow to reach the target. In addition, the system proposes apps based on the user's location in order to assist tourists in exploring the virtual space.

(iii) We used the implementation to carry out an evaluation of the approach. The results from a user study provide initial evidence that the app successfully addresses several disadvantages we identified in the survey such as disengagement caused by mobile guides and missing signage. Nevertheless, localization and orientation inaccuracies are key issues. The feature for proposing apps is promising but needs to address a user's activity.

It would be interesting to fully understand the degree of disengagement caused by our application. The Resource Competition Framework (Oulasvirta et al. 2005) measures a user's attention paid to a task performed on a mobile phone and the environment. This approach allows comparing disengagement caused by the app with standard mobile guides. Similarly, Kiefer and Giannopoulos (2015) measure a user's attention by using eye tracking. Another aspect is to examine references to other types of objects as proposed by the participants in the evaluation. Venues and buildings potentially complement the signs used in this work. Finally, an improved system might also be subject to a quantitative study in another city with a different setting and a larger cohort of participants in order to validate findings reported here.

References

Beeharee AK, Steed A (2006) A natural wayfinding exploiting photos in pedestrian navigation systems. In: Proceedings of the 8th conference on human-computer interaction with mobile devices and services (MobileHCI '06). ACM, p 8188

Benyon D (2012) Presence in blended spaces. Interact Comput 24(4):219–226

Brown B, Chalmers M (2003) Tourism and mobile technology. In: Proceedings of the eighth European conference on computer supported cooperative work (ECSCW 2003). Springer, Netherlands, pp 335–354

Dantzig G, Fulkerson R, Johnson S (1954) Solution of a large-scale traveling-salesman problem. J Oper Res Soc Am 2(4):393–410. Informs

Giannopoulos I, Kiefer P, Raubal M (2015) GazeNav: gaze-based pedestrian navigation. In: Proceedings of the 17th international conference on human-computer interaction with mobile devices and services (MobileHCI '15). ACM, pp 337–346

Kiefer P, Giannopoulos I (2015) A framework for attention-based implicit interaction on mobile screens. In: Proceedings of the 17th international conference on human-computer interaction with mobile devices and services adjunct (MobileHCI '15). ACM, pp 1088–1093

Lovelace KL, Hegarty M, Montello DR (1999) Elements of good route directions in familiar and unfamiliar environments. In: International conference on spatial information theory (COSIT '99). Springer, Berlin, pp 64–82

Mokey S, Nalbandian A, O'Keefe B (2013) Location as interaction: exploring blended spaces in the global village. In: Proceedings of the 27th international BCS human computer interaction conference (BCS-HCI '13). British Computer Society, pp 52:1–52:5

Montello DR (2005) Navigation. The Cambridge handbook of visuospatial thinking. Cambridge University Press, pp 257–294

Montello DR, Sas C (2006) Human factors of wayfinding in navigation. International encyclopedia of ergonomics and human factors. CRC Press/Taylor & Francis, Ltd, pp 2003–2008

Müller J, Jentsch M, Kray C, Krüger A (2008) Exploring factors that influence the combined use of mobile devices and public displays for pedestrian navigation. In: Proceedings of the 5th nordic conference on human-computer interaction: building bridges (NordiCHI '08). ACM, pp 308–317

Norrie M, Signer B (2005) Overlaying paper maps with digital information services for tourists. In: Proceedings of the international conference information and communication technologies in tourism 2005. Springer, Vienna, pp 23–33

O'Keefe B, Benyon D, Chandwani G, Menon M, Duke II R (2014) A blended space for heritage storytelling. In: Proceedings of the 28th international BCS human computer interaction conference on HCI 2014—sand, sea and sky holiday HCI (BCS-HCI '14). British Computer Society, pp 90–99

Oulasvirta A, Tamminen S, Roto V, Kuorelahti J (2005) Interaction in 4-second bursts: the fragmented nature of attentional resources in mobile HCI. In: Proceedings of the SIGCHI conference on human factors in computing systems (CHI '05). ACM, pp 919–928

Robinson S, Jones M, Eslambolchilar P, Murray-Smith R, Lindborg M (2010) "I Did It My Way": moving away from the tyranny of turn-by-turn pedestrian navigation. In: Proceedings of the 12th international conference on human computer interaction with mobile devices and services (MobileHCI '10). ACM, pp 341–344

Schirmer M, Hartmann J, Bertel S, Echtler F (2015) Shoe me the way: a shoe-based tactile interface for eyes-free urban navigation. In: Proceedings of the 17th international conference on human-computer interaction with mobile devices and services (MobileHCI ’15). ACM, pp 327–336

Szymczak D, Rassmus-Gröhn K, Magnusson C, Hedvall P (2012) A real-world study of an audio-tactile tourist guide. In: Proceedings of the 14th international conference on human-computer interaction with mobile devices and services (MobileHCI ’12). ACM, pp 335–344

Vaittinen T, Salminen M, Olsson T (2013) City scene: field trial of a mobile street-imagery-based navigation service. In: Proceedings of the 15th international conference on human-computer interaction with mobile devices and services (MobileHCI ’13). ACM, pp 193–202

Wither J, Au CE, Rischpater R, Grzeszczuk R (2013) Moving beyond the map: auto-mated landmark based pedestrian guidance using street level panoramas. In: Proceedings of the 15th international conference on human-computer interaction with mobile devices and services (MobileHCI ’13). ACM, pp 203–212

Perspectives in Externalizations of Mental Spatial Representations

H. Löwen, A. Schwering, J. Krukar and S. Winter

Abstract Place is a core component of human spatial knowledge and therefore a central topic in GI Science. People use externalizations of mental spatial representations to communicate about space. Textual descriptions and graphical descriptions are the two main modes of communication. In this paper a distinction of three scales of spatial descriptions is assumed and textual and graphical descriptions are collected and analyzed in order to investigate the differences between the spatial descriptions. Thereby the focus lies on the properties and perspectives of the descriptions. It is found that within the textual descriptions people tend to not consistently use one perspective, but switch perspectives and predominantly apply the route perspective. For the graphical descriptions there has been no clear categorization of description perspectives. However, there are differences in the properties of these descriptions that indicate different perspectives.

Keywords Spatial information extraction · Communication modes · Spatial description perspectives

1 Introduction

Externalizing *mental spatial representations* always involves cognitive transformation processes, which include invoking parts of the mental representation and encoding it into a chosen modality (Richter and Winter 2014). The two main communication modes for externalizing mental spatial representations are on the one hand spoken or written language and on the other hand graphical configura-

H. Löwen (✉) · A. Schwering · J. Krukar
Institute for Geoinformatics, University of Muenster, Heisenbergstraße 2,
48149 Muenster, Germany
e-mail: loewen.heinrich@uni-muenster.de

S. Winter
Department of Infrastructure Engineering, The University of Melbourne,
Melbourne, VIC 3010, Australia

A. Bregt et al. (eds.), *Societal Geo-innovation*, Lecture Notes
in Geoinformation and Cartography, DOI 10.1007/978-3-319-56759-4_7

tions, i.e. drawn sketches. In order to externalize spatial information into one of the communication modes, people must take a perspective on it (Taylor and Tversky 1996). The perspectives of spatial descriptions imply certain properties of how space is described, e.g. viewpoint, frame of reference[1] and terms of reference[2] (Taylor and Tversky 1996). These might differ with respect to the scales of spatial descriptions. Three scales of spatial descriptions are distinguished here, which will be outlined in the following section: (1) place descriptions, describing where something is located, (2) route descriptions, describing a suggested path between locations, and (3) region descriptions, describing the configuration of whole regions.

In order to improve the understanding of how spatial descriptions are structured and how they differ from each other, the following questions will be investigated:

(1) Do the properties of spatial description differ between the modes of communication?
(2) Does the choice of perspectives differ between the scales of spatial descriptions?
(3) What are the preferred perspectives that people choose in spatial descriptions?

The underlying hypothesis of this paper is that the driving force for choosing a particular perspective is the mode of communication and not the scales of spatial descriptions. The perspectives of externalizations of mental spatial representations will be investigated by comparing the textual and the graphical descriptions within the three scales of spatial descriptions. Empirical data will be collected for all modes and scales of spatial descriptions.

This understanding is applicable in the field of Human-Computer Interaction, for designing systems which communicate spatial information to its users across different scales and context of use. Similarly, Volunteered Geographic Information applications can benefit from better understanding the perspectives naturally preferred by humans to communicate spatial information across different scales. The next two sections will review the related work and expound the methodology with respect to previous research in this area. This will imply the design of a user experiment to collect different spatial descriptions. The descriptions will be analyzed by the properties of textual and graphical descriptions and the results will be presented in Sect. 4 and discussed in Sect. 5.

[1] Taylor and Tversky distinguish three frames of reference: (1) *relative*, where the origin of the coordinate system is one of the participants, the speaker or the addressee, (2) *intrinsic*, where the origin of the coordinate system is a specific object, and (3) *extrinsic*, where the origin of the coordinate system is external to the scene.

[2] The two different terms of reference are (1) LRFB = left, right, front, back, and (2) NSWE = north, south, west, east.

2 Related Work

This section will review the relevant literature on communication modes for externalizing mental spatial representations, perspectives of spatial descriptions and scales of space. The latter will be the base for the distinction between scales of spatial descriptions, which will be presented in the Sect. 3.

2.1 Communications Modes

As mentioned above, the two main communication modes for externalizing mental spatial representations are the textual and the graphical mode. The process of externalizing mental spatial representations into communication modes involves cognitive processes which lay a filter between the mental spatial representations and the spatial descriptions. These cognitive processes consist mainly of (1) invoking portions of the mental spatial representation from long-term memory into working memory and (2) mapping selected elements of the working memory into the particular communication mode (Richter and Winter 2014). Moreover, there is no one-to-one correspondence of the working memory and the expression, but there are many possible expressions so that they do not give a direct clue on the mental spatial representations (Richter and Winter 2014).

Sketch maps are two-dimensional pictorial representations and descriptions of spatial locations, spatial configurations and routes. They depict a filtered, abstracted and schematized subset of a mental spatial representation and reflect cognitive distortions of mental spatial representations. In contrast to sketch maps, textual descriptions are linear descriptions of locations, configurations or routes (Richter and Winter 2014). Richter and Winter state that textual descriptions will have less impact on route descriptions, as routes are linear and textual descriptions have a linear structure as well. Therefore, textual descriptions have a stronger impact on location and configuration descriptions because they require cognitive linearization strategies (Richter and Winter 2014).

2.2 Perspectives of Spatial Descriptions

2.2.1 Textual Descriptions

In literature three kinds of reference frames as well as three kinds of perspectives are distinguished for textual spatial descriptions (e.g. Buhler 1982; Carlson-Radvansky and Irwin 1994; Levelt 1984, 1989; Levinson 1996; Taylor and Tversky 1996). Levinson summarizes the three reference frames as follows:

Table 1 Properties of types of description perspectives (reproduced from Taylor and Tversky 1996)

Properties	Description perspective		
	Gaze	Route	Survey
Viewpoint	Fixed, external	Changing, internal	Fixed, external
Verbs	Stative	Active	Stative
Referent	Object (or person)	Person	Object
Terms of reference	LRFB	LRFB	NSEW
Frame of reference	Relative	Intrinsic	Extrinsic
World analog	View entire scene from fixed point, horizontally displaced	View while exploring	View entire scene from fixed point, vertically displaced (map)

(1) The *relative* reference uses one of the participants as the origin of the coordinate system and describes the locations of an object in relation to that individual's *front, back, left* and *right*, with respect to some other object in the scene (ternary relation), e.g. "the man is to the left of the house".
(2) The *intrinsic* reference frame uses a specific object as origin of the coordinate system and describes the location of the other objects in relation to the object's intrinsic *front, back, left, right, top* and *bottom* (binary relation), e.g. "the man is in front of the house". In this case, the origin could also be a person as for example in route descriptions.
(3) The *extrinsic* or *absolute* reference frame uses an origin of the coordinate system that is external to the scene and most commonly describes the location of objects in relation to the cardinal directions *north, south, east* and *west* (binary relation), e.g. "the man is north of the house".

Taylor and Tversky suggest the distinction of three perspectives of describing environments. These are related to the three reference frames and reflect a natural way of experiencing and describing an environment (Table 1):

(1) The *gaze* perspective takes a stationary viewpoint from which the entire scene can be viewed (horizontally displaced) and applies the *relative* reference frame for the description. It is restricted to the vista space.
(2) The description in the *route* perspective corresponds to the *intrinsic* frame of reference and describes the scene from a changing viewpoint within the environment, analogous to viewing an environment by exploration.
(3) The *survey* perspective describes an environment from a fixed external viewpoint (vertically displaced) and corresponds to the *extrinsic* reference frame. It is analogous to the descriptions from the birds-eye-view (e.g. Ehrich and Koster 1983; Levelt 1982; Taylor and Tversky 1996).

Taylor and Tversky discuss that different characteristics of the environment affect the selection of perspectives. Two of these characteristics are the number of paths through the environment and the sizes of the environment (Taylor and Tversky 1996). Moreover, Taylor and Tversky showed that people mix perspectives and in a later study they investigated why people mix perspectives in textual descriptions of the environment about half the time (Taylor and Tversky 1996; Tversky et al. 1999). They outline that there are cognitive costs related to the descriptions for both retaining a perspective and switching perspective. Parameters of the descriptions are the referent object, the viewpoint and the terms of reference. These are related to the cognitive costs and may change when the perspective changes. Possible reasons for mixing perspective might be that at some point the costs for retaining perspective might not be higher than changing perspective and switching perspective may be more effective in communication than not switching perspective (Tversky et al. 1999). Another reason for switching perspective is that people perceive and represent the environment from multiple perspectives simultaneously. However, when environments are well-learned and presumable abstracted into a perspective-free representation, cognitive costs of switching perspectives disappear (Tversky et al. 1999; Vasardani et al. 2013).

2.2.2 Graphical Descriptions

As verbal and graphical descriptions are fundamentally different, graphical descriptions do not necessarily have to fit the properties and types of verbal descriptions. Graphical descriptions mainly depict objects in an environment and the spatial relations among them. Bryant and Tversky distinguish between the *inside* and *external* perspective within graphical descriptions (Bryant and Tversky 1999). The inside perspective uses an references frame that is centered to a person and describes relations with respect to the person's *front, back, left, right, head* and *feet*. The external perspective is looking from an external position onto the scene and uses a reference frame that is centered to an external object or a person. It describes the directions according to the objects' *front, back, left* and *right*. Other studies term the two perspectives *route perspective*, which adopts a first-person perspective from within the scene (e.g. a mental tour), and *survey perspective*, which adopts a third-person top-down perspective onto the scene (e.g. a map) (e.g. Galea and Kimura 1993; Hund et al. 2012; Kato and Takeuchi 2003; Lawton 1996; Lawton and Kallai 2002; Pazzaglia and Beni 2001; Sholl et al. 2000). However, the boundaries might not always be that sharp because people might depict the environment from the top-down perspective but include for example detailed facades of buildings. The properties and perspectives that will be used for the investigation in this paper will be outlined in more detail in the following section.

2.3 Scales of Space

The term *space* is often structured with respect to different scales, however, a clear categorization is challenging because different disciplines have a different understanding of the term and there are often no clear boundaries between the classes (Montello 1993). Freundschuh and Egenhofer, for example, distinguish between small- and large-scale spaces and Montello distinguishes four different categories of space (Freundschuh and Egenhofer 1997; Montello 1993):

- *Figural space* is defined as the space that is projectively smaller than the human body and can be directly perceived from one place without appreciable locomotion (e.g. small objects or pictures).
- *Vista space* is defined as the space that is larger than the human body but can be visually apprehended from one place without appreciable locomotion (e.g. a single room).
- *Environmental space* is defined as the space that is larger than the human body but can not be visually apprehended without considerable locomotion (e.g. a building or a city district).
- *Geographical space* is defined as the space that is larger than the human body but can not be apprehended directly through locomotion and has to be learned from representations such as maps (e.g. a country).

In the following section a distinction between three scales of spatial descriptions will be presented, which is related to Montello's categorization of space.

3 Method

As a central part of this paper an experiment will be designed, which will collect spatial descriptions. For each scale of spatial descriptions that will be presented in the first part of this section (place, route and region description), two spatial descriptions will be collected: one textual description and one graphical description. The descriptions will be collected from the same group of participant and the order of tasks will be randomized. In contrast to the experiments of Taylor and Tversky (1992a), where participants studied maps and recalled the maps in spatial descriptions, textual and graphical descriptions will be collected from participants by asking them to recall the information from memory only. The analysis of the description perspectives will, among others, involve categorizing the descriptions and measuring the frequencies of the different properties and perspectives of the descriptions. Thereby, the differences in the perspectives within the different scales of spatial descriptions and within the two modes of externalization will be investigated. Moreover, it will reveal which perspective participant preferably choose for the particular scales and modes.

3.1 Scales of Spatial Descriptions

A categorization of different scales of spatial descriptions is related to the scales of space, as outlined in the previous section. A clear categorization is challenging because different disciplines have a different understanding of the term *space* and there are often no clear boundaries between the classes (Montello 1993). A categorization of different scales of spatial descriptions might therefore not be universal, but is defined here as follows:

(1) *Place descriptions* support the localization of objects in the environment and identify locations. Place descriptions are centered to one point which is the location of a clear figure on the ground of the environment. In case of an emergency, it might be the description of the visual surroundings of the particular place or the description of the location in relation to nearby (global) landmarks. The intention of a place description is the identification of a particular location.

(2) A *route description* is sequential and focuses on the guidance through an environment or to a particular destination (Shanon 1979). It thereby describes the location of a moving figure on the ground of the environment.

(3) *Region descriptions*, in contrast to place descriptions (and route descriptions), describe larger environments by describing the location and configuration of several objects (Emmorey et al. 2000; Taylor and Tversky 1992b). It might be categorized by varying figure-ground relations. The intention of a region description is to answer the *where* question for several objects and locations and to answer the *how* question—"how is the environment structured?"

This categorization will be used for the data collection to request the different types of descriptions from the participants.

3.2 Perspective

The textual and the graphical modes are fundamentally different and it was reviewed that there exists no such clear categorization of perspectives in graphical descriptions as presented by Taylor and Tversky (1996) for textual descriptions. The suggested properties in Table 1 will be used to classify and compare the textual descriptions.

For the graphical descriptions, there are properties that can be applied for a distinction. The *reference frame*, which was mentioned by Bryant and Tversky (1999), might either be intrinsic or extrinsic. However, it is expected that participants will predominantly apply the extrinsic reference frame, which corresponds to the fixed, external viewpoint in Taylor and Tversky's distinction. Another property will be the *referent*, which will be distinguished here in three categories:

(1) The referent will be considered to be a *person*, if the participant, the addressee or the imagined location of a person is explicitly depicted and other objects in the sketches will be closely aligned to the location of the person. Moreover, the person will be predominantly, but not exclusively, depicted in the center of the sketch.

(2) The referent will be considered to be a *route*, if the sketch will be clearly limited to the space between the origin and the destination and most objects in the sketches will be closely related to the route itself (e.g. landmarks along the route or at decision points Anacta et al. 2014). However, routes might still be enriched with global landmarks, which are not sketched directly next to the routes (Anacta et al. 2014).

(3) If there is no clear referent or focus in the sketch, but the description might be categorized by multiple objects that form a larger two dimensional extend, the referent will be considered to be the *region*. In this case the purpose of the sketch will be the depiction of the configuration of the whole region.

A further property to analyze graphical descriptions of the environment is the *alignment*. It will be distinguished here between *north alignment*, *heading alignment* and *no alignment*. The alignment will be classified as heading aligned, if the sketch is aligned to the imagined viewing direction of the participant or the addressee, e.g. from the origin of the route to the destination. These properties are not exhaustive, as for example the *indication of the cardinal direction* or the *indications of the route* to follow might also be considered in graphical description.

3.3 Experiment

3.3.1 Subjects

A total of 30 people (13 male, 17 female) participated in the experiment. All participants were native German speakers between 20 and 30 years (M = 24.07, SD = 2.36). The participants were required to have lived in the city of Münster for at least 6 months. This was to ensure the familiarity of the participants with the layout of the city. Most of the participants were students from various disciplines. They received €10 allowance for their participation.

3.3.2 Design and Procedure

The experiment setup was a simple experiment room where only the participant and the experimenter were present. The experiment was designed to last for approximately 1 h. After signing a consent form, participants were handed out the experiment material, consisting of three parts. Part 1 asked the participants some general questions regarding their familiarity with the city, in Part 2 participants gave in total six spatial descriptions and in Part 3 they answered the Questionnaire Spatial Strategies (Münzer and Hölscher 2011). During the experiment, there was no further interaction between participants and experimenter, but the participants were provided with all necessary information and questions by the experiment material.

The order of the questions for the spatial descriptions in the second part was randomized. For each spatial description the participants were provided with a short explanation, some context, and the task. There were no further restrictions to the tasks, except an approximate time of 5 min per task was assigned, which, however, was not monitored. For the textual and the graphical descriptions the participants were provided with one plain page each, but they had access to further pages to extend their descriptions.

For the *place descriptions* participants were asked to give a description, where they were asked to imagine standing at a certain location in the city and describing their own location in case of an emergency. It was expected that the participants would give spatial descriptions that are more or less restricted to the vista space and are centered to the particular location of the person and would not have a large two-dimensional extend.

For the *route descriptions* the participants were asked to provide a route description between two prominent global landmarks in the city Münster to a cyclist. The bike is one of the main means of transport in Münster and the infrastructure is reasonably well developed. Geographically, the two landmarks are on opposing sides of the inner city, which itself is surrounded by a promenade. The route descriptions were expected to have a linear structure and it was expected that the participants would not only include landmarks along the route but also spatial information distant to the route that support the global orientation (Schwering et al. 2013).

For the *region descriptions* participants were asked to provide a spatial description of the inner city of Münster to an unfamiliar person. In contrast to the place and route descriptions, it was expected that the region descriptions would have a larger two-dimensional extend but no linear structure.

4 Results

4.1 Perspective in Textual Descriptions

For the textual route descriptions participants applied the route perspective consistently (Table 2). More diverse are the results for the place and the region descriptions. Instead of using one perspective consistently, participants used different perspectives in the textual descriptions and often switched perspectives.

In place descriptions almost half of the people chose to describe the place within the route perspective. However, these descriptions in the route perspectives are

Table 2 Perspectives chosen by participants in textual descriptions

	Gaze	Route	Survey	Mixed
Place	5 (17%)	14 (47%)	4 (13%)	7 (23%)
Route	0	30 (100%)	0	0
Region	0	11 (37%)	12 (40%)	7 (23%)

different from the route descriptions because route descriptions generally describe just one route from an origin to a destination. The place descriptions in the route perspective, however, describe the routes from some arbitrary user-chosen origin to the destination of the particular imaginary location of the user. Moreover, the participants described in average 1.36 (SD = 0.74) routes instead of just one route. Considering the seven place descriptions with mixed perspectives, it was found that six of these descriptions started within the survey perspective and then switched towards the route perspective. One of the descriptions started within the gaze perspective and then switched to the route perspective. None of the perspectives changed from the route perspective towards other perspectives.

For the region descriptions participants used either route, survey or mixed perspectives. Approximately a third of the people applied the route perspective and a third of the people applied the survey perspective. The remaining participants mixed route and survey perspectives. For the mixed perspective it was again noted that six in seven participants started the description within the survey perspective and then switched to the route perspective. One participant switched from the survey perspective to the route perspective and gave a short summary in the survey description again at the end of their description. As in place descriptions, the textual region descriptions in the route perspective were different from the route descriptions. The participants described the region through several routes. The descriptions consisted in average of 4.36 (SD = 2.54) branches[3] and 0.73 (SD = 0.79) circuits.[4]

4.2 *Perspective in Graphical Descriptions*

The analysis of the *referent* property in the graphical descriptions shows clear results for the route and the region descriptions. As can be seen in Table 3, for all route descriptions the referent is classified to be the above mentioned *route* referent and for all region descriptions the referent is classified to be the above mentioned *region* referent. For the place descriptions the results are diverse: for the majority of descriptions the referent is classified as *person* (73%), however, for 13% of the descriptions the referent is classified as route or region. For the same amount of descriptions it can not be clearly distinguished between the person and region

[3] A *branch* is defined here as an alternative path through the region. A description with two branches thereby consists of two integrated route descriptions.

[4] A *circuit* is defined here as a route description through a region with the destination equals the origin.

Table 3 Referent property in graphical descriptions

	Person	Route	Region	Ambiguous
Place	22 (73%)	1 (3%)	3 (10%)	4 (13%)
Route	0	30 (100%)	0	0
Region	0	0	30 (100%)	0

Table 4 Alignment property in graphical descriptions

	Heading	North	Not aligned
Place	0	17 (57%)	13 (43%)
Route	17 (57%)	6 (20%)	7 (23%)
Region	0	23 (77%)	7 (23%)

referent. Three of the place descriptions (10%) for which the referent was unambiguously classified as person, the location of the person was not depicted in the center of the sketch, but at the bottom or at the top. However, all the other objects in the sketch were closely aligned to this location.

The *alignment* property was more diverse than the referent property. For the place descriptions slightly more than half of the descriptions were north aligned (57%), but almost the same number of descriptions (13 out of 30) were not north aligned. Slightly more than half of the route descriptions (57%) were aligned to the users heading direction. For the majority of these descriptions (15 out of 17) the origin of the route is depicted at the bottom and the destination at the top of the sketch. For the remaining two descriptions the depiction of origin and destination are swapped. The route descriptions that are not aligned to the users heading direction are split between north alignment (6 out of 30) and no alignment (7 out of 30). The most significant results regarding the alignment property are obtained for the graphical region description, where more than 75% of the descriptions are aligned to north and less than 25% of the descriptions are not aligned to north (Table 4).

The *viewpoint* of all graphical descriptions was external to the scene. However, in total 10 descriptions contained objects that were not sketched according to this viewpoint, but were horizontally displaced: 6 (20%) place descriptions, 1 (3%) route description, and 3 (10%) region descriptions. Furthermore, it was found that only one out of 90 graphical descriptions in total contains an indication of cardinal directions, although almost 50% of all descriptions are aligned to north. Moreover, 80% of the graphical route descriptions contained indications of the route to follow, e.g. in the form of arrows.

5 Discussion

5.1 Perspective in Textual Descriptions

Taylor and Tversky (1996) investigated the perspectives of textual spatial descriptions and presented three ways of textually describing space with different properties (see Table 1). In this paper it was expected to find all three perspectives of textual spatial descriptions applied by the participants of the experiment. Moreover, it was expected that there would be a connection of the place descriptions towards the gaze perspective, of the route perspectives towards the route descriptions and of the region perspectives towards the survey perspectives. Table 2 shows that all three perspectives are applied in the textual descriptions of the experiment. The gaze perspective is exclusively used for place descriptions, however, is not the predominant choice of perspective for the place descriptions. 40% of the region descriptions implement exclusively the survey description. The other 60% of the region descriptions are either found to be in the route perspective or a mix of survey and route perspective. The most prominent perspective is the route perspective as a large degree of participants applied this perspective. Moreover, a considerable degree of participants who started the descriptions within the survey or the gaze perspectives in the place and region descriptions switched towards the route perspective.

In their study, Taylor and Tversky investigated the perspective for each landmark in textual region descriptions and found that people often switch perspectives more than once (Taylor and Tversky 1996). In contrast to that, only one description in the experiment presented in this paper shows two switches of perspectives, whereas all other descriptions, which were evaluated as mixed perspective, show only one switch. Besides a switch from the gaze perspective towards the route perspective in one description, the switches are exclusively from the survey perspective towards the route perspective. An investigation of the features that are described in the different perspectives, as Taylor and Tversky did, is not performed. However, qualitatively looking at the descriptions permits the assertion that participants start to give an overview of the environment or to locate one particular point in the environment within the survey perspective before they switch towards the route perspective. This suggestion requires further qualitative or even quantitative investigations in future work.

Another difference to the experiment of Taylor and Tversky is that participants in this study are provided with a context that relates to real world situations. Participants had to give the descriptions from memory without explicitly studying a map in advance. The way of studying perspectives in descriptions of fictitious environments that are learned from a sketch, like in the experiment of Taylor and Tversky, is most likely to be influenced by the sketch, by the perspective of the sketch, and by the mental ability of the individual to take a viewpoint within a scene that was exclusively learned from a map (Taylor and Tversky 1996).

The main question that has to be asked at this point is why people predominantly use the route perspective, even to a large degree for place and region description, and why most switches within the mixed perspective are towards the route perspective. Again it can be referred to Richter and Winter who stated that textual descriptions will have less impact on route descriptions, as they have a linear structure (Richter and Winter 2014). It is suggested to extend this statement to "textual descriptions have less impact on spatial descriptions in the route perspective, as they have a linear structure and do not require further linearization." The cognitive costs that are required for the linearization might be low enough for people to predominantly use the route perspective or switch towards the route perspective instead of consistently using the survey or gaze perspective for the description.

5.2 *Perspective in Graphical Descriptions*

Regarding the graphical descriptions a set of properties to analyze the perspectives in the sketches was presented in the methods section and results were presented in the previous section, respectively. For the *referent* property, it was expected that there would be a preference of the participants to apply the person referent to the place descriptions, the route referent to the route descriptions and the region referent to the region descriptions. The results in Table 3 meet these expectations, however, the result for the place descriptions are not as clear as for the route and the region descriptions. The four ambiguous descriptions, where the referent could not be clearly classified between the person and the region, can be explained by the objects in the sketches. The objects covered an area that was not necessarily centered to the location of the person only, so that the referent might already be considered to be the region. However, the location of the person was explicitly depicted in the sketch, which suggests the referent to be a person instead of a region.

For the *alignment* property the results show that, although not exclusively, only route descriptions are aligned to the heading of the person, whereas place and region descriptions show only differences with respect to the north alignment. The other way around, a heading alignment might clearly identify a description as a route description. Noting that only one in 90 descriptions contains a indication of the cardinal direction, whereas almost 50% of all descriptions are aligned to north, shows that a significant amount of people apply the cardinal directions for the alignment of the sketches. In contrast to the textual mode, an explicit indication of the cardinal directions within the graphical mode is not mandatory. Moreover, participant did not consider the indication of the cardinal directions as important, which contrasts the indication of the routes in the route descriptions. An indication of the route is as well not mandatory, which is affirmed by the 20% of route descriptions that do not contain indications, however the majority of participants must have considered the indication of the route to follow as important for the graphical route description.

Table 5 Criteria for a categorization of graphical descriptions (LRFB (HF) = left, right, front, back (head, feet); NSEW = north, south, east, west)

Properties	Description perspectives		
	Place	Route	Region
Viewpoint	Fixed, external	Fixed, external	Fixed, external
Frame of reference	Extrinsic	Extrinsic	Extrinsic
Referent	Person	Route	Region
Alignment	None or north	Heading	None or north
Indication	Persons location	Route to follow	

Considering the properties that have been outlined for the graphical descriptions, a distinction between three description perspectives in graphical descriptions might be suggested here: (1) place perspective, (2) route perspective, and (3) region perspective. These perspectives are clearly related to the three scales of spatial descriptions and shall constitute an equivalent of the graphical descriptions to the suggested perspectives for textual descriptions as shown in Table 1.

The description within the *place perspective* might mainly be identified by the referent classified as person. The indication of the person's location in the sketch and a small number of objects that are centered to the location of the person distinguished it from the region perspective. A *route perspective* for graphical descriptions might be classified by the referent in the sketch, which is the route, and the alignment of the descriptions from the origin to the destination. Moreover, an indication of the route in the descriptions might confirm the classification of the route perspective. A *region perspective* for graphical descriptions, in contrast, might be identified by a referent that is regarded to be the region and a preference of an alignment to north. However, the north alignment itself does not explicitly related to one of the perspectives. In Table 5 the suggested description perspectives and their properties are listed.

5.3 General Discussion

In general, there were differences found between the properties of different spatial descriptions and it is assumed that there exist different "perspectives" in both textual and graphical descriptions. For the textual descriptions the perspectives that were presented by Taylor and Tversky were used to classify the descriptions and it was found that participants chose different perspectives for the three different scales of spatial descriptions and even switched perspectives. For the graphical descriptions it was shown that there are differences between the description properties and it was suggested to summarize these properties to different perspectives of graphical spatial descriptions. These perspectives differ between the three scales of spatial descriptions, however not all graphical descriptions can clearly be classified towards one perspective.

Considering the two modes of externalizations of mental spatial representations, the textual and graphical descriptions are hardly comparable, because they are different in their underlying structure and their properties. On the one hand textual descriptions might differ with respect to their viewpoint, whereas graphical descriptions do usually not. On the other hand graphical descriptions might have a clear alignment, whereas the alignment property does not apply to the textual descriptions as it applies to the graphical descriptions. Within both modes of communication a tendency towards different perspective was shown and discussed. Moreover, the perspectives seem to be related to the scales of spatial descriptions. Therefore it can be said that there are differences in externalizations of mental spatial representations, both, between the three scales of spatial descriptions and between the two modes of communication.

6 Conclusion

In this paper it was assumed that there are three different scales of spatial descriptions and that people mainly externalize mental spatial representation within the textual and the graphical communication mode. It was proposed that there are differences in the perspectives of spatial descriptions, that are mainly induced by the structural differences of the communication modes and not by the different scales of spatial descriptions. However, it was found that in both modes of spatial descriptions there are differences between the descriptions of the different scales of spatial descriptions. Moreover, the descriptions of the two modes of externalization show structural differences with respect to their properties and perspectives.

Overall, these finding do only relate to one part of human communication about space, which is the externalization of mental spatial relations. This, however, does not allow any inferences about how people receive spatial information. The best and most natural way of people to externalize spatial information might not necessarily be the best way to receive and understand spatial information. However, as already mentioned above, the knowledge about the naturally preferred way of humans to communicate spatial information across context and scale is applicable to the fields of Human-Computer Interaction and Volunteered Geographic Information. Future work will have to further investigate on the one hand the proposed perspectives of graphical descriptions and on the other hand how people naturally receive, process and understand spatial descriptions.

References

Anacta VJ, Schwering A, Li R (2014) Determining hierarchy of landmarks in spatial descriptions. In: Eighth international conference on geographic information science, Vienna, Austria

Bryant DJ, Tversky B (1999) Mental representations of perspective and spatial relations from diagrams and models. J Exp Psychol Learn Mem Cogn 25(1):137–156

Buhler K (1982) The deictic field of language and deictic words. In: Jarvella RJ, Klein W (eds) Speech, place, and action. Wiley, New York, pp 9–30

Carlson-Radvansky LA, Irwin DE (1994) Reference frame activation during spatial term assignment. J Mem Lang 33(5):646–671

Ehrich V, Koster C (1983) Discourse organization and sentence form: The structure of room descriptions in Dutch. Discourse Process 6(2):169–195

Emmorey K, Tversky B, Taylor HA (2000) Using space to describe space: Perspective in speech, sign, and gesture. Spat Cogn Comput 2(3):157–180

Freundschuh SM, Egenhofer MJ (1997) Human conceptions of spaces: implications for geographic information systems. Trans GIS 2(4):361–375

Galea LAM, Kimura D (1993) Sex differences in route-learning. Pers Individ Differ 14(1):53–65

Hund AM, Schmettow M, Noordzij ML (2012) The impact of culture and recipient perspective on direction giving in the service of wayfinding. J Environ Psychol 32(4):327–336

Kato Y, Takeuchi Y (2003) Individual differences in wayfinding strategies. J Environ Psychol 23 (2):171–188

Lawton CA (1996) Strategies for indoor wayfinding: The role of orientation. J Environ Psychol 16 (2):137–145

Lawton CA, Kallai J (2002) Gender differences in wayfinding strategies and anxiety about wayfinding: a cross-cultural comparison. Sex Roles 47(9):389–401

Levelt WJM (1982) Cognitive styles in the use of spatial direction terms. In: Jarvella RJ, Klein W (eds) Speech, place, and action. Wiley, Chichester, United Kingdom, pp 251–268

Levelt WJM (1984) Some perceptual limitations on talking about space. In: van Doorn AJ, van de Grind WA, Koenderink JJ (eds) Limits in perception. VNU Science Press, Utrecht, The Netherlands, pp 323–358

Levelt WJM (1989) Speaking: from intention to articulation. The MIT Press, Cambridge

Levinson SC (1996) Frames of reference and Molyneux's question: crosslinguistic evidence. In: Bloom P et al (eds) Language and space. The MIT Press, Cambridge, pp 109–156

Montello DR (1993) Scale and multiple psychologies of space. In: Frank AU, Campari I (eds) Spatial information theory: a theoretical basis for GIS, Proceedings COSIT '93. Lecture notes in computer science. Springer, Berlin, pp 312–321

Münzer S, Hölscher C (2011) Entwicklung und Validierung eines Fragebogens zu räumlichen Strategien. Diagnostica 57(3):111–125

Pazzaglia F, Beni R De (2001) Strategies of processing spatial information in survey and landmark-centred individuals. Eur J Cogn Psychol 13(4):493–508

Richter K-F, Winter S (2014) Landmarks: GIScience for intelligent services. Springer

Schwering A, Li R, Anacta VJ (2013) Orientation information in different forms of route instructions. In: Proceedings of the 16th AGILE conference on geographic information science, Leuven, Belgium

Shanon B (1979) Where questions. In: Proceedings of the 17th annual meeting on association for computational linguistics. Association for Computational Linguistics, La Jolla, California, pp 73–75

Sholl MJ et al (2000) The relation of sex and sense of direction to spatial orientation in an unfamiliar environment. J Environ Psychol 20(1):17–28

Taylor HA, Tversky B (1992a) Descriptions and depictions of environments. Mem Cogn 20 (5):483–496

Taylor HA, Tversky B (1992b) Spatial mental models derived from survey and route descriptions. J Mem Lang 31(2):261–292

Taylor HA, Tversky B (1996) Perspective in spatial descriptions. J Mem Lang 35(3):371–391

Tversky B, Lee PU, Mainwaring S (1999) Why do speakers mix perspectives? Spat Cogn Comput 1(4):399–412

Vasardani M et al (2013) From descriptions to depictions: a conceptual framework. In: Tenbrink T et al (eds) Spatial information theory: 11th international conference, COSIT 2013. Springer, pp 299–319

Personal Dimensions of Landmarks

Eva Nuhn and Sabine Timpf

Abstract Landmarks are crucial elements of route instructions given by humans. The currently accepted qualification of an object as a landmark is dependent on spatial dimensions, i.e. visual, semantic and structural dimensions. However, even if an object qualifies as a landmark because of its spatial dimensions, its selection by a traveller as wayfinding aid is further influenced by the knowledge and experience of the traveller. A geographical object that is salient for one person may have no importance at all for another person. Thus, there is a need to incorporate more personal dimensions into a comprehensive model of landmarks. We propose such a multidimensional model for landmarks for pedestrians. The main contribution of this paper is the definition of the personal dimensions of landmarks as basis for a multidimensional model. The following personal dimensions are introduced and debated: spatial knowledge, interests, goals and background. Further, attributes and attribute values describing the personal dimensions are identified and data collection is discussed.

Keywords Landmarks · Pedestrian navigation · Wayfinding · Spatial and personal dimensions

1 Introduction

Assume that you just arrived at the main station of a city unfamiliar to you and you would like to go to your hotel. Not knowing the route, you ask someone on the street for help. What you will most likely receive is an instruction that includes descriptive elements, visual cues that you will encounter along the way and special

E. Nuhn (✉) · S. Timpf
Geoinformatics Group, University of Augsburg, Alter Postweg 118,
86159 Augsburg, Germany
e-mail: eva.nuhn@geo.uni-augsburg.de

S. Timpf
e-mail: sabine.timpf@geo.uni-augsburg.de

A. Bregt et al. (eds.), *Societal Geo-innovation*, Lecture Notes
in Geoinformation and Cartography, DOI 10.1007/978-3-319-56759-4_8

objects, i.e. landmarks (Daniel and Denis 1998). Landmarks are references for wayfinding, which are external to the observer (Lynch 1960). This means that any distinct object or geographic feature that is noticed, remembered and stands out from the background may serve as a landmark (Presson and Montello 1988; Caduff and Timpf 2008). Which route and landmarks are chosen for route instructions is dependent on the mode of travel, the desired route characteristics and the presumed level of knowledge of the recipient (Lovelace et al. 1999). In our example the person you asked did not notice that you were not familiar with the city and thus s/he refers to a landmark called the "corncob". Since you do not know this landmark you are not able to navigate using it. If you would inform the informant of this fact s/he would either refer to another landmark or change the description to a more general one (e.g. "high-rise Hotel Tower"). While a human informant may be able to tailor route instruction to your needs, current landmark generation algorithms are not yet there. Informants base the decisions on which route information to give to you on perceived personal dimensions, e.g. your probable familiarity with the environment, your goal for navigation or your perceived background.

The incorporation of such personal dimensions into route generation algorithms for pedestrians is a challenging task. While much research has been published about the spatial dimensions of landmarks (see Sect. 2: Related Work), the personal dimensions have been neglected so far. There is no consideration of the personal dimensions of landmarks and no investigation of how to integrate this information directly into the routing algorithm.

To tackle these problems we propose a multidimensional model for personalized landmarks for pedestrians (Nuhn and Timpf 2016). The model considers three different inputs (see Fig. 1): the established spatial dimensions of landmark candidates (see Sect. 2.1), start and destination of the route and, in addition, personal dimensions (see Sect. 3). The model allows for the assessment of the inputs and the determination of their influence on the landmarkness or salience (Caduff and Timpf 2008) of a landmark candidate. The model results in a measure of the personal salience of a landmark candidate dependent on the personal dimensions. This measure can then be integrated in the generation of a route between the defined start and destination. The result of the algorithm is a personalized route for a pedestrian with personal landmarks.

The personal dimensions of a landmark are a fundamental part of our model for the assessment of the personal salience of a landmark candidate. The remainder of this paper is organized as follows. After a review of existing methods to determine landmarks based on spatial and personal dimensions, we identify personal dimensions for our model and describe them in detail. Subsequently, we discuss possible data acquisition methods for the personal dimensions, and finally, we conclude with an outlook and future work.

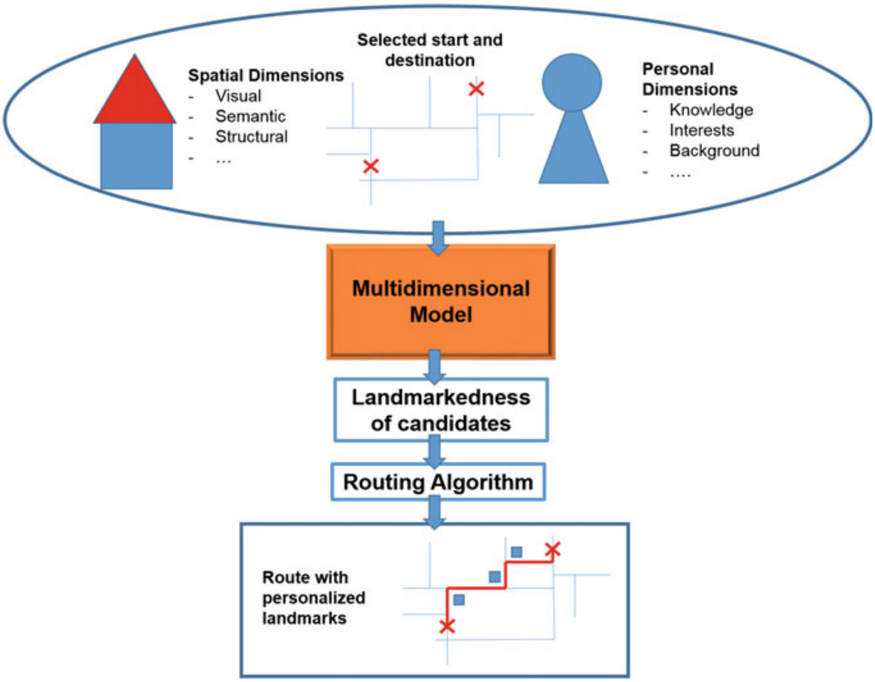


Fig. 1 Model configuration

2 Related Work

In this section methods for determining landmarks based on spatial and personal dimensions respectively are investigated.

2.1 Determination of Landmarks Based on Spatial Dimensions

In existing research, landmarks are characterized according to their spatial dimensions. The most influential characterization was proposed by Sorrows and Hirtle (1999). Their framework defines three spatial dimensions of a landmark:

1. Visual Prominence, which describes the visual characteristics of a landmark (e.g. façade area, shape or colour),
2. Structural Significance—which gives information about the location of a landmark within the spatial environment (e.g. location directly at the street) and
3. Semantic Salience, which focuses on the meaning of a landmark (e.g. through explicit marks on a building).

This classification is not mutually exclusive. Normally, a geographic feature shows more than one characteristic that classifies it as a landmark and determines its salience as a landmark.

Raubal and Winter (2002) proposed the first approach towards a formal model of landmark salience of a feature. The authors suggest measures for each attribute of the spatial dimensions of built-up and network features. This work was further extended and tested by Nothegger et al. (2004). The results showed that the model is an applicable assessment of landmark salience. Further, Winter (2003) extended Raubal and Winter's model by including the advance visibility of an object. Because the weak point of these approaches is the number of available data sources, Elias (2003) proposed to use existing spatial databases. She focused on buildings as landmark candidates and used spatial attributes from topographic and cadastral data sets to automatically extract landmarks using data mining methods. Another approach using OSM (Open Street map) data was proposed by Nuhn et al. (2012). The study focuses on buildings derived from OSM and 3D city model data and assesses the suitability of an object as a landmark with the help of a landmark index based on spatial dimensions.

Apart from such landmark identification approaches, landmark integration approaches also exist, which focus on the determination of landmarks for a specific route. In these approaches the focus again is on spatial parameters (e.g. position along the route, uniqueness in a given environment or visibility from the route). Klippel and Winter (2005) proposed a first approach to automatically integrating salient objects dependent on a specific route into routing instructions. The model takes into account advance visibility, the configuration of the street network and the route along the network. Another approach by Richter Richter and Klippel (2006), Richter (2007) used qualitative descriptions of landmarks' locations. Ordering information to determine a landmark's relative location is used and demonstrated on different landmark geometries (point landmarks as well as linear and areal landmarks) and different spatial situations. Duckham et al. (2010) addressed the incorporation of cognitively salient landmarks in computer-generated routing instructions. They developed a weighting system that assigns weights to points of interests (POIs) and an algorithm for annotating routing instructions with landmarks based on the weighting system. A first approach investigating the integration of landmark information directly into the routing algorithm was proposed by Caduff and Timpf (2005). The *Landmark-Spider-Algorithm* calculates the clearest route in terms of spatial references and uses selected landmarks to describe the route. Their model selects spatial cues based on distance and orientation of the traveller with respect to the landmark and the salience of spatial objects. The results of this algorithm are presented in a spatio-analogical way, which supports wayfinding decisions. However, they also did not consider the personal dimensions of landmarks within their work.

2.2 Determination of Landmarks Based on Personal Dimensions

The landmark salience of a feature is not only dependent on spatial but also on personal dimensions. An object that is salient for one person may have no importance at all for another person. This is because people add salience to the objects they perceive due to their knowledge, background and interests. That means, the landmarkness of an object is also dependent on personal dimensions such as mobility, gender, age, education, hometown and other socio-demographic properties (Winter et al. 2012). There is a large body of research about the adaptation of the content and appearance of a map depending on a user's preferences (e.g. Sarjakoski et al. 2007; Sarjakoski and Sarjakoski 2008; Reichenbacher 2007; Wiebrock 2011). The knowledge-based system from Sarjakoski et al. (2007) for example considers aspects such as the use case for which the map is needed (e.g. cycling, emergency or experts), the time (e.g. summer or winter), the device on which the map is displayed or the user's age group (e.g. a teenager or elderly person). However, no other parameters are discussed about traveller's knowledge or experience, and especially no landmarks are considered.

There is only little work that deals with the idea that salience of a landmark is not the same for every person. Burnett et al. (2001) were the first to recognize that travellers familiar with the environment choose different landmarks for route instructions than those who are unfamiliar with the environment. More recent studies confirm this assumption and show that persons that are familiar with a specific environment prefer objects with a special personal meaning as landmarks (Quesnot and Roche 2015). Meng (2005) showed that the usability of an egocentric mobile map is not only dependent on objective but also on subjective parameters. Subjective parameters concern, amongst others, the user's emotion (e.g. joyfulness or irritation) during map interaction.

Balaban et al. (2014) are also focusing on emotions. They consider the mood condition (positive, negative, neutral) and show that negatively laden landmarks are better recollected than positively laden landmarks and that positive-laden landmarks are better remembered than neutral landmarks. The emotion dependent selection of landmarks is also a kind of personalization, but neglects other personal dimensions. Götze and Boye (2016) recently proposed to learn individual salience models for landmarks used in route instructions. A mathematical model of salience is automatically derived directly from route instructions given by humans. Each landmark that a person can refer to in a given situation is modelled as a vector of features. The salience associated with each landmark is calculated as a weighted sum of these features. Because the feature vectors consider only spatial attributes (distance and angle to a landmark as well as name and type extracted from OSM data) this approach is restricted to the spatial dimensions and lacks an answer on how to consider the personal dimensions of landmarks.

3 Personal Dimensions

The main contribution of this paper is the definition of personal dimensions of landmarks for the inclusion within a multidimensional model. We identify personal dimensions and their attributes by taking into account five dimensions when viewing a person as an individual (Brusilovsky and Millán 2007):

- Personal knowledge,
- Personal interests,
- Personal goals,
- Personal background and
- Individual traits.

For the provision of personal landmarks probably the most important dimension to consider is "personal knowledge". In our use case this refers to the *spatial knowledge* of the traveller. Highly personalized landmarks such as the "corncob" in the abovementioned example can be used only if a traveller has spatial knowledge of an environment. Another crucial dimension is the consideration of *interests* of a traveller, because one's level of interest could enhance memory for some information (McGillivray et al. 2015). The *goal* of the traveller is the most changeable dimension from the above-mentioned ones and has an impact on the amount of instructions that are required and on the distribution of landmarks. The personal *background* is a significant dimension because it influences the way objects are recognized and perceived. *Individual traits* are features that define a person as an individual. For example personality traits, cognitive styles or factors (Brusilovsky and Millán 2007). Unlike the other dimensions, individual traits can only be determined through especially designed psychological tests. For this reason we currently do not consider individual traits of a traveller. The other dimensions are discussed in detail in the following sections.

3.1 Spatial Knowledge

For the provision of personal landmarks the most important dimension to consider is the spatial knowledge of the traveller. Spatial knowledge is commonly divided into three levels with interdependent contents (Siegel and White 1975; Thorndyke 1981; Herrmann et al. 1998; Golledge 1999). Attributes for the multi-dimensional model for dimension Spatial Knowledge comprise *no knowledge*, *landmark*, *route* and *survey knowledge*. The attributes of spatial knowledge and how these attributes may be measured can be seen in Table 1. A Boolean value is assigned to each attribute: 'True', if the attribute holds true and 'False' otherwise. At least one of the attribute values must hold 'True'. The attributes are not mutually exclusive; how they influence each other is explained below.

Table 1 Attributes for spatial knowledge of the traveller

	Example	Measurement
No knowledge	$K_n = F$	$K_n \in \{T, F\}$
Landmark knowledge	$K_l = T$	$K_l \in \{T, F\}$
Route knowledge	$K_r = T$	$K_r \in \{T, F\}$
Survey knowledge	$K_s = F$	$K_s \in \{T, F\}$

Landmark knowledge. During spatial knowledge acquisition landmarks are the first salient geographic features, which are available in no particular order on a cognitive map (Couclelis et al. 1987). If a traveller knows landmarks within the environment for navigating, these personal landmarks can be used within the route instructions, to link known with unknown elements along the route. If only landmark knowledge is available the other attributes *no*, *route* and *survey knowledge* are 'False'.

Route knowledge. Route knowledge is needed to navigate from a starting point to a destination. On this level of spatial knowledge, the environment is represented as a set of routes. Route knowledge includes the knowledge of a sequence of landmarks along a route and the knowledge of how to get from one landmark to another. The availability of route knowledge also has an influence on the granularity of the route instructions (Tenbrink and Winter 2009). The instructions can be coarser, merely enriched with some personal landmarks, or more detailed in areas where no knowledge is available. Route knowledge implies landmark knowledge and therefore the attribute value for *landmark knowledge* is also set to 'True'. In contrast, *survey* and *no knowledge* are set to 'False'.

Survey knowledge. Survey knowledge is the result of the mental integration of two or more routes, in contrast to route knowledge, which is related to only one route. According to Schmauks (1998) survey knowledge is obtained when the global structure of an environment is known, so that new routes can be generated and shortcuts can be detected. The survey knowledge is usually generated from route knowledge through integration into a "mental map" (Herrmann et al. 1998). Survey knowledge implies that the traveller is familiar with an environment. Quesnot and Roche (2015) showed that persons that are familiar with a specific environment prefer objects with personal meaning as landmarks. Such objects have landmarkness solely because of their semantics, e.g. "my home" or "my work place" (Richter and Winter 2014). The higher the degree of familiarity the higher is the possible degree of personalization of a landmark. If a traveller has *survey knowledge*, then *route* and *landmark knowledge* are also set to 'True'. *No knowledge* is set to 'False".

No spatial knowledge. If someone has never been to the environment to navigate, he has no spatial knowledge at all. Quesnot and Roche (2015) showed that people unfamiliar with a specific environment prefer landmarks because of their visual or structural salience. In those cases, highly visible landmarks located at strategic points of the route should be provided. If the attribute value of *no knowledge* is 'True' then all the other attribute values are 'False'.

The personal spatial knowledge of a traveller can change over time. It can either increase (through learning) or decrease (forgetting) (Brusilovsky and Millán 2007). The familiarity of a traveller with an environment can increase when the traveller gets to know the environment and gains some route or survey knowledge. The familiarity of a traveller can decrease when the traveller has not visited the familiar environment for a longer time and objects did change during this time period (e.g. explicit marks on geographic features, new buildings, disappeared objects, …).

3.2 Personal Interests

To support wayfinding of a traveller it is helpful to include features (i.e. landmarks) in a map, which match the traveller's interest (Reichenbacher 2007). The level of interest can enhance memory for some information (McGillivray et al. 2015). Some studies differentiate between *interests* and *preferences*. According to Weißenberg et al. (2004) interests are static and application specific e.g. interest in shopping, sightseeing or entertainment. In contrast, preferences are situation dependent and can be divided into simple and complex preferences (Weißenberg et al. 2004). For example, consider a traveller not much interested in arts. In a foreign country and on holidays s/he prefers artificial monuments to get to know the culture—this changes the simple preference to a complex one. This change is dependent on external factors not directly related to the traveller (Weißenberg et al. 2004). In our work we focus on interests that the traveller can specify and propose a method to capture and update the interests if needed. Situation dependence e.g. if the traveller is travelling (which would also imply different interests than in daily life) are not treated within this work.

Personal interests. How personal interests are measured can be seen in Table 2. Personal interests can be manifold and there are many objects in an urban environment interesting to different people. People interested in nature may prefer landmarks like rivers or trees, while people interested in gastronomy would focus on restaurants, bars and cafés. We propose to apply scalar models to measure personal interests, which estimates the traveller's interest in a subject by a single value on a specific scale. A distinction can be drawn between quantitative methods (e.g. numbers from 0 to 5) or qualitative approaches (e.g. good, average, poor, none) (Brusilovsky and Millán 2007). Since a traveller should specify his interests

Table 2 Attributes for personal interests

	Example	Measurement
Personal interests	Interest in arts = 2 Interest in nature = 5 Interest in gastronomy = 4	I_i (i = 1…n), n = number of interests $I_i \in \{1, 2, 3, 4, 5\}$ Scale of interest: 1 (no), 2 (little), 3 (moderate), 4 (strong), 5 (very strong)

himself, a qualitative approach seems more suitable. It is more intuitive to rate an interest with qualitative values than with quantitative methods.

3.3 *Personal Goals*

Landmarks qualify as attractive if they are helpful aids to achieve a goal. The landmarkness of a geographic object itself is not dependent on the traveller's goal because it depends much more on the traveller's interests or knowledge of the environment. Nonetheless it is important to know the traveller's intention, because it might make a difference in the number and distribution of landmarks. In human wayfinding three goals are distinguished (Golledge 1999): Traveling with the goal of

- reaching a familiar destination,
- exploring the environment and
- finding the way to a novel unknown destination.

Wiener et al. (2009) divide the reason for traveling through space into traveling with a specific spatial goal and traveling with a non-spatial goal. Table 3 shows the attributes for personal goals considered within our multidimensional model and describes how these properties are measured. We propose to assign a Boolean value for each goal: 'True', if the goal holds true and 'False' otherwise. At least one of the attribute values must hold 'True'. The goals are mutually exclusive, that means one holds 'True' and the others are 'False'.

Known goal. The focus of traveling with a specific spatial goal is primarily on reaching a particular location, e.g. a predefined destination that is known. This is a very common case: e.g. commuting between home and work place is an example for traveling to a known place (Golledge 1999). In this case the traveller needs no landmarks around the destination. In addition, landmarks at important decision points and for confirmation along the route can be provided, dependent on his spatial knowledge of the route.

New goal. This type of wayfinding is mostly carried out with different kinds of wayfinding aids e.g. with the help of a routing service. As Michon and Denis (2001) have shown, the frequency with which landmarks are mentioned increases in the vicinity of the destination. Thus, there is a need for more landmarks around the destination when traveling towards a spatial goal that is unknown. Confirmatory landmarks are needed at points where a change in direction is required or along long segments, because it can be assumed that if the goal is unknown the route or at least parts of the route are unknown as well.

Table 3 Attributes for personal goals

	Example	Measurement
Known goal	$G_k = F$	$G_k \in \{T, F\}$
New goal	$G_n = F$	$G_n \in \{T, F\}$
Exploratory travel	$G_e = T$	$G_e \in \{T, F\}$

Exploratory travel. In contrast to traveling with a specific spatial goal in mind, the reason for traveling with a non-spatial goal is for example to explore a new environment (e.g. after moving to a new town, to discover a city or just to walk on the beach (Wiener et al. 2009)). In case of traveling with a non-spatial goal to explore the environment extra landmarks along the route for additional information or to get to know the environment are helpful.

3.4 Personal Background

The personal background of a traveller is mainly described by demographic data, which are "objective facts" (see Kobsa et al. 2001). The personal background is not a strictly personal dimension and it is also possible to relate it to specific user groups. In this case preferences don't have to be specified for each person individually. Important demographic data for the incorporation in our multidimensional model are discussed below.

Gender. The gender of participants is considered because of the known differences regarding spatial cognition between women and men (Coluccia and Louse 2004). Further, a difference between women and men in the importance of structural salience has been found (Quesnot and Roche 2015). As shown in Table 4 a binary value is assigned to this variable, 'm' for male and 'f' for female.

Age. The age of persons was found to be an important attribute in spatial cognition, because of strong differences in orientation abilities (Jansen-Osmann et al. 2007). Age may also have a particular impact on the structural salience of landmarks (Kattenbeck 2016). The attribute value of age is a natural number.

Country of residence. Travellers, not living within the country of the environment to navigate, may be used to environments and objects shaped differently (Kattenbeck 2016). For example, if a German refers to a "telephone box" an Englishman would search for a completely different object than it is known in Germany. The attribute value for place of residence is a Boolean value. 'True' if the place of residence is in the country in which the traveller wants to navigate, 'False' if the traveller is not residing in the same country.

Cultural background. Travellers, who did not grow up within the environment to navigate may be also used to completely different shapes. For example someone who grew up in a small village in Africa has different background compared to somebody who grew up in a modern city (e.g. in central Europe). The attribute

Table 4 Attributes for personal background

	Example	Measurement
Gender	G = m	G € {m, f}
Age	a = 34	a € ℕ
Country of residence	C = T	C € {T, F}
Cultural background	B = T	B € {T, F}
Education	E = Fireman	E (string)

value is also a Boolean value. 'True' if the cultural background is the same as the environment to navigate, 'False' otherwise.

Education. The education of the traveller can influence the way visual and structural dimensions are perceived (Kattenbeck 2016). Consider e.g. surveyors that have a perspective on measuring points or firemen who take special note of hydrants, while others do not notice these spatial objects. The attribute value for education is a string.

4 Data Collection for the Personal Dimensions

In order to provide personalized landmarks the attribute values of the personal dimensions must be collected. This can be accomplished by implicitly observing travellers' interaction or by soliciting explicit direct input from the traveller (Poslad et al. 2001). Explicit methods are the simpler of the two approaches. However, a casual user is not willing to spend much time and effort in such explicit methods to specify personal attributes. In this section we review existing acquisition methods for personal dimensions. We review both, explicit and implicit methods, but keep in mind that implicit methods should be preferred whenever explicit methods get to be too exhaustive and time-consuming.

Spatial knowledge. A way to determine which spatial knowledge a traveller has of an environment to navigate is asking true/false questions. For example, for *survey knowledge*: "I know the area very well and I am able to detect shortcuts and find new routes". In case of affirmation no further questions are needed, because this automatically implies route and survey knowledge and excludes no knowledge. If the question is answered in the negative, further questions are necessary. If a traveller is answering positively to a question asking about route knowledge then s/he can additionally be asked where this route started and ended to learn where the traveller has been before and therefore might be quite familiar with the environment. In case of only landmark knowledge the traveller can define which landmarks s/he already knows (e.g. in Paris the Eiffel Tower or in London the London Bridge), which can then be incorporated into the route instructions.

Capturing spatial knowledge by asking questions could be impractical because users tend to be unwilling to make active personalizations. Instead of defining preferences an automatic capturing method would be helpful. A possible method is to store already navigated routes. If a traveller has to go a route or parts of a route again he already has more knowledge of the environment. Also landmarks that were used in former route instructions could be stored and incorporated into new instructions.

Personal Interests. The simplest forms to acquire the personal interests of a traveller are scalar models, which estimate the traveller's interest in a subject by a single value on some scale. In Sect. 3.2 we stated that a qualitative approach is suitable, because of intuitive rating. However, questions about personal interests can be manifold, if the interest should be captured in the most complete and comprehensive way. There can be a lot of different fields of interests occurring in a city, e.g. gastronomy, architecture or arts (to name just a few) necessitating many questions. A method has to be identified that is not too exhaustive and time-consuming.

An implicit feedback method to obtain personal interests is the usage of sensors. These sensors can be integrated in smartphones and deliver invisible and practically imperceptible information about the traveller's behaviour. Sensors can give information about the traveller's position (via GPS) and heading (via inbuilt compass sensors). This allows for the calculation of a visibility area, the retrieval of associated geographic data (Wolfensberger and Richter 2015), and permits conclusions regarding the traveller's interests.

Another implicit method is the use of a learning system. In a first step the traveller is presented with landmark proposals based on spatial dimensions. After following the route s/he is asked if these landmarks were helpful or not. This can easily be achieved by clicking on a landmark using color-coding (green for helpful landmarks, red for rather unhelpful landmarks). Additionally, the traveller may supply landmarks that s/he would have preferred along the route (again by just clicking on them). This information can be analysed and the system can learn which landmarks best fit the traveller.

Personal goals. One method to capture the personal goal is to allow the traveller to specify it by himself. "Goal" is a dynamic personal dimension and can change from session to session (Brusilovsky and Millán 2007). A way to determine it, as in the case of spatial knowledge, is asking true/false questions. For example for *known goal*: "I am on the way to a familiar destination (e.g. to my working place, to my favourite shop)". The other goals can be queried following the same pattern. Because "personal goal" is a dimension whose attributes are mutually exclusive, at most two questions are necessary. If the first question is 'True' the others are 'False' and only one question is needed. If two questions are 'False' it can be concluded that the third must be 'True'.

Another method to capture the traveller's goals is to provide a possibility to select the goal at the time when the start and the destination of the route are selected. This would be an easy and intuitive method, which requires little time and effort for the traveller.

Personal background. Similar to spatial knowledge the personal background delivers information about the traveller's familiarity with an environment. The background of the traveller is a static variable, which is unchanging during the navigation. Additionally, it is nearly impossible to deduce by sensors or by simply watching the traveller (see also Brusilovsky and Millán 2007). As a result, the personal background of the traveller must be provided explicitly, by entering the required values.

5 Outlook and Future Work

This paper proposes a multidimensional model for landmarks that incorporates spatial and personal dimensions. The main contribution of this paper is the definition of the personal dimensions of landmarks. The personal dimensions *spatial knowledge*, *interests*, *goals* and *background* were defined and debated. Further, attributes and attribute values describing the personal dimensions were identified and possible methods for their acquisition were discussed.

In future work we want to put together the attributes of the personal dimensions to integrate them in a multidimensional model for personalized landmarks. With such a model it will be possible to calculate the landmarkness or salience of a geographic feature, which is dependent on the personal dimensions of landmarks. Particular emphasis during this work will be on the method for the interconnection of the spatial dimensions and the personal dimensions of landmarks within one modelling framework. Therefore existing approaches to model the spatial dimensions of landmarks (e.g. Raubal and Winter 2002) will be used and extended by the personal dimensions.

Another focus will be on further investigating potential methods for collecting data on the personal dimensions. Advantages and disadvantages of different methods need to be investigated. We aim to select an appropriate method for the collection of the attribute values of the personal dimensions.

Finally, we intend to apply our multidimensional model to an actual wayfinding scenario with different travellers, having different personal requirements and needs, to show the applicability and usefulness of our approach.

References

Balaban CZ, Röser F, Hamburger K (2014) The effect of emotions and emotionally laden landmarks on wayfinding. In: Bello P, Guarini M, Mcshane M, Scassellati B (eds) Proceedings of the 36th annual conference of the cognitive science society. Austin, USA, pp 1880–1885

Brusilovsky P, Millán E (2007) User models for adaptive hypermedia and adaptive educational systems. In: Brusilovsky P, Kobsa A, Nejdl W (eds) The adaptive web—methods and strategies of web personalization. Springer, Heidelberg

Burnett G, Smith D, May A (2001) Supporting the navigation task: characteristics of 'good' landmarks. In: Hanson, MA (ed) Contemporary ergonomics. Taylor & Francis

Caduff D, Timpf S (2005) The landmark spider: representing landmark knowledge for wayfinding tasks. In: Barkowsky T, Freksa C, Hegarty M, Lowe R (eds) AAAI spring symposium: reasoning with mental and external diagrams: computational modeling and spatial assistance. AAAI Press, Stanford, pp 30–35

Caduff D, Timpf S (2008) On the assessment of landmark salience for human navigation. Cogn Process 9:249–267

Coluccia E, Louse G (2004) Gender differences in spatial orientation: a review. J Environ Psychol 24:329–340

Couclelis H, Golledge RG, Gale N, Tobler W (1987) Exploring the anchor-point hypothesis of spatial cognition. J Environ Psychol 7:99–122

Daniel M-P, Denis M (1998) Spatial descriptions as navigational aids: a cognitive analysis of route directions. Kognitionswissenschaft 7:45–52

Duckham M, Winter S, Robinson M (2010) Including landmarks in routing instructions. J Locat Based Serv 4:28–52

Elias B (2003) Extracting landmarks with data mining methods. In: Kuhn W, Worboys M, Timpf S (eds) International conference on spatial information theory. Foundations of geographic information science, COSIT 2003, Ittingen, Switzerland. Springer, pp 375–389

Golledge RG (1999) Human wayfinding and cognitive maps. John Hopkins University Press, Baltimore

Götze J, Boye J (2016) Learning landmark salience models from users' route instructions. J Locat Based Serv 10:47–63

Herrmann T, Schweizer K, Janzen G, Katz S (1998) Routen-und Überblickswissen-konzeptuelle Überlegungen. Kognitionswissenschaft 7:145–159

Jansen-Osmann P, Schmid J, Heil M (2007) Spatial knowledge of adults and children in a virtual environment: the role of environmental structure. Euro J Dev Psychol 4:251–272

Kattenbeck M (2016) Empirically measuring salience of objects for use in pedestrian navigation. Dissertation, University of Regensburg

Klippel A, Winter S (2005) Structural salience of landmarks for route directions. In: Cohn AG, Mark DM (eds) International conference on spatial information theory, COSIT 2005, Ellicottville, USA. Springer, pp 347–362

Kobsa A, Koenemann J, Pohl W (2001) Personalised hypermedia presentation techniques for improving online customer relationships. Knowl Eng Rev 16:111–155

Lovelace KL, Hegarty M, Montello DR (1999) Elements of good route directions in familiar and unfamiliar environments. In: Freksa C, Mark DM (eds) International conference on spatial information theory. Cognitive and computational foundations of geographic information science, COSIT '99, Stade, Germany. Springer, pp 65–82

Lynch K (1960) The image of the city. MIT, Boston

McGillivray S, Murayama K, Castel AD (2015) Thirst for knowledge: the effects of curiosity and interest on memory in younger and older adults. Psychol Aging 30:835

Meng L (2005) Egocentric design of map-based mobile services. Cartographic J 42:5–13

Michon P-E, Denis M (2001) When and why are visual landmarks used in giving directions? In: Montello DR (ed) International conference on spatial information theory. Foundations of geographic information science, COSIT 2001, Morro Bay, USA. Springer, pp 292–305

Nothegger C, Winter S, Raubal M (2004) Selection of salient features for route directions. Spat Cogn Comput 4:113–136

Nuhn E, Reinhardt W, Haske B (2012) Generation of landmarks from 3D city models and OSM data. In: Gensel J, Josselin D, Vandenbroucke D (eds) Multidisciplinary research on geographical information in Europe and beyond. Proceedings of the AGILE'2012 international conference on geographic information science. Avignon, France, pp 365/392–369/392

Nuhn E, Timpf S (2016) A multidimensional model for personalized landmarks. In: Gartner G, Huang H (eds) 13th international conference on location based services. Austria, Vienna, pp 4–6

Poslad S, Laamanen H, Malaka R, Nick A, Buckle P, Zipl A (2001) Crumpet: creation of user-friendly mobile services personalised for tourism. 3G 2001. Second international conference on 3G mobile communication technologies, London UK. IET, pp 28–32

Presson CC, Montello DR (1988) Points of reference in spatial cognition: Stalking the elusive landmark. Br J Dev Psychol 6:378–381

Quesnot T, Roche S (2015) Quantifying the significance of semantic landmarks in familiar and unfamiliar environments. In: Fabrikant SI, Raubal M, Bertolotto M, Davies C, Freundschuh S, Bell S (eds) 12th international conference on spatial information theory, COSIT 2015, Santa Fe, USA. Springer, pp 468–489

Raubal M, Winter S (2002) Enriching wayfinding instructions with local landmarks. In: Egenhofer MJ, Mark DM (eds) Geographic information science. Springer, Berlin

Reichenbacher T (2007) The concept of relevance in mobile maps. In: Gartner G, Cartwright W, Peterson MP (eds) Location based services and telecartography. Springer Berlin Heidelberg, Heidelberg

Richter K-F (2007) A uniform handling of different landmark types in route directions. In: Winter S, Duckham M, Kulik L, Kuipers B (eds) International conference on spatial information theory, COSIT 2007, Melbourne, Australia. Springer, pp 373–389

Richter K-F, Klippel A (2006) Before or after: prepositions in spatially constrained systems. In: Barkowsky T, Knauff M, Ligozat G, Montello DR (eds) Spatial cognition V reasoning, action, interaction. Springer

Richter K-F, Winter S (2014) Landmarks. Springer

Schmauks D (1998) Kognitive and semiotische Ressourcen fur die Wegfindung. Kognitionswissenschaft 7:124–128

Sarjakoski LT, Sarjakoski T (2008) User interfaces and adaptive maps. Encyclopedia of GIS. Springer US, Boston, MA

Sarjakoski LT, Koivula T, Sarjakoski T (2007) A knowledge-based map adaptation approach for mobile map services. In: Gartner G, Cartwright W, Peterson MP (eds) Location based services and telecartography. Springer Berlin Heidelberg, Heidelberg

Siegel A, White S (1975) The development of spatial representations of large-scale environments. In: Reese H (ed) Advances in child development and behavior. Academic, London

Sorrows ME, Hirtle SC (1999) The nature of landmarks for real and electronic spaces. In: Freksa C, Mark DM (eds) International conference on spatial information theory. Cognitive and computational foundations of geographic information science, COSIT'99, Stade, Germany. Springer, pp 37–50

Tenbrink T, Winter S (2009) Variable granularity in route directions. Spat Cogn Comput 9:64–93

Thorndyke PW (1981) Spatial cognition and reasoning. In: Harvey JH (ed) Cognition, social behavior, and the environment. Lawrence Erlbaum Associates, Hillsdale, New Jersey

Weiβenberg N, Voisard A, Gartmann R (2004) Using ontologies in personalized mobile applications. Conference on information and knowledge management. ACM, Arlington, USA, pp 2–11

Wiebrock I (2011) Zur kontextbasierten Visualisierung von Geodaten auf Basis von standardisierten Webdiensten. Universität der Bundeswehr München

Wiener JM, Büchner SJ, Hölscher C (2009) Taxonomy of human wayfinding tasks: a knowledge-based approach. Spat Cogn Comput 9:152–165

Winter S (2003) Route adaptive selection of salient features. In: Kuhn W, Worboys M, Timpf S (eds) International conference on spatial information theory. Foundations of geographic information science, COSIT 2003, Ittingen, Switzerland. Springer, pp 113–136

Winter S, Janowicz K, Richter K-F, Vasardani M (2012) Knowledge acquisition about places. SIGSPATIAL Spec 4:20–21

Wolfensberger M, Richter K-F (2015) A mobile application for a user-generated collection of landmarks. In: Gensel J, Tomko M (eds) International symposium on web and wireless geographical information systems, Grenoble, France. Springer, pp 3–19

Personal Activity Centres and Geosocial Data Analysis: Combining Big Data with Small Data

Colin Robertson, Rob Feick, Martin Sykora, Ketan Shankardass and Krystelle Shaughnessy

Abstract Understanding how people move and interact within urban settings has been greatly facilitated by the expansion of personal computing and mobile studies. Geosocial data derived from social media applications have the potential to both document how large segments of urban populations move about and use space, as well as how they interact with their environments. In this paper we examine spatial and temporal clustering of individuals' geosocial messages as a way to derive personal activity centres for a subset of Twitter users in the City of Toronto. We compare the two types of clustering, and for a subset of users, compare to actual self-reported activity centres. Our analysis reveals that home locations were detected within 500 m for up to 53% of users using simple spatial clustering methods based on a sample of 16 users. Work locations were detected within 500 m for 33% of users. Additionally, we find that the broader pattern of geosocial footprints indicated that 35% of users have only one activity centre, 30% have two activity centres, and 14% have three activity centres. Tweets about environment were more likely sent from locations other than work and home, and when not

C. Robertson (✉)
Department of Geography and Environmental Studies, Wilfrid Laurier University, Waterloo, Canada
e-mail: crobertson@wlu.ca

R. Feick
School of Planning, University of Waterloo, Waterloo, Canada
e-mail: robert.feick@uwaterloo.ca

M. Sykora
Centre for Information Management, School of Business and Economics, Loughborough University, Loughborough, UK
e-mail: M.D.Sykora@lboro.ac.uk

K. Shankardass
Department of Health Sciences, Wilfrid Laurier University, Waterloo, Canada
e-mail: kshankardass@wlu.ca

K. Shaughnessy
Department of Psychology, University of Ottawa, Ottawa, Canada
e-mail: Krystelle.Shaughnessy@uottawa.ca

A. Bregt et al. (eds.), *Societal Geo-innovation*, Lecture Notes
in Geoinformation and Cartography, DOI 10.1007/978-3-319-56759-4_9

directed to another user. These findings indicate activity centres defined from Twitter do relate to general spatial activities, but the limited degree of spatial variability on an individual level limits the applications of geosocial footprints for more detailed analyses of movement patterns in the city.

Keywords Geosocial · Personal activity centres · Clustering · Spatial analysis

1 Introduction

The proliferation of Internet and communications technologies (ICTs) and their associated information infrastructures are having transformative impacts on how people use, perceive, and co-develop urban spaces and places. Massive data streams are providing new ways to monitor, deliver, and analyze a variety of urban services (e.g. water and electricity consumption), community resources (e.g. transit and greenspace) and human activities (e.g. financial transactions and commuting flows) (Miller 2010). For researchers, new data streams with embedded geographic coordinates produce digital traces of individuals' interactions with each other and their surroundings and provide new opportunities to understand how urban communities operate and evolve, often at much finer spatial and temporal resolutions than previously possible (Batty et al. 2012).

Many of these digital traces result from user-generated and geosocial media which consist largely of photos, videos, text messages and tags, along with metadata related to locational references, time stamps and links to users' profiles. There are clear challenges to using these data, as data quality, coverage, locational accuracy and thematic relevance may be uneven, unknown, or limited (e.g., Robertson and Feick 2015). However, these data are often the only source of information available that can describe routine activities, patterns of movement, and observations of events and surroundings for large numbers of people (Goodchild 2007; Poorthuis et al. 2016). In this light, it is not surprising that a sizeable body of research has coalesced around using geosocial media from Twitter, Flickr, and Foursquare, for example, to gain new insights on topics as varied as fine-grained mobility patterns, place-sensing, vernacular geographies, and public sentiment, among others (e.g., Hollenstein and Purves 2013; Crampton et al. 2013; Mitchell et al. 2013).

In this paper, we explore the use of geosocial data for analysis of personal activity of individuals through the development of individual spatial and temporal clusters of granular geosocial media traces. An individual-based approach to spatial analysis of geosocial data is in contrast to more commonly used analyses of spatial aggregate patterns of social media activity. Aggregate approaches can highlight areas that display comparatively high or low levels of personal and work-related social communication (e.g. business, tourism and entertainment districts of large cities) as illustrated in Fig. 1). However, aggregate approaches can also obscure our understanding of how individuals' use of urban space varies and contributes to overall patterns. As well, individual-level patterns can be aggregated to examine

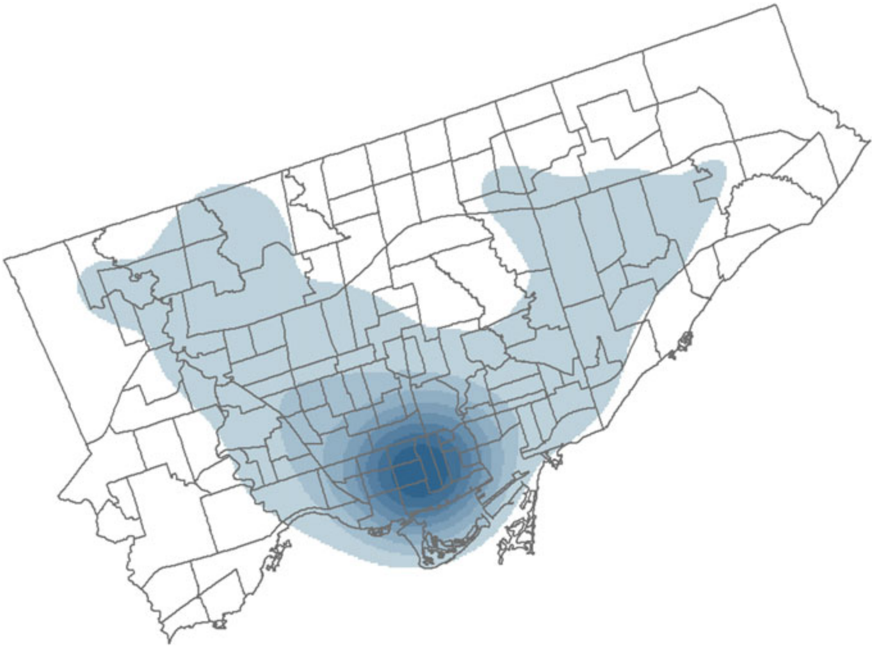


Fig. 1 Kernel density heatmap for 2.6 million geotagged Tweets in Toronto, Canada

broader-scale patterns more common to ‘big data’ analyses. The goal of this work is to examine the use of spatial and temporal clusterings for exploring place-use within urban complexes and identifying locations that are of personal or functional significance to individuals. To demonstrate the possible value of this approach, we apply it to operationalize the concepts of home (primary), work (secondary) and “third” places as a way of delineating locations of social and functional importance to individuals and groups and untangling these spaces from global patterns (Hickman 2013; Oldenburg and Brissett 1982; Soukup 2006). We use the term “activity centre” to mean the spatial expression of an individual’s most important locations that they interact with on a regular basis (Golledge and Stimson 1997).

Given the increasingly varied work-life arrangements of urban populations as they respond to, and spur, changing economic, social and technological conditions, traditional place- and time-specific divisions between home, work and leisure time become more heterogeneous. Growing numbers of people are engaged in telecommuting, are “always connected” to work via mobile devices, and augment face-to-face socialization with digital alternatives (e.g. Facebook, online gaming, etc.) (Steinkuehler and Williams 2006; Sykora et al. 2015). Such changes have implications for planners and researchers alike. For example, there are public health implications of shifting activity centres in relation to exposure to environmental hazards as well as opportunities for interventions. Geosocial data offer possibilities to examine these developments in ways that are not possible with traditional data sources, such as censuses and surveys.

The use of user-generated and geosocial media in the social sciences has not been without criticism and challenges. Common to other forms of “big” geodata, early analyses often conflated data set size and frequency with meaning and assumed that simple mapping of thousands or even millions of geosocial data points would shed light on broader social and urban processes (Crampton et al. 2013). As Kitchin (2014) notes, many forms of big data are by-products or “the exhaust” of specific activities. In contrast, small data are assembled based on carefully designed sample frames and variable selections. This has raised important questions related to data quality in multi-authored data sets, how the availability of massive data streams influences research foci, and the role of theory in analysis (Miller and Goodchild 2015).

More balanced and critical approaches to data-driven research using geosocial media have emerged recently in response to these criticisms. The representativeness of geosocial media is now recognized as being highly variable since its use and creation is dominated by relatively few advantaged groups (Haklay 2010; Miller and Goodchild 2015) and is concentrated spatially in core urban areas and places of widespread popularity (Li et al. 2013).

A growing suite of papers describe activity centre analysis from geolocated Tweets. Huang and Wong (2016) used the DBSCAN method to generate spatial clusters for users, and then inferred activity zone types from urban land use data. Huang et al. (2014) took a similar approach using DBSCAN to derive spatial clusters and then auxiliary data and metadata to infer activity space details. One of the challenges of relying on social media data to derive functional activity spaces of individuals is the huge uncertainty in the specificity of detection and how that uncertainty is distributed geographically and by demographics. In this study, we confront this challenge by adopting a bifurcated approach to geosocial activity centres. Firstly, we explore the use of DBSCAN to generate both spatial and temporal clusters for individuals at the population (e.g., big data) scale. Secondly, we examine a small sample of individuals who were recruited directly to provide information to characterize the relationships between our cluster-based methods and the true locations of participants’ home, work, and other locations. Our specific research objectives in this paper are to: (1) provide a comparative analysis of two clustering approaches to generate personal activity centres (PACs) from geosocial data (big data), (2) compare the outputs of these algorithms to personally defined activity centres for a small subset of user-reported data (small data), and (3) examine whether participants’ activity on social media are directly related to their surroundings at the time. We see this as a first step towards developing a robust methodology with known providence that allows individuals’ routine use of urban space to be examined independently. As well, we aim to provide some degree of validation and contextual enrichment of ‘big data’ approaches by integrating reports from individual participants captured within such data streams.

2 Methods

2.1 Data Sources

Data were obtained from the public Twitter Streaming API during the year of September 2013 to August 2014 within the boundaries of the city of Toronto, Canada. There were a total of 2.6 million geocoded messages and more than 99,000 unique users in the dataset after duplicates were removed. As we were interested in highly active users and given the skewed distributions of user-generated content in general, we constrained our analysis to users with a minimum of twenty-five messages over the year. Users with extremely high numbers of tweets (more than 2500) were also removed from the database since inspection revealed these users represented automated accounts ("bots") and/or businesses. This reduced our database to a final dataset of 2 million messages distributed across just over 16,000 unique users.

A secondary data source was individual survey responses from a subset of active Twitter users who resided in the study area and also had records within the larger database of tweets. We recruited participants through direct messaging and public posts on Twitter and other social media networks as part of a larger study investigating geolocated sentiment analysis and urban form. For the analysis reported here, survey data for a total of 16 participants were used. Details of the methodology for these data are reported in Sykora et al. (2015). In short, participants filled out a short entry survey upon initiation into the project that provided a baseline demographic profile including home and work locations. Over a series of weeks, each participant received short follow-up surveys that were triggered by their social media activity (e.g., posting a message to Twitter). Variables collected in these surveys included their location at the time of Tweeting (home, work, other) and the activity they were engaged in at the time of Tweeting (working, relaxing, etc.). The participant data cover a period from August 2015 to November 2016, slightly after the larger database described above. The temporal displacement between the small and big datasets used here is due to technical problems with data storage which limited our ability to collect Tweets during the concurrent period. However, we consider the impacts of this misalignment to be marginal as we collected length-of-residence data for participants in the survey, which indicated the majority had not changed residence in the period since the collection of Tweets from the API.

2.2 Analysis Methods

Recognizing some of the issues inherent in aggregate spatial analysis of geolocated social media data described above, we aimed to detect spatial areas of significance to individual users in the dataset. Two simple methods were used to derive clusters from individual social media users' geosocial data—clustering by space, and clustering by time. Our intuition is that user-activity may be similar in space (e.g., Tweeting from home or work) or similar in time (e.g., Tweeting at lunch hour from the same or from different locations), and these may be used as indicators of how people are behaving. With granular activity detection, scaled up to the population level, we may be able to investigate interesting questions about the use of space, and link expressions derived from social media content to their geographic context in a meaningful way for individuals.

2.2.1 Personal Activity Centres and Big Data Analysis of Georeferenced Tweets

The clustering algorithm used was the density-based clustering of applications with noise (DBSCAN) method (Ester et al. 1996), one of the more widely used methods for simple clustering of points. The purpose of DBSCAN is to find spatial clusters of high density and to identify points in low-density regions as outliers (or noise). The density in DBSCAN is defined by two key parameters, the neighbourhood size and the minimum number of points belonging to a cluster. Together, these parameters define the types of clusters that will be found by the algorithm. To find irregularly shaped clusters, the algorithm distinguishes between core points and border points through the concept of whether points are density-reachable, such that point p in a spatial pattern is density-reachable from point q if there is a chain of points connecting them with density above the threshold determined by the two parameters.

We operationalized the DBSCAN algorithm to find spatial clusters with a minimum number of five points, and a maximum neighbourhood size of 100 m. We set these parameters after exploratory analysis and consideration for GPS error and local mobility within the same functional place within an urban setting (e.g. movement within a single property).

For temporal clustering we set the neighbourhood size to 30 min, and again the minimum number of points to five points per cluster. Connected segments of Twitter activity within 30-min intervals would then be connected into clusters of minimum density. To capture clusters occurring over midnight, we transformed the hour of the Tweet to two dimensions (cosine and sine transforms) and used these as input into the DBSCAN algorithm. Note that because clusters are defined exclusively at the individual level, clusters can overlap spatially across users. Varying of parameters and re-running these analysis did not significantly change our overall results.

Derived clusters were ranked for each user based on spatial density. For each user, we took the set of points belonging to cluster K for user i, and computed the maximum distance separating the points. The density for user i and cluster k was computed as:

$$d_{i,k} = \frac{N_{i,k}}{Max\|P_n - P_m\| \forall P \in k}$$

and ranked such that,

$$PAC_{i,1} = P_i \in k$$

where,

$$d_{i,k} > d_{i,k+1} > d_{i,k+n}$$

Thus the set $\mathrm{PAC}_{\mathrm{i1}}$ is the highest density clustering of points for user *i*, followed by $\mathrm{PAC}_{\mathrm{i2}}$ and so on up to the number of clusters for user *i*. This ranking of individual-level clusters therefore maps onto our sociological-derived notions of space-use based on function: home, work, and other. Here, we focused mostly on the analysis of the first two orders (i.e., highest density locations). We hypothesize here that the densities will follow from highest density Tweeting at home, second highest at work, and third and higher order rankings at third places.

Note that for both spatial and temporal clusters we use the spatial distance in the denominator, as we are ultimately interested in functional areas of the city at the individual level. To distinguish between cluster types, spatial PACs are referred to here as PAC-Ss, while temporal clusters are noted as PAC-Ts. Correspondence between rank orders of PAC-Ss across the dataset therefore is seen to indicate equivalent types of spatial areas at the individual level and similarly for PAC-Ts. We use this framework to investigate spatial patterns of activity centres, and to compare to the true functional areas for the smaller subset of study participants.

Spatial and temporal clustering methods were compared using the variance of information (VI) distance for comparing clustering methods (Meila 2007). The VI statistic is based on the difference in entropy introduced by the different clustering methods. The entropy of a clustering *C* of *k* clusters, is defined as:

$$H(C) = -\sum_{k=1}^{k} P(k) log P(k)$$

where *P(K)* is the proportion of points in cluster *k*. Given two clustering methods, we can compute the joint-entropy, otherwise known as the mutual information of clustering *C* and clustering *C′* as:

$$I(C, C') = \sum_{k=1}^{k} \sum_{k'=1}^{k'} P\left(k, k'\right) log \frac{P\left(k, k'\right)}{P(k) P'\left(k'\right)}$$

which is the amount of information common to the two types of clusters. We can therefore measure the similarity between two clustering methods using the following statistic as defined by Meila (2007);

$$VI\left(C,C'\right) = [H(C) - I(C,C')] + [H(C') - I(C,C')].$$

The VI is combined entropy, minus the mutual information (i.e., the entropy common to both clustering methods). As such the VI is a metric that measures how dissimilar cluster sets are, and has a value of zero when the two cluster sets are identical. We will use the VI to measure the degree to which spatial and temporal clusterings of social media activity in Toronto are similar, and how similarities vary across space. As all analysis was done at an individual level, cluster comparisons were made for individual users, yielding a distribution of cluster distances for the two clustering methods across the approximately 16,000 users represented in the dataset. We examined the magnitude, distribution, and spatial patterns of cluster distances.

2.2.2 Comparison of Clustering Methods and Validation Data

For the subset of 16 users who participated in the validation study, we had both spatial and temporal clusters (PACs) calculated from the larger Twitter database (n = 3738), as well as data obtained from their participation in the wider study of social media use in Toronto (n = 125). Using our PAC ordering framework described earlier, we compared home, work, and other locations as reported by participants, to PAC orders as derived from clusters in space and time.

Deriving clusters for users from the Twitter database allowed us to map hypothesized functional areas for each individual in the wider population of users. We examined the proportion of cluster types in relation to the total number of Tweets measured in each neighbourhood of Toronto.

2.2.3 Examining Tweets Pertaining to the Environment

Participants who were sent a survey in response to their Twitter activity were asked to report if their Tweet related to their immediate environment. The purpose of this question was in order to better understand the explicit linkage between the content shared on Twitter and their surroundings at the time of Tweeting. Social media have been postulated as a potential tool for researching person-place linkages, especially in the context of health and epidemiological research. To explore this idea further, we estimated the probability that a submitted Tweet was about the environment in relation to the place it was sent from, whether it was a message directed to someone, whether the user used Twitter for professional or personal purposes, and the age of the user. To explore these relationships, a logistic regression model was constructed with a random effect for users. In this subset of the data, there were 59 unique users and 772 messages.

3 Results

A total of 2,090,637 messages by 16,793 unique users were analyzed for spatial and temporal clustering. In general, the number of clusters per user was higher for temporal clustering than for spatial clustering (Table 1), and temporal clusters had higher densities and higher numbers of Tweets. In terms of distribution, only 1.9% of users did not have temporal clusters, which signifies no dominant time periods in which these users posted messages. Of the remaining users, 16.9% had only one temporal cluster, 23.7% had two and 20.3% had three as their highest order temporal cluster.

In contrast to the PAC-T findings, only 2.5% of users had zero spatial clusters. Some 35.1% had a maximum of one spatial cluster, 29.8% had only two, and 14.1% had three PAC-Ss. Overall, 18.5 and 37.0% of users had four or more spatial and temporal clusters respectively. In general, users' clusters tended to be found at only a handful of discrete locations, whether defined spatially or temporally. This finding is in line with research in the urban sociological, geography and planning fields that has found consistency in urban space use due to stability in home, work and often social activity spaces (Oldenburg and Brissett 1982). Note that the spatial footprint of temporal clusters could vary significantly, as only time was used as a criterion for clustering, although density ranking was based on spatial densities. The distributions for the maximum number of clusters are shown in Fig. 2.

The distribution of VI scores ranged from 0 to 5.14, with a mean of 2.18 and standard deviation of 0.78. Randomly sampling from the upper and lower tails of the VI distribution, we present the groupings of two similar and dissimilar PAC clusterings in Fig. 3. In this figure, red points belong to first order PAC-Ss and PAC-Ts, blue points signify second order PACs, while points classed as third or higher order PACs are green. Grey points do not belong to a PAC cluster. Different behaviours are captured temporally compared to spatially. In the case of User A, the locations of the first (red) and second order (blue) clusters are reversed when their PACs are defined based on space (Fig. 3a) as opposed to time, however the points that are members of the clusters are very similar. For User B, where spatial clustering reveals two distinct clusters for order one and order two, temporal clustering captures what is likely a commuting pattern as part of the secondary cluster. In general, the more clusters that were discovered for a user, the lower their cluster agreement scores. However, by constraining comparisons between the two lower order clusters and investigating individual users, the differences and meaning behind the different clusters becomes more apparent.

Table 1 Descriptive statistics of spatial and temporal clusters of Twitter messages

Metric	Spatial	Temporal
Number of clusters total	42,606	52,351
Avg. number of points per cluster	35.80	30.42
Mean density	0.19	1.41
Mean number of clusters per user	2.54	3.23

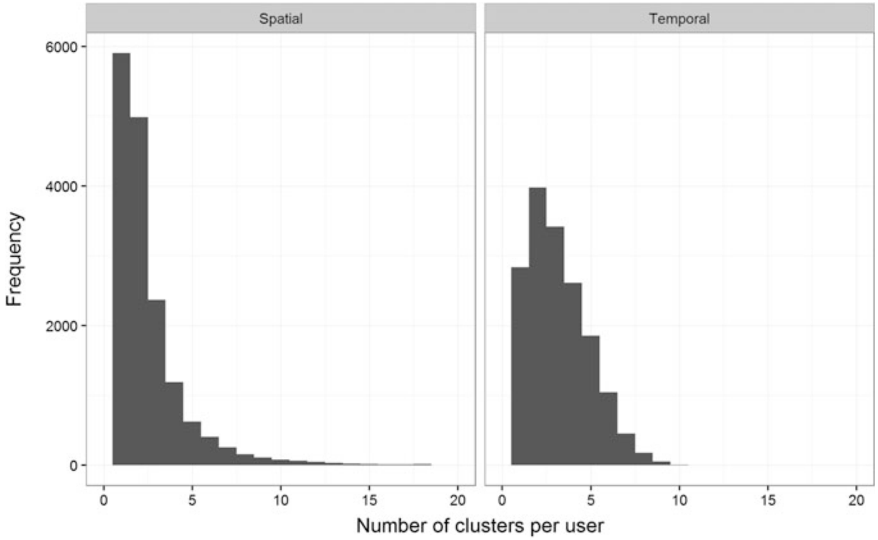


Fig. 2 Number of spatial and temporal clusters per user

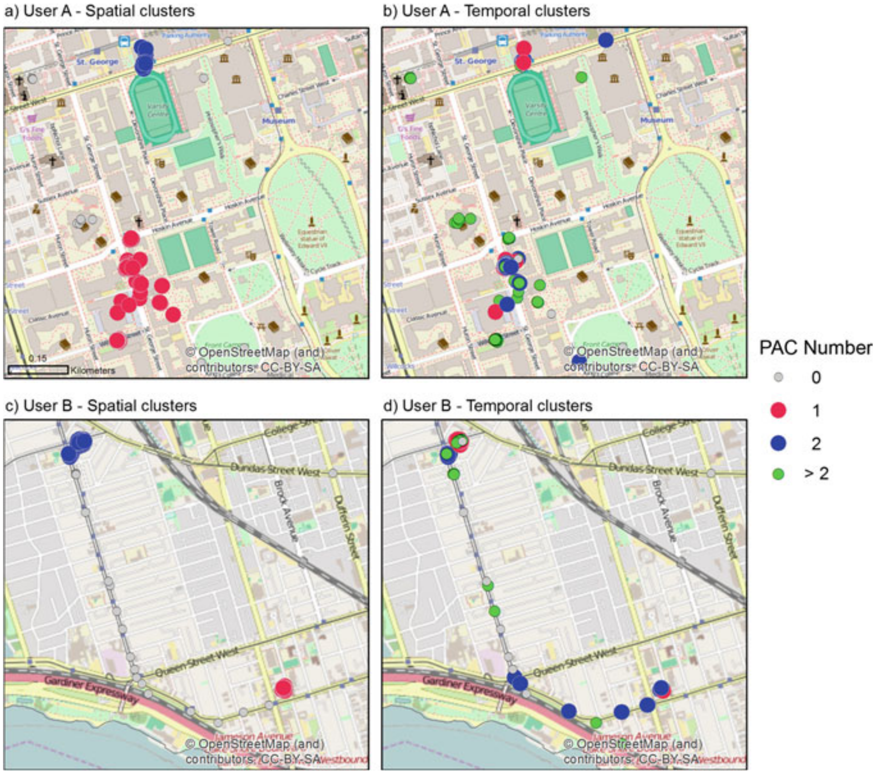

a) User A - Spatial clusters
b) User A - Temporal clusters
c) User B - Spatial clusters
d) User B - Temporal clusters
PAC Number
0
1
2
> 2
© OpenStreetMap (and) contributors, CC-BY-SA
0.15 Kilometers

Fig. 3 Spatial and temporal clusters of Twitter messages from DBSCAN; similar (**a**, **b**) and dissimilar (**c**, **d**) clusterings

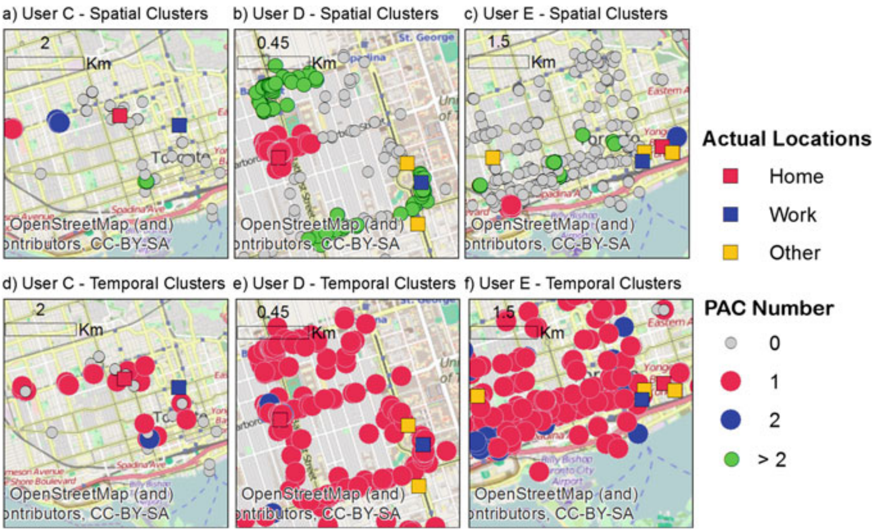

a) User C - Spatial Clusters
b) User D - Spatial Clusters
c) User E - Spatial Clusters
d) User C - Temporal Clusters
e) User D - Temporal Clusters
f) User E - Temporal Clusters
2
0.45
1.5
Km
OpenStreetMap (and)
ontributors, CC-BY-SA
Actual Locations
Home
Work
Other
PAC Number
0
1
2
> 2

Fig. 4 Spatial (**a**–**c**) and temporal (**d**–**f**) clusters for 3 users in relation to their reported home, work and other locations

In Fig. 4, we see spatial and temporal clusters in relation to users' true home and work locations as reported by the 16 study participants. In this way, we see how the spatial expression of the activity centres differs from the functionally important areas they identified. For User C's spatial clusters (Fig. 4a), we see that the tweets around their home location (red square) are spread out over a range of about 1 km in their local neighbourhood. Conversely, their highest density spatial PAC was at a distant location. Exploratory analyses of PAC-S-1 and PAC-S-2 for User C revealed these locations to be both houses in residential areas.

With temporally defined clusters (Fig. 4d), User C's PAC-T-1 (red) is between 4 and 6 pm, and PAC-T-2 is later in the evening around 8–10 pm (blue). The varied spatial pattern associated with these temporal clusters indicates that while this user regularly tweets at these times, they do so from different locations. For User D, the spatial cluster (Fig. 4b) identifies their home location well (red), while no tweets were recorded or predicted at work. Their work location is associated with a number of tweets (green), but these did not constitute their PAC-S-2. The pattern shows several areas of high activity as well as some commuting patterns, over a very small area (~1 km). For User D's temporal clusters (Fig. 4e), the majority of tweets were in PAC-T-1 (red) which was a cluster of quite regular tweeting activity that spans over the entire working day from 8 am until around midnight. User D had the most tweets of the users we investigated. User E had a lower density overall spatial pattern (Fig. 4c) and a temporal pattern of tweeting (Fig. 4e) similar to user D.

In terms of spatial clusters, for user E (Fig. 4c), PAC-S-1 was located near an urban park in the west side of the city, while their home location was located 4 km east in the central district. PAC-S-2 was located 500 m west of their home, while their work location was 500 m east of their home. This is within the range of locational accuracy provided by the postal code reference used to locate home and work locations. Temporal clustering for User E (Fig. 4f) found a PAC-T-1 to be morning, between 7 and 9 am, and PAC-T-2 was evening, between 7 and 10 pm, both of which were highly dispersed at the neighbourhood scale (~3 km).

The distribution of spatial PACs computed over the full dataset differed significantly from the pattern in Fig. 1. Figure 5 presents the proportional mapping of Tweets by neighborhood and the proportion of PAC-S-1 tweet clusters relative to the total number of tweets. Figure 5a shows a pattern similar to Fig. 1 with a familiar high concentration of messages in the downtown central core, and incrementally fewer as one moves out to suburban and non-core parts of the city.

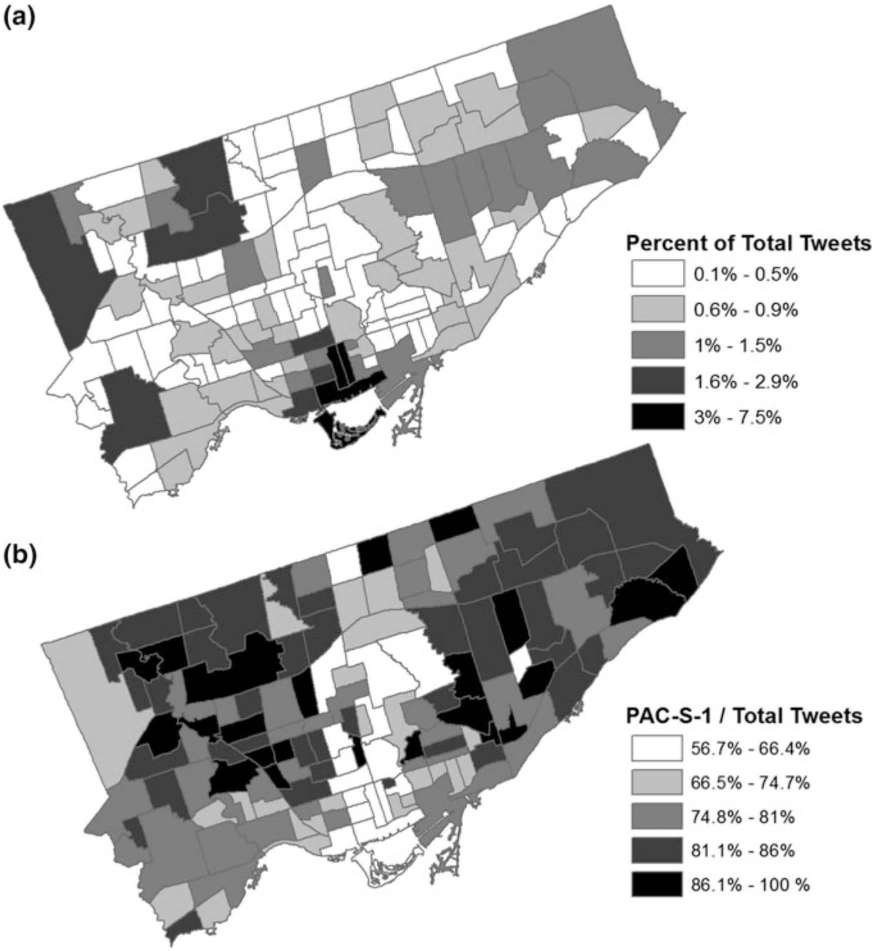


Fig. 5 Spatial distribution of: **a** total Tweet variation, and **b** spatial cluster order 1 (PAC-S-1), aggregated by neighbourhoods

Alternatively, mapping PAC-S-1 tweets as a proportion of the total tweets in a neighbourhood shows the inverse pattern where the highest values are in outlying areas. This shows that in aggregate, the spatial PAC-1 clustering captures an intuitive demarcation of predominantly working and entertainment areas of the city and predominantly residential areas.

The relationships of PAC-S-1 to home and work locations for all study participants are provided in Fig. 6. The figure shows that, in general, spatial PAC-1 clusters (PAC-S-1) are very close to home locations. Over 50% of PAC-S-1s are within 500 m of the study participants' geocoded home location. For PAC-S-2, the median distance to work locations is 1.4 km, with about 20% within 500 m and 33% within 1 km. PAC-S-3 clusters show the largest median distances to both home and work locations. For temporal clustering, the PAC-T-1 was closest to home locations, with a median distance of 1.4 km, and a median distance of 2.2 km to work locations.

For the 59 users and 772 messages investigated here, 28.0% were reported to be about users' surroundings at the time of Tweeting. By users, the 1st and 3rd

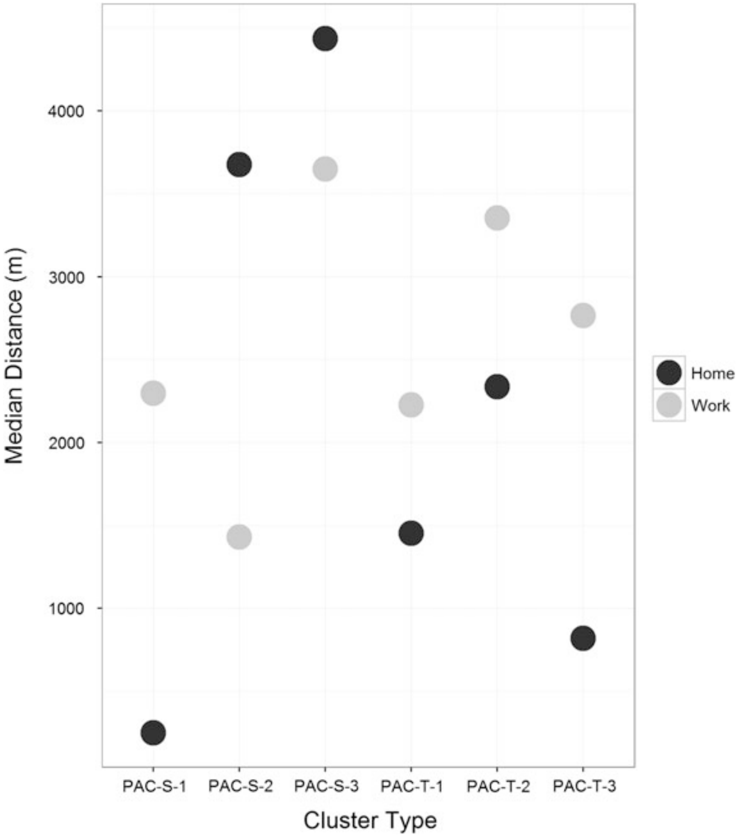


Fig. 6 Median distance between cluster types, home and work locations for all participants

Table 2 Logistic regression models results

Model term	Estimate	Standard error	P-value
Intercept	−1.92	0.607	0.001
Use for work	0.175	0.502	0.727
Tweet directed	−1.059	0.234	<0.001
Age of user	0.022	0.017	0.195
Place-work	−0.071	0.233	0.760
Place-other	1.583	0.234	<0.001

quartiles of this proportion were 7.0% and 42.9% respectively, indicating a high degree of individual variation. The results of the regression model that examined Tweets that were about users' environments are reported in Table 2. From this table we see that two variables are significant predictors of the log-odds of an environmental Tweet, whether the Tweet was directed at another user (negative association) and whether the Tweet was sent from a third place (positive association). The variance of the random effects term was 0.789 and the mixed-effects model reported here performed better than a regular logistic model based on AIC (however same effects were present plus a slight positive effect for age).

We visualize these results in Fig. 7 which shows the probability of a Tweet being about the environment as reported by users, which shows some discriminative ability of the factors reported in Table 2.

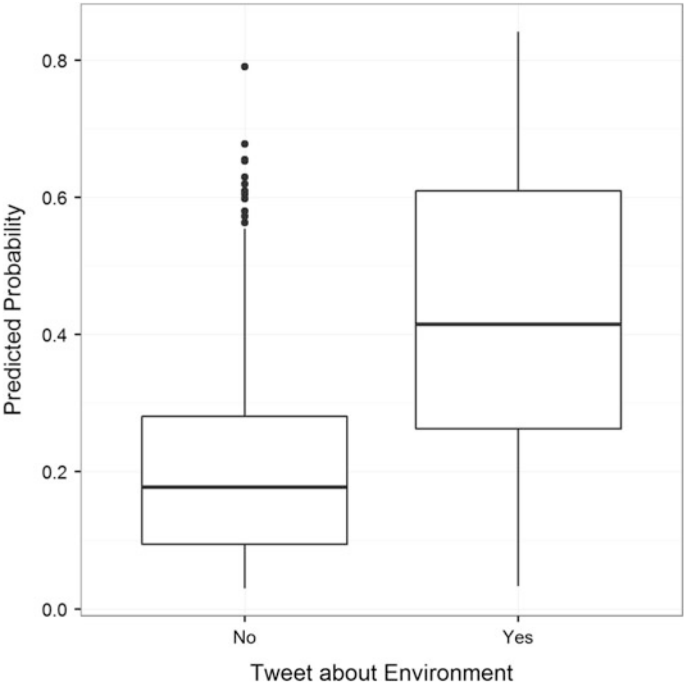


Fig. 7 Model probabilities of environment-related Tweets in participant data

4 Discussion

Our analysis provides a comparative clustering of geosocial footprints for both a large database of geolocated tweets in the City of Toronto and a subset of study participants selected for more in-depth analysis. The spatial and temporal clusterings provide reasonable estimates of personal activity centres for many individuals. From the distribution of cluster orders, we see that on average Twitter users are active from only a handful of locations, usually within close proximity to their home and work locations (e.g. Fig. 3a, b). For some users, there is also evidence of commuting behaviours in their geosocial footprints (Fig. 3d) and temporal consistency in messaging behaviour.

Spatial clustering of individual tweet locations provided a more realistic estimate of space use at the individual level. We discovered clusters that aligned with our hypothesized activity space categories of home and work locations as evidenced in Fig. 6. The hypothesis that personal activity centres derived from geosocial data can estimate real functional areas seem to be supported by the analysis. Our focus on using clustering to delimit the spatial and temporal extents of ranked PACs complements other approaches that derive individuals' activity spaces through social network connections or message content. Study participants' reported home, work and other activity locations provided a means to examine how well the calculated PACs corresponded to what may be considered true functional areas for individuals. In this way, we sought to couple the analytical advantages of big geodata, including extensive sampling of populations and unobtrusive data collection methods, with more structured small data that serve as an indicator of the validity of the clusterings to delimit meaningful locations at the individual level.

There are important limitations to this work that we are seeking to address in ongoing research. Like many types of big data, geosocial media is partial in nature. For instance, social media users are not representative of a city's socio-demographic composition, these data can only be created under certain conditions (e.g. not while driving), and only a small fraction of the data are encoded with GPS coordinates (Morstatter et al. 2013). In our study, these realities were accentuated by the small set of active participants that produced the validation data and, in recognition of this limitation, ongoing work is focused on expanding the participant pool within Toronto and selected other cities in North America. As well, even though our analyses focused on individuals, we aimed to understand both how people use social media in the city (generalizing to the broader population of Twitter users), and how individuals' digital expressions are impacted by their environment. In general however, representativeness issues related to analysis of social media may be alleviated as these technologies become more utilized and accessible by more of the population (Boyd 2014).

There are several interesting implications of this work that merit further study. First, the approach demonstrated for deriving spatial and temporal footprints from geosocial data streams offers new information to understand individuals' use of urban space. This could be of particular value for examining the dynamics of space use in response to evolving work-life patterns, changes to the urban fabric (e.g. promotion of mixed land uses), or seasonal conditions (e.g. snow storms, heat waves). Second, an individual's PACs can be enriched with complementary data extracted from their social media user profiles (e.g. interests, profession, demographic characteristics) or from analysis of the content of their messages (e.g. sentiment analysis). This could help researchers to understand some of the reasons that underlie a person's routine activity patterns. In particular, combining detailed analysis of message content with PACs may hold potential for public health planning and evaluation in cities, examining spatial health and equity issues related to congestion, pollution, stress, and fear of crime. Finally, there is a clear need to consider methods to protect personal privacy given the growing sources of geospatial traces and new methods to rapidly derive associated outputs. This could include, for example, limited random displacement of data points prior to developing PACs or post-process randomization of data points within the convex hull of a PAC.

Figure 5 demonstrated the stark contrast in spatial pattern when mapping aggregate tweet density versus mapping aggregate patterns of PACs. There is potential to provide more nuanced spatial analysis of geosocial data by providing functional meaning to otherwise disaggregate patterns. For example, in an urban analytics setting, we could constrain an analysis of place-based issues to only those messages located within the vicinity of individuals' PAC-1s to help filter out noise in the signal and provide meaningful spatial context to other forms of public engagement tools for urban managers and planners.

Finally, model results indicate that the 'environmental content' in tweets may be limited, and importantly, may vary systemically. Significant associations with the place from which the message was sent, as well as whether it was directed could be used as filtering criterion when doing environmental analyses of Twitter data. Deeper understanding of these forms of variability in the content and intention in geosocial data is needed before such data can be used to their greater potential for understanding human activities and interactions with the environment.

Our analysis provided comparative clustering using existing algorithms to derive personal activity centres from geosocial media data. We demonstrated the effectiveness in locating home and work locations from simple spatial clustering for a majority of users investigated. Ongoing studies and validation data will provide further insight into the preliminary patterns investigated here.

Acknowledgements The authors gratefully acknowledge our study participants as well as the Social Sciences and Humanities Research Council of Canada for funding this research.

References

Batty M, Axhausen KW, Giannotti F, Pozdnoukhov A, Bazzani A, Wachowicz M et al (2012) Smart cities of the future. Eur Phys J Spec Top 214(1):481–518

Boyd D (2014) It's complicated: the social lives of networked teens. Yale University Press, New Haven

Crampton JW, Graham M, Poorthuis A, Shelton T, Stephens M, Wilson MW, Zook M (2013) Beyond the geotag: situating "big data" and leveraging the potential of the geoweb. Cartogr Geogr Inf Sci 40(2):130–139

Ester M, Kriegel HP, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd, vol 96, no 34, pp 226–231

Golledge RG, Stimson RJ (1997) Spatial behavior: a geographic perspective. Guilford Press

Goodchild MF (2007) Citizens as sensors: the world of volunteered geography. GeoJournal 69:211–221

Haklay M (2010) How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. Environ Plan Des B 37:682–703

Hickman P (2013) "Third places" and social interaction in deprived neighbourhoods in Great Britain. J Hous Built Environ 28(2):221–236

Hollenstein L, Purves R (2013) Exploring place through user-generated content: using Flickr tags to describe city cores. J Spat Inf Sci 1(January):21–48

Huang Q, Cao G, Wang C (2014) From where do tweets originate?: a GIS approach for user location inference. In: Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks, ACM, pp 1–8

Huang Q, Wong DWS (2016) Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us? Int J Geogr Inf Sci 30:1873–1898

Kitchin R (2014) The real-time city? Big data and smart urbanism. GeoJournal 79(1):1–14

Li L, Goodchild MF, Xu B (2013) Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. Cartogr Geogr Inf Sci 40:61–77

Meilă M (2007) Comparing clusterings—an information based distance. J Multivar Anal 98 (5):873–895

Miller HJ (2010) The data avalanche is here. Shouldn't we be digging? J Reg Sci 50(1):181–201

Miller HJ, Goodchild MF (2015) Data-driven geography. GeoJournal 80(4):449–461

Mitchell L, Frank MR, Harris KD, Dodds PS, Danforth CM (2013) The geography of happiness: connecting twitter sentiment and expression, demographics, and objective characteristics of place. PLoS ONE 8(5):e64417

Morstatter F, Pfeffer J, Liu H, Carley KM (2013) Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. arXiv preprint arXiv:1306.5204

Oldenburg R, Brissett D (1982) The third place. Qual Soc 5(4):265–284

Poorthuis A, Zook M, Shelton T, Graham M, Stephens M (2016) Using geotagged digital social data in geographic research. In: Clifford N, French S, Cope M, Gillespie T (eds) Key methods in geography. Sage, London, pp 248–269

Robertson C, Feick R (2015) Bumps and bruises in the digital skins of cities: unevenly distributed user-generated content across US urban areas. Cartogr Geogr Inf Sci 1–18

Soukup C (2006) Computer-mediated communication as a virtual third place: building Oldenburg's great good places on the world wide web. New Media Soc 8(3):421–440

Steinkuehler CA, Williams D (2006) Where everybody knows your (screen) name: online games as "third places". J Comput-Mediat Commun 11(4):885–909

Sykora MD, Robertson C, Shankardass K, Feick R, Shaughnessy K, Coates B, Lawrence H, Jackson T (2015) Stresscapes: validating linkages between place and stress expression on social media. Published by CEUR Workshop Proceedings

Part II
Spatio-Temporal Analysis

Spatio-Temporal Road Coverage of Probe Vehicles: A Case Study on Crowd-Sensing of Parking Availability with Taxis

Fabian Bock, Yuri Attanasio and Sergio Di Martino

Abstract Finding a parking space is a key mobility problem in urban scenarios. Parking Guidance Information (PGI) systems could mitigate this issue, but they require information about on-street parking availability. An encouraging solution discussed in the literature is crowd-sensing by a fleet of probe vehicles, which can continuously scan the current state of parking lanes during their regular trips. Nevertheless, the achievable spatio-temporal coverage of such a fleet is still an open point. In this paper, we present an evaluation of the suitability of a fleet of taxis as probe vehicles for parking crowd-sensing. In particular, we exploited a dataset of real-world trajectories collected from about 500 taxis over 3 weeks in San Francisco (USA), to extract their movement patterns. The quality of achievable parking information is determined by combining these patterns with availability data collected from parking sensors in about 400 road segments. For that, the last sensing of a taxi is considered as an estimate of parking availability in a road segment. Results of movement patterns show a heterogeneous distribution in time and space. Nevertheless, already about 500 taxis are enough to provide availability information with a maximal deviation of one parking space per road segment in about 90% of time steps. Thus, taxis show a high suitability as probe vehicles for crowd-sensing parking information.

Keywords Parking search · On-street parking · Taxi trajectories · Crowd sourcing

F. Bock (✉)
Institute of Cartography and Geoinformatics, Leibniz University,
Hannover, Germany
e-mail: fabian.bock@ikg.uni-hannover.de

Y. Attanasio
University of Naples “Federico II”, Naples, Italy
e-mail: y.attanasio@studenti.unina.it

S. Di Martino
Department of Electrical Engineering and Information Technologies,
University of Naples “Federico II”, Naples, Italy
e-mail: sergio.dimartino@unina.it

A. Bregt et al. (eds.), *Societal Geo-innovation*, Lecture Notes
in Geoinformation and Cartography, DOI 10.1007/978-3-319-56759-4_10

1 Introduction

The search for parking spaces in large cities is considered to have a strong socio-economic impact, including traffic congestion, air pollution, and drivers' wasted time. Shoup (2006) found that, on average, 30% of the traffic in investigated areas was due to drivers looking for a parking space. A root cause of this problem is that drivers have no knowledge about where there could be a free parking space matching their expectations, and so they have to roam. *Parking Guidance Information* (PGI) systems, i.e. systems able to spread up-to-date information about the state of parking infrastructure could significantly reduce this problem, since drivers could be guided directly to an area (or a parking facility) with a high likelihood of free spaces (Ma et al. 2014; Richter et al. 2014). The challenge with PGI is how to obtain detailed real-time on-street parking space availability information (Ma et al. 2014; Xu et al. 2013).

To date, there are mainly two solutions: either instrument parking infrastructure with special sensors (magnetometers, cameras, etc.) (SFMTA 2014), or leverage crowd-sensing solutions, which can be either participatory or opportunistic (Ganti et al. 2011), like mobile apps (Ma et al. 2014) or probe vehicles (Mathur et al. 2010). These solutions have contrary pros and cons. The instrumentation leads to continuous and highly accurate parking information, but it is very expensive to install and maintain. Thus, it is not scalable to cover entire urban scenarios (Xu et al. 2013). Smartphone apps can be very cheap, but they have significant problems in acquiring sufficient real-time information about the parking availability (Bock et al. 2016a). Probe vehicles can be a promising compromise, since series sensors, like side-scanning ultrasonic sensors or windshield-mounted cameras, can be profitably used to determine the state of the parking stalls as the vehicle moves across a road segment (Bock et al. 2016a). The problem is that accessing these sensors is not possible without the involvement of the car manufacturer. Researchers proved that it is possible to effectively retrofit new, cheap sensors just on special types of vehicles with a high annual mileage, like buses or taxis and collect the data on their regular trips (e.g. Mathur et al. 2010).

Summarizing, there is evidence that some kinds of vehicles, like taxis, can be profitably turned into probe vehicles to detect the availability of on-street parking spaces. However, the movements of these vehicles do not uniformly cover the city, but it can be assumed that they follow some spatial distribution with higher coverage on main roads, and temporal distribution with peak hours. While the spatio-temporal patterns of taxi pick-up and drop-off locations are well studied (e.g. Liu et al. 2012a; Hoque et al. 2012), the spatio-temporal suitability of parking sensing with taxi probe vehicles is not investigated yet.

To fill this knowledge gap, we conducted an empirical investigation, by exploiting two public mobility datasets from the municipality of San Francisco (USA). The first one contained nearly one month of trajectories coming from 536 taxis. The second one contained information from about 8,000 parking stalls, collected every 5 min from sensors embedded in the asphalt. By analyzing these datasets, we aimed at understanding the potential spatio-temporal coverage of taxi used as probe vehicles and evaluating the potential performance of parking availability sensing. Thus, we aimed at addressing the following research question: *Is crowd-sensing using a fleet of taxi as probe vehicles appropriate to obtain current parking availability information?*

The main contributions of this paper to the body of knowledge are: (I) an empirical analysis of the spatio-temporal coverage of a fleet of taxis, including the identification of typical spatial and temporal moving patterns of taxis and achievable update frequency of parking state information, (II) a comparison of this spatio-temporal coverage with the on-street parking behavior, (III) a quality assessment of the parking information that could be obtained by the taxi probe vehicles.

2 Preliminary Concepts and Related Work

Given the relevant socio-economic impact of the parking problem, many research efforts have been directed towards improving parking search (Teodorović and Lučić 2006). In the following, we provide an overview of the technology to detect free parking spaces with probe vehicles and of existing research analyzing taxi trajectories.

2.1 Detecting Free Parking Spaces with Probe Vehicles

As mentioned in the introduction, real-time on-street parking availability information can be collect either by sensors in the infrastructure at constant time intervals or by crowd-sensing solutions in an irregular frequency. Focusing on probe vehicles, they can be utilized in multiple ways to detect free parking spaces. The most investigated methods involve the use of ultrasonic sensors (Degerman et al. 2006; Mathur et al. 2010; Satonaka et al. 2006) or vision-based sensors (Houben et al. 2013). As a probe vehicle drives by a road segment, its sensors are constantly scanning the sides of the vehicle, like the vehicle named "Transmitter" in Fig. 1, looking for gaps sufficiently large to fit a vehicle. Each gap is a potential free parking space, whose location and size can be sent to a back-end server via cellular network. Learning methods can be applied to distinguish between free parking spaces and gaps with parking restrictions (Bock et al. 2016b). This information is aggregated and processed further into live digital maps of parking availability trends, that are subsequently broadcasted to all the interested users like the vehicle "Receiver" in Fig. 1.

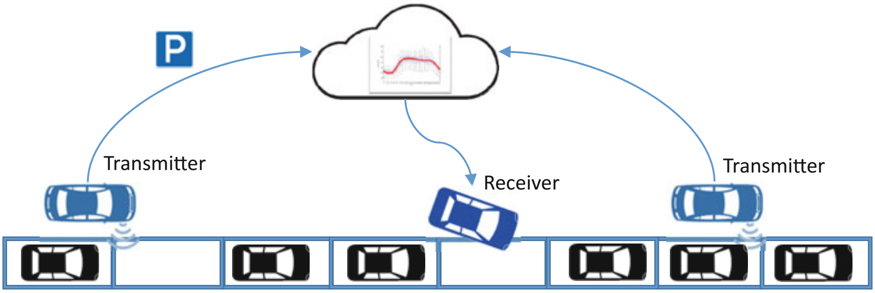


Fig. 1 The architecture based on probe vehicles to obtain a dynamic map of parking space availabilities (based on Robert Bosch GmbH (2016))

2.2 Taxi Trajectory Analysis

In the literature, there are many papers investigating taxi trajectories, but most of them are focused on pick-up and drop-off events as basis for different kinds of socio-economic or mobility studies. For example, in the paper by Liu et al. (2012b), the authors exploit taxi trajectories collected in Shanghai to determine interurban land use variations by focusing on temporal variations of taxis' pick-ups and drop-offs. In another paper (Liu et al. 2012a), the authors investigate the correlation between taxi trajectory and characteristics of human mobility, also taking into account geographical impacts. Taxi trajectories have also been analyzed in order to determine the levels of attractiveness of a certain area, using a clustering approach to group spatio-temporally similar pick-ups and drop-offs points (Yue et al. 2009).

To the best of our knowledge, there is only one paper aimed at evaluating the quality of parking availability crowd-sensing by using taxis as probe vehicles (Mathur et al. 2010). They also use the taxi trajectory dataset from San Francisco (Piorkowski et al. 2009) to determine the city coverage. Anyhow, there are several differences with respect to our investigation. As that paper represents a first attempt to quantify the suitability of taxis for parking detection, there are some simplifications in the process that we tried to overcome. In particular, the authors did not perform a real map-matching of the taxi trajectories, but assigned the trajectories to regular "cells" of 175 × 190 m. They did not consider the presence of intersections, the driving directions, and the number of lanes per road. Also, they defined a simple average function to estimate the temporal coverage of taxis passing in a cell without consideration of the time of day. All these differences with our investigation will be discussed in detail in Sects. 4 and 5.

3 Spatio-Temporal Processing of the Data Sources

To evaluate the potential of taxis as probe vehicles for parking availability estimation, we exploited two public datasets about mobility in San Francisco (USA): a dataset of taxi trajectories and a dataset of infrastructure-based on-street parking availability information.

Since these two datasets have been collected from different providers, they adopt a very different representation of the data. For instance, taxi trajectories are expressed in terms of raw GPS points while parking information is related to some specifically defined identifiers of road segments. As a consequence, to use them in combination, some non-trivial data cleansing and pre-processing steps were required.

In Fig. 2, we describe the processing pipeline we applied to perform our investigation. More in detail, as for the parking data, we started from a dataset containing raw availability data. We performed a map overlay on top of the *OpenStreetMap* road network and applied some data cleansing to get a dataset useful to compute parking availability, for each road segment and each time instant of the considered time frame. As for the taxi data, the GPS points are map matched to the OpenStreetMap road network and the underlying taxi routes identified. In the next step, we performed a spatial, temporal, and logic filtering of the data. Measures are calculated describing the coverage of the taxis and results are aggregated temporally for time of the day or week and spatially for road classes. This spatio-temporal coverage represents at what times the parking data could be obtained if the taxi were used as probe vehicles. Thus, we downsampled the original parking dataset according to that coverage. Finally, we compared the full dataset of parking information with the downsampled one to quantify the difference in terms of parking infrastructure state detection that could be achieved with the use of probe vehicles.

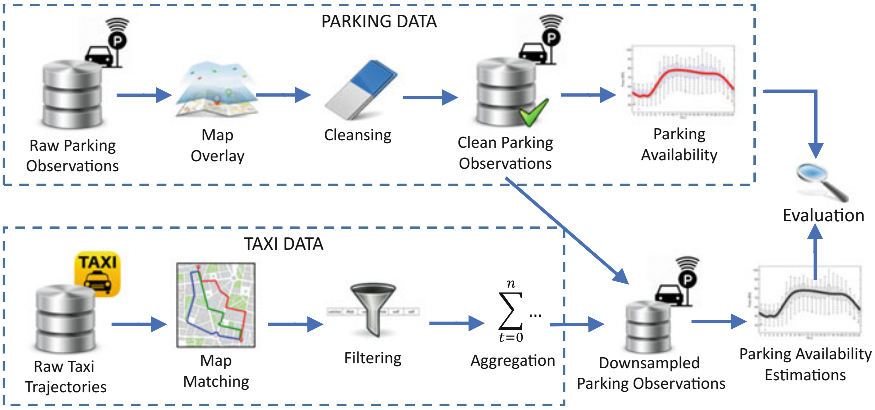


Fig. 2 Schematic view of the evaluation pipeline

In the following we provide a detailed description of each step of this pipeline, about taxi data in Sect. 3.1 and parking data in Sect. 3.2. The results will be presented and discussed in Sects. 4 and 5.

3.1 Spatio-Temporal Processing of Taxi Trajectories

The taxi trajectories contain GPS coordinates of 536 vehicles from the *Yellow Cab* company, collected over 25 days in the San Francisco Bay Area from 2008/05/17 until 2008/06/10, and are publicly available (Piorkowski et al. 2009).

The data was collected within the *Cabspotting* project, which aimed at the extraction of socio-economic properties of regions from the taxi patterns. Each taxi periodically provided information on its latitude and longitude, timestamp, and occupancy (1 = occupied, 0 = free) to a central server. There are 11,219,955 GPS points in total, with a median time gap between two subsequent GPS measures of 60 s and time gaps between 30 and 120 s for 86% of all records.

3.1.1 Map Matching of Taxi Trajectories

The low frequency and the noise of the GPS records require an elaborate map matching to align the sequence of GPS points with the road network (from *OpenStreetMap* in our case) and to identify the visited road segments. To this aim, we employed the map matching approach described by Axer et al. (2015), which is based on Lou et al. (2009).

More in detail, the sequence of GPS points of every taxi is split into shorter sequences, with a split every time there is a change in the taxi occupation status, or whenever there is a time gap between subsequent GPS points longer than 3 min. Short sequences of less than 5 GPS points and sequences with implausible velocities between two GPS points are discarded. Then, for every GPS point of a sequence, candidate street segments are identified from the road network. The road segments between two subsequent projected GPS points are identified via a shortest path search. In a following global optimization step, the road candidates achieving the highest score, based on spatial and temporal criteria, are selected.

As a result, 8,839,942 GPS points (78.8% of the dataset) were successfully matched to the road network. The analysis of taxi movements in the following sections is a conservative estimation, since better positioning information could add up to 27% more points to our dataset.

3.1.2 Filtering of Taxi Trajectories

After map matching, some temporal and spatial filtering of the trajectories has been performed on the dataset. We excluded the first and last day of the dataset, since records do not cover these days completely. We reduced the data interval to a 3-week period, in order to represent each day of the week equally, leading to an observation period ranging from 2008/05/18 (Sunday) to 2008/06/07 (Saturday). Also, we removed all the taxis not moving for more than a week, because it is not a common behavior for this kind of vehicle, resulting in a final set of 486 taxis. Finally, we removed all the map matched trajectories not overlapping road segments covered by parking availability information (see Sect. 3.2).

In addition, when probe vehicles sense on-street parking occupancy using the technology described in Sect. 2.1, only those vehicles moving on a lane immediately next to a parking lane can obtain information about the parking availability. Therefore, the number of lanes and the driving direction need to be considered. For our investigation area, the road network from OpenStreetMap provides the number of lanes only for about 52% of the road segments. We assumed that, if the number of lanes is even on a two-way road, half of the lanes belong to each driving direction. For road segments with uneven lane count or missing lane information, we manually identified the number of lanes and directions using Google StreetView.

Since the lane choice of the taxi drivers is unknown, in the rest of the paper we estimate a pessimistic and an optimistic case:

- **Pessimistic case**: we consider a uniform distribution of the taxis over the lanes, and therefore, only 1/# (lanes) of the visits (randomly chosen) in the corresponding driving direction will be used for the subsequent calculations.
- **Optimistic case**: the taxis are rewarded for driving on the lanes next to the parking lanes. Thus, we consider all road segment visits. Just in one-way roads with multiple lanes and parking on both sides, observation of both sides is not possible and only half of the visits are assigned randomly to each parking lane.

3.1.3 Aggregation of Taxi Trajectories

For the evaluation of the suitability of taxis as probe vehicles, we need to measure the temporal resolution of the taxi visits for each road segment. In the study of Mathur et al. (2010), the authors used the average time between two subsequent taxi visits $\bar{T}_s$ at a location, defined as:

$$\bar{T}_s = \frac{1}{N_{taxi} - 1} \sum_{i=1}^{N_{taxi}-1} (t_{i+1} - t_i) \qquad (1)$$

where t_i are the ordered timestamps of taxi visits and N_{taxi} is the total number of visits at a road segment in the investigated period.

In our opinion, even if this measure provides an insight on the phenomenon, it can be biased if taxis do not pass at regular time intervals. To clarify, let us consider the likely scenario of a road segment where a parking stall has just been occupied. Few instants after this parking event, a taxi passes on that road segment, scanning the parking situation and detecting the important information that there is a new occupied stall. If this taxi is immediately followed by some other taxis, all the subsequent ones will read exactly the same parking situation, thus not providing any additional information (assuming correct sensors). Nevertheless, in Eq. (1), the additional taxis reduce the average time $\bar{T}_s$, thus leading to better statistics, even if in reality the additional taxis would not provide any information gain. Therefore, the average time $\bar{T}_s$ is not able to capture the differences in coverage for irregular visit times.

To overcome this limitation, we propose a different way to quantify the temporal resolution of taxi coverage. We aim at measuring the *temporal interval* between the instant of time t when a mobility-related event could happen (for instance a new parking space becomes available) and the instant of time of the next taxi visit $t_{next}(t)$ (also called *time gap* in the following). Since the time of the event is unknown in most scenarios, we average this temporal interval over all instants of time:

$$\bar{T}_g = \frac{1}{N_{steps}} \sum_{i=0}^{N_{steps}-1} (t_{next}(t_0 + i \cdot \Delta t) - (t_0 + i \cdot \Delta t)) \tag{2}$$

where $t_{next}(t)$ is the timestamp of the next taxi visit after time t and N_{steps} is the number of time steps with interval Δt (= 5 min in our computations).

In addition to this measure, we compute also the total number of road segment visits per time interval. It describes the activity of the taxis and can be compared to the parking fluctuation measure, defined in Sect. 3.2.2.

The taxi visits are aggregated by hour of day, days of the week, and the road classes given by the attribute 'highway' in OpenStreetMap.[1] The considered values are 'primary', 'secondary', 'tertiary', 'unclassified', 'residential', and 'service' (in descending importance of the road). Note that 'unclassified' is a defined class and does not stand for an unknown class assignment.

3.2 Spatio-Temporal Processing of Parking Availability Data

On-street parking availability data was obtained from the SFpark project (SFMTA 2014). The main focus of this project, ran between 2011 and 2014, and whose costs exceeded $46 million, was the improvement of on-street parking management in San Francisco (SFMTA 2014).

In this project, more than 8,000 parking spaces were equipped with static sensors embedded in the asphalt. The number of current occupied parking spaces and

[1] Definitions are available under: http://wiki.openstreetmap.org/wiki/Key:highway.

total number of provided parking spaces per road segment was collected by querying their publicly provided REST API every 5 min from 2013/06/16 (Sunday) until 2013/07/06 (Saturday). Within the SFpark project, a road segment (also named *block face*) is defined as one side of a road between two intersections, and on average has about 6 parking spaces.

3.2.1 Data Cleansing of Parking Data

In the final report of the SFpark project (SFMTA 2014), some sensor issues were described. To remove road segments with implausible sensor values, we analyzed the data regarding time gaps, constant sensor signals, and low parking occupancy rates. In particular, a road segment is removed from our investigation if the number of occupied parking spaces is constant or sensor data is missing for more than three days. Also, road segments are removed that never show an occupancy rate of at least 85% in the full investigation period, based on the assumption that either some sensors failed or parking is not competitive there and thus less relevant to monitor. A map of the remaining road segments is shown in Fig. 3.

3.2.2 Matching to Road Network and Data Aggregation

The locations of the parking lanes in the SFpark project are described by a line geometry. The road segment in OpenStreetMap, that is closest to the middle point of the parking lane and not intersecting the parking lane, is assigned to this parking lane. Visual evaluation showed that this simple approach was already sufficient to correctly match all the road segments.

Let us observe that road segments and times of day with higher parking fluctuations need also higher probe vehicle coverage. Thus, the required coverage of the probe vehicles is related to the frequency of parking occupancy changes. The *parking fluctuation* is measured as the difference in occupied parking spaces between two subsequent observations within the 5-min record interval. Note that more parking changes might happen within every 5-min interval. For example, if the difference is zero, it is either possible that there was no parking event or that one or more vehicles left the parking lane and the same number of vehicles parked there within this time interval. Thus, this measure does not represent the actual turnover rate. However, as a relative measure, it still indicates times and locations of higher and lower parking fluctuation.

Analogously to the aggregation of the taxi data in Sect. 3.1.3, parking data is aggregated by hour of day, day of the week, road classes.

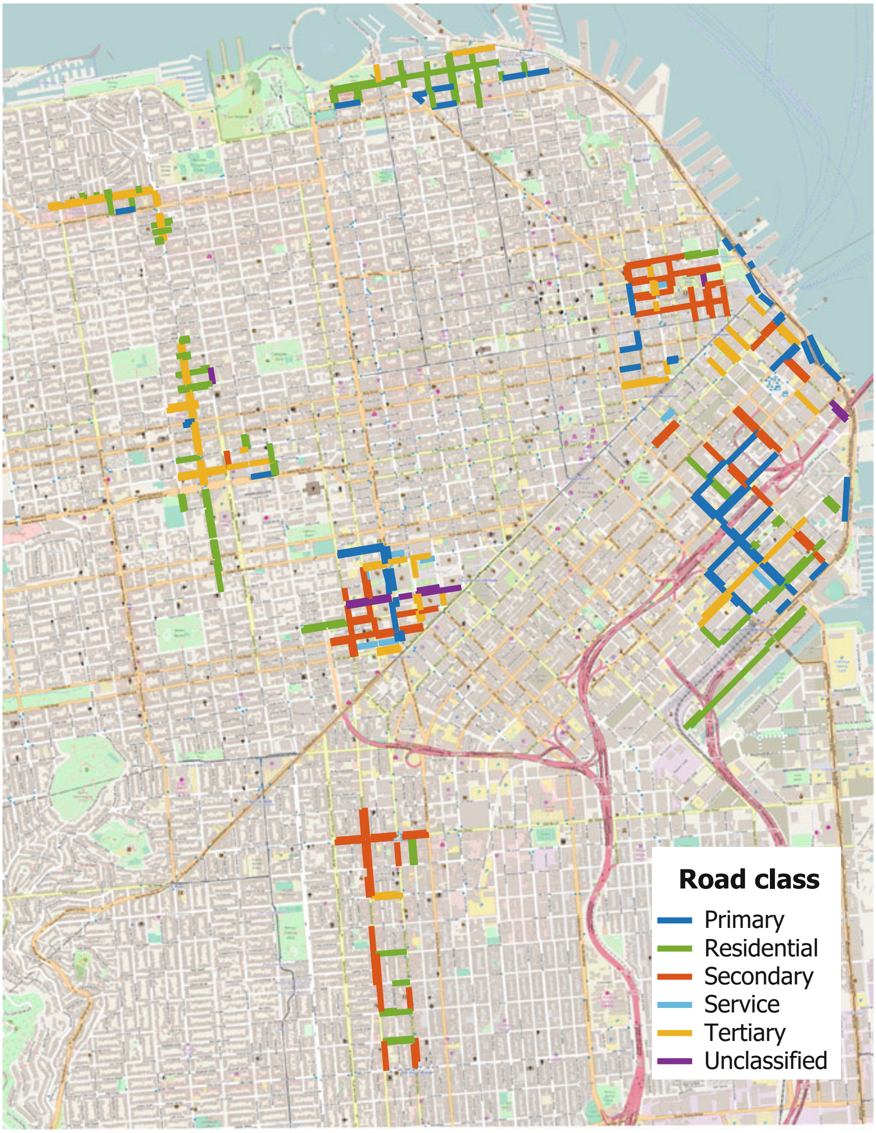


Fig. 3 Map of the investigation area showing all road segments with parking availability data. The *colors* indicate the road classes. The background map is taken from OpenStreetMap

4 Results and Discussion on Analysis of Taxi Trajectories

The taxi trajectories of the 486 taxis contributed to nearly 190,000 visits on road segments per day. The distribution over the days of the investigation period has some variation, between 140,806 and 209,619 road segment visits per day. All road segments covered by the SFpark project sensors were visited by at least one taxi during the investigation period. Except for three road segment, all the others were visited on average more than four times per day, up to a maximum of more than 1,400 times per day. Figure 4 shows a map of the mean time gaps until the next taxi visit.

4.1 Distribution of Time Gaps

The cumulative distribution function of time gaps for all time steps and road segments is shown in Fig. 5. We provide this distribution for the optimistic and pessimistic cases of lane coverage, as defined in Sect. 3.1.2. As expected, the optimistic case shows shorter time gaps than the pessimistic one throughout the full curve. Both curves increase quickly for small time gaps, with 64% and 57% of time steps on road segments observed in less than 20 min, respectively.

In comparison to Mathur et al. (2010), the shapes of the curves of the cumulative distribution function are very similar: many road segments (‘cells’ in their case) are visited within a small time gap (‘average inter-polling time’) of several minutes. Only for a few road segments, the time gaps become longer than 1 h. However, for the absolute values, Mathur et al. (2010) reports “an average inter-polling time of 25 min for 80% of the cells” for 300 taxis, while in our case, for the optimistic case, we observe a time gap of 42 min for 80% of the road segments and instants of time with 486 taxis. We believe that the main reasons for these deviations, besides the modified time measure, are due to the fact that we considered driving directions (less than 40% of the road segments are one-way roads) and we performed a map matching, compared to their cell approach which overestimates the visits, especially at intersections. Nevertheless, also the other differences described in Sect. 2 contribute to these deviations.

4.2 Temporal Coverage of the Taxis over the Day

The temporal coverage of the taxis shows a high variability over the day, as described by the boxplot of Fig. 6. Indeed, taxis show very little activity (and therefore we have longer time gaps) in the early morning hours, around 4 a.m., while the best coverage and lowest times can be achieved around noon and in the evening, around 8 p.m. The plot also shows that the variance is much higher during the night than during the day.

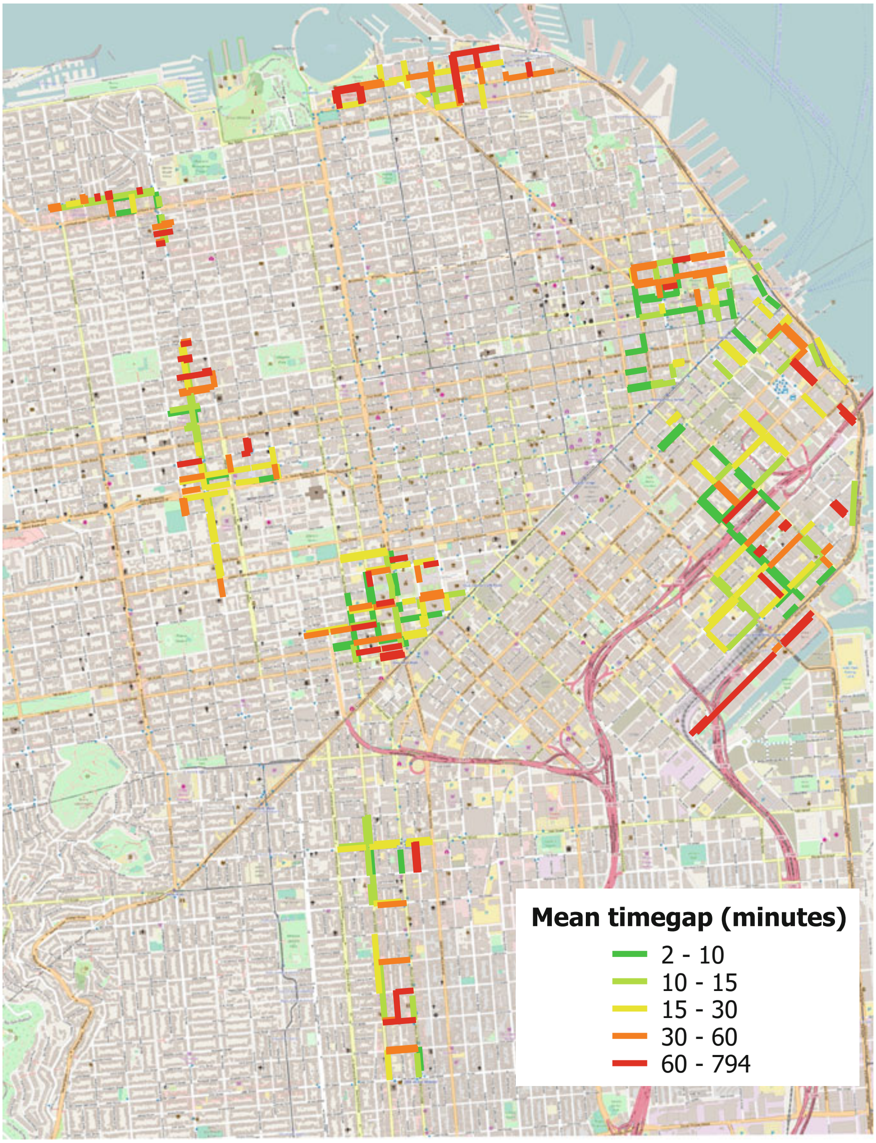


Fig. 4 Map of the road segments colored according to the mean time gap. The background map is taken from OpenStreetMap

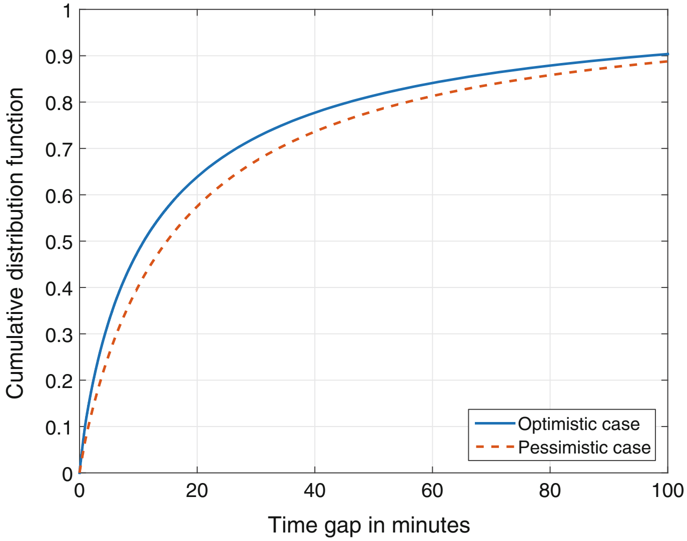


Fig. 5 Cumulative distribution function of the time gap until the next taxi visit, accumulated over all time steps and road segments. Optimistic case: taxis drive next to the parking lanes. Pessimistic case: taxis drive in all the lanes with equal probability

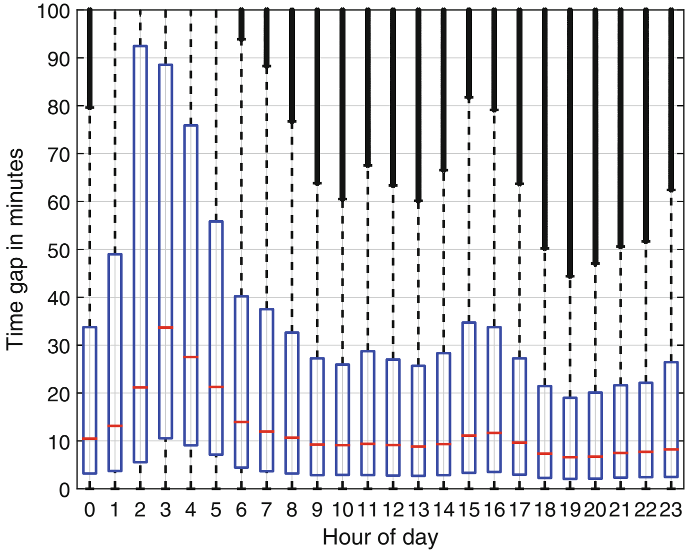


Fig. 6 Boxplot of the time gaps for all road segments, grouped by the hours of day. Due to the magnitude of the considered data, the outliers appear as a *solid black line*

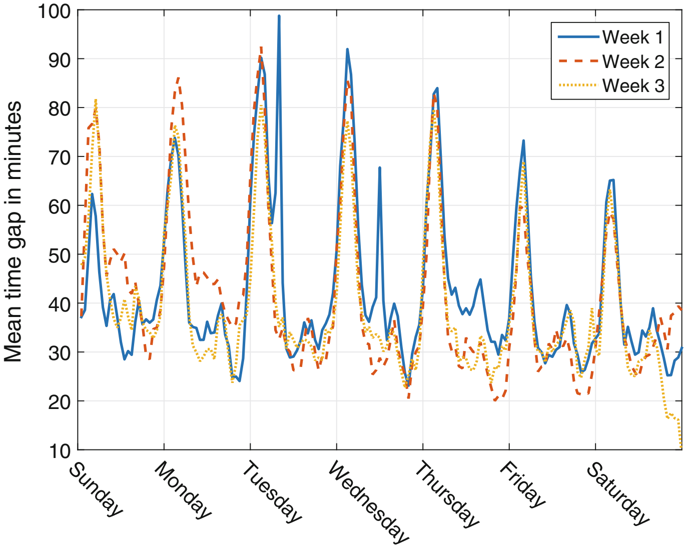


Fig. 7 Mean values of time gaps over the days of the week

This result is consistent with temporal analysis of taxi trips in the literature, like in Liu et al. (2012a) for Shanghai or NYC TLC (2014) for New York.

In Fig. 7, we present the average time gaps over the days of the week. From that figure, we can see that there is a very similar behavior among weeks, with time gap peaks during the night. We observed four cases of strong deviations: on Tuesday and Wednesday of the first week, there are irregular peaks with high time gaps since no GPS points were recorded for 80 and 50 min, respectively. Monday of the second week shows larger time gaps. This is most likely due to the fact that it was the Memorial day, a US holiday. Finally, the time gaps drop at the end of the third week, due to the way we computed gaps (there is no 'next taxi visit' anymore in some road segments). Since all 3 weeks show some atypical behavior, a median week pattern is calculated. For each hour, the median week is determined and the taxi visits of this hour are used as the 'typical' weekly taxi coverage in the following section.

4.3 Spatial Coverage of the Taxis over Road Classes

The coverage of taxis is also highly related to the road class, as defined by OpenStreetMap (see Fig. 8). The most important urban road classes 'Primary', 'Secondary', and 'Tertiary' have a time gap 11 min, 16 min, and 27 min on average over the day, respectively. The average values per hour are always less than 50 min, even during the night. About 65% of all road segments belong to these road classes. On the contrary, small roads of the classes 'Unclassified' and 'Service' never present

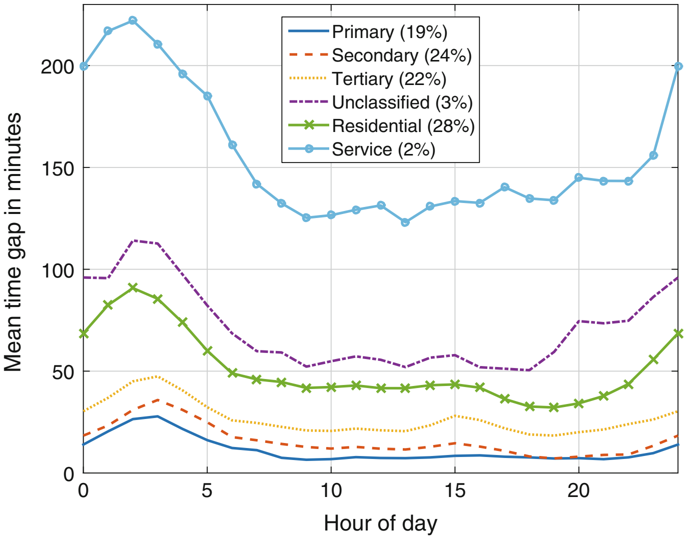


Fig. 8 Mean values of time gaps over the hours of day for different road classes. The *numbers* in *brackets* are the percentages of the road classes

gaps smaller than 50 min, even during the most active hours and going up to nearly 4 h during the night. However, only 5% of all investigated road segments belong to these two classes.

The higher number of taxi visits on main roads shows that taxi drivers prefer those roads to reach the destination. This corresponds to the general traffic flow, which is usually higher on main roads than on side roads.

5 Evaluation of Taxi Suitability for Parking Crowd-Sensing

The pre-processed datasets of taxi trajectories and parking availability are combined in this section to evaluate the suitability of taxis for parking crowd-sensing. In Sect. 5.1, we compare the taxi activity and parking fluctuation over the hours of day. Then, we assess the suitability of the crowd-sensing by assuming that taxis observe the parking availability according to their movement pattern. As quality measure, we compare the status of parking availability of the last taxi visit with the current availability (Sect. 5.2). Finally, in Sect. 5.3, we discuss the threats to validity of our approach in detail.

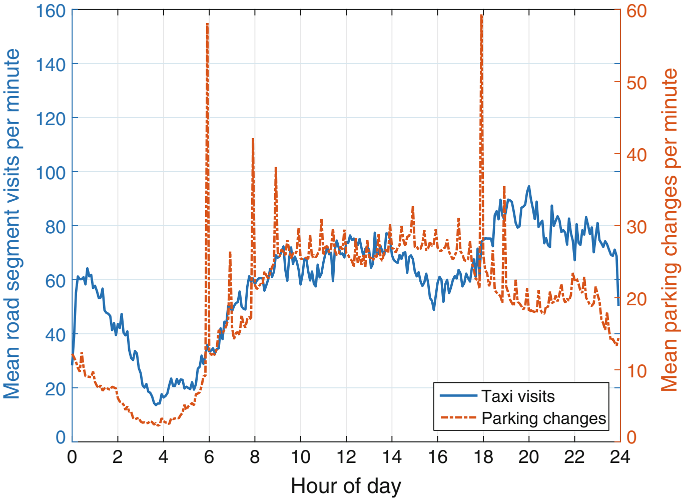


Fig. 9 Comparison of day patterns for the mean number of road segment visits by the taxis and the mean count of parking occupancy changes

5.1 Comparison of Taxi Rate and Parking Fluctuation

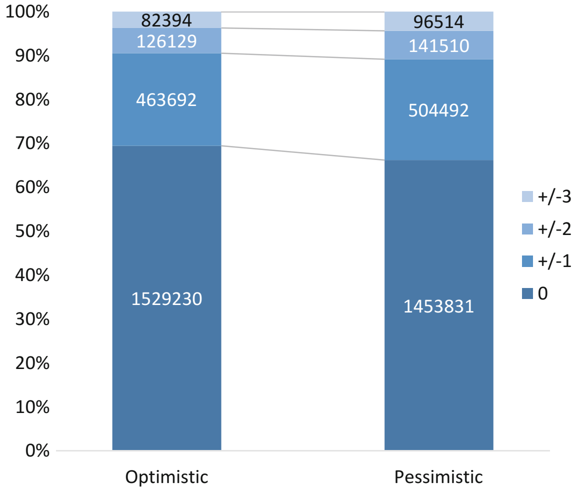
As seen in Sect. 4, the time gaps among taxi visits are strongly dependent on the time of day. In Fig. 9, we show the mean number of road segment visits by the taxis in a typical week (see Sect. 4.2) and the mean count of parking occupancy changes in dependence on the time of day. Both curves show the lowest values during the night at around 4 a.m. and a strong increase in the morning. While taxis are mostly active around noon and in the evening hours (around 8 p.m.), parking activity is highest during the day between 10 a.m. and 4 p.m. and decreases at evening hours. In addition, distinct peaks appear at full and half hours; most distinct at 6 a.m. and 6 p.m.. These peaks are caused by changes in parking permission. The parking occupancy increases drastically after the end of tow-away periods. Therefore, sensing at these times is very relevant to capture the new occupancy rate.

5.2 Quality of Taxi Coverage for Parking Sensing

To assess the effectiveness of taxis as probe vehicles, and thus to answer our research question stated in the introduction, we compare the actual parking occupancy against the parking occupancy observed by the last taxi visit, for each 5-min time step and each road segment.

To this aim, we consider that the parking occupancy is observed by a taxi in the same time step, if the time between the parking occupancy measurement and the

Fig. 10 Differences between observed and actual parking occupancy counted for all time steps for the optimistic and pessimistic case



next taxi visit is less than 2.5 min. If the time to the next taxi visit is between 2.5 and 5 min, the visit is assigned to the subsequent time step. This evaluation is executed for every street segment separately over the 3 weeks of parking data. As taxi movement pattern, we use the 'typical' median movements, as defined in Sect. 4.2.

The differences between actual parking occupancy and occupancy at the last taxi visit are computed for both the optimistic and pessimistic cases and they are shown in Fig. 10. Results show that probe vehicles are providing correct observation in 69.5% and 66.2% of the cases, respectively. If we accept an error of ±1 parking space, called *sufficient* in the following, the taxis are able to provide a sufficient observation in 90.6% and 89.2% of the cases. Only for 3.7%/4.4% of time steps, a difference of at least 3 parking spaces is observed. Considering the observation as a binary decision (i.e. either the parking lane is full or at least one parking space is available) a precision of 80.1% and a recall of 81.0% is achieved for the optimistic case (pessimistic case: 78.5 and 79.5%).

These results show that parking availability sensing with 486 taxis is already promising with the simple extrapolation of the last taxi observation state. A more sophisticated parking availability estimation like a Kalman filter including historical trends might even further improve these results as shown by Xu et al. (2013) for smartphone crowd-sensing. On the other side, we did not consider sensor mistakes in this evaluation, which surely decrease the quality of obtained results. To improve the coverage in lower class roads, the system is also easily scalable by including more taxis or other vehicles, like for instance delivery vehicles. The differences between the pessimistic and optimistic cases are notably small. Therefore, an incentive system to motivate drivers to use the lanes next to the parking lanes appears to be less relevant. More promising could be an incentive system for drivers to change the route for a higher information gain.

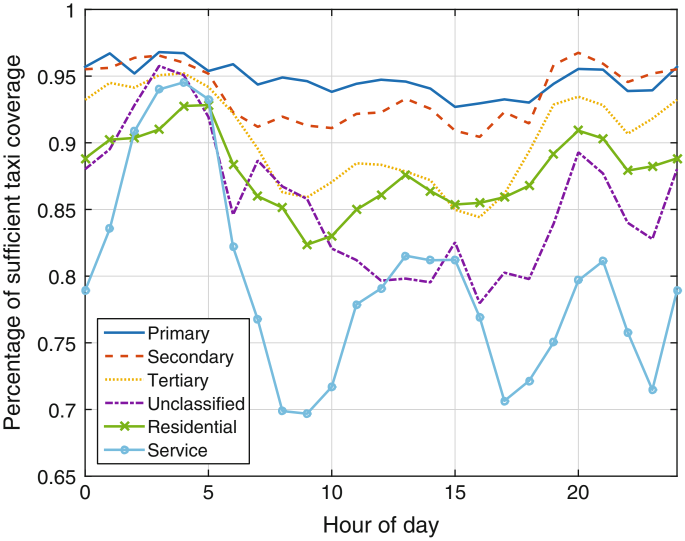


Fig. 11 Percentage of time steps whose difference between last occupancy observation and actual parking occupancy is sufficient (within ±1), for different road classes

Since the taxi observation pattern shows strong spatial and temporal dependence, the quality is also evaluated for time of the day and road classes (see Fig. 11 for the optimistic case). The percentage of time steps with sufficient taxi coverage is highest for primary roads and lowest for service roads at most times during the day, which corresponds to the taxi time gaps evaluation in Sect. 4.3. Interestingly, the ratio of correct parking occupancy observations is highest during the night for all road classes although only few taxis drive at these times since also parking changes are low then.

5.3 Threats to Validity

The main threat to validity of our empirical analysis comes from the fact that the employed datasets are from different years (May/June 2008 vs. June/July 2013). It is possible that socio-economic changes, seasonal factors, and the emergence of mobility services like Uber[2] may have influenced the taxi movement patterns. Nevertheless, the infrastructure of San Francisco did not change significantly during these few years and we found that the traffic remained nearly constant.[3] Also, the generalizability of these results to other cities needs to be evaluated, as the city structure and the mobility behavior vary among cities and countries.

[2] https://www.uber.com/.

[3] http://www.dot.ca.gov/trafficops/census/.

The analysis of taxi trajectories in San Francisco is based on a dataset of about 500 taxis which is about one third of all taxis in this area. While the movement patterns might be different to other taxis' patterns, we assume they are representative since they cover the fleet of the largest taxi provider at that time, namely *Yellow Cab*. We assumed that taxis choose either the lane next to the parking lane (optimistic case) or randomly (pessimistic case). However, taxi drivers might prefer specific lanes with faster traffic flow. In few road segments, taxi lanes exist, but they are ignored in our evaluation.

For the parking data, we applied an intensive data cleansing to avoid roads with irregular parking pattern. However, the sensor performance also suffered from environmental influences like nearby tram lines (SFMTA 2014) which could not be considered in the evaluation. The dataset consists of data in a 5-min interval. Parking changes between the observations and thus the probe vehicle performance could not be determined. However, a parking information service with a 5-min interval can already be considered as very detailed since the duration to reach a specific road segment is also of the order of several minutes.

6 Conclusion

In this paper, we analyzed the spatio-temporal movement pattern of taxis in San Francisco and evaluated their suitability as probe vehicles for parking availability monitoring. Results show that taxi movements reveal a strong dependence on the hour of day and the road class. Despite this heterogeneous distribution in time and space, already about 500 taxis are enough to reach a sufficient coverage (meaning a maximal deviation of one parking space) for about 90% of road segments and time steps.

As future research, it would be interesting to evaluate the probe vehicle approach in combination with different availability estimation methods and different sensor quality estimates. Furthermore, we would like to investigate the potential of varying the routes of probe vehicles to increase the knowledge gain, allowing for minor detours.

In conclusion, taxis as probe vehicles are a promising solution to reduce parking search traffic by sensing parking availability. Since the sensing technology is already advanced, traffic management authorities could apply this solution either by installing additional sensors in the probe vehicles soon or by accessing the car sensors in cooperation with the car manufacturer in the longer term.

Acknowledgements We gratefully thank Steffen Axer for computing the map matching of the taxi trajectories. This research has been supported by the German Research Foundation (DFG) through the Research Training Group SocialCars (GRK 1931). The focus of the SocialCars Research Training Group is on significantly improving the citys future road traffic, through cooperative approaches. This support is gratefully acknowledged.

References

Axer S, Pascucci F, Friedrich B (2015) Estimation of traffic signal timing data and total delay for urban intersections based on low-frequency floating car data. In: Proceedings of the 6th mobility TUM 2015

Bock F, Di Martino S, Sester M (2016a) What are the potentialities of crowdsourcing for dynamic maps of on-street parking spaces? In: Proceedings of the 9th ACM SIGSPATIAL international workshop on computational transportation science, IWCTS'16, pp 19–24

Bock F, Liu J, Sester M (2016b) Learning On-Street Parking Maps from Position Information of Parked Vehicles, Springer, pp 297–314

Degerman P, Pohl J, Sethson M (2006) Hough transform for parking space estimation using long range ultrasonic sensors. Technical report, SAE Technical Paper

Ganti RK, Ye F, Lei H (2011) Mobile crowdsensing: current state and future challenges. IEEE Commun Mag 49(11):32–39

Hoque MA, Hong X, Dixon B (2012) Analysis of mobility patterns for urban taxi cabs. In: 2012 international conference on computing, networking and communications (ICNC), pp 756–760

Houben S, Komar M, Hohm A, Luke S, Neuhausen M, Schlipsing M (2013) On-vehicle video-based parking lot recognition with fisheye optics. In: Proceedings of IEEE international conference on intelligent transportation systems, pp 7–12

Liu Y, Kang C, Gao S, Xiao Y, Tian Y (2012a) Understanding intra-urban trip patterns from taxi trajectory data. J Geogr Syst 14(4):463–483

Liu Y, Wang F, Xiao Y, Gao S (2012b) Urban land uses and traffic source-sink areas: Evidence from gps-enabled taxi data in shanghai. Landscape Urban Plann 106(1):73–87

Lou Y, Zhang C, Zheng Y, Xie X, Wang W, Huang Y (2009) Map-matching for low-sampling-rate gps trajectories. In: ACM SIGSPATIAL GIS 2009

Ma S, Wolfson O, Xu B (2014) Updetector: sensing parking/unparking activities using smartphones. In: Proceedings of the 7th ACM SIGSPATIAL international workshop on computational transportation science, pp 76–85

Mathur S, Jin T, Kasturirangan N, Chandrasekaran J, Xue W, Gruteser M, Trappe W (2010) Parknet: drive-by sensing of road-side parking statistics. In: Proceedings of 8th international conference on mobile systems, applications, and services, pp 123–136

NYC TLC (2014) 2014 Taxicab Fact Book. http://www.nyc.gov/html/tlc/downloads/pdf/2014_taxicab_fact_book.pdf. Accessed 13 Nov 2016

Piorkowski M, Sarafijanovic-Djukic N, Grossglauser M (2009) CRAWDAD dataset epfl/mobility (v. 2009-02-24). Downloaded from. http://crawdad.org/epfl/mobility/20090224

Richter F, Di Martino S, Mattfeld DC (2014) Temporal and spatial clustering for a parking prediction service. In: 2014 IEEE 26th international conference on tools with artificial intelligence (ICTAI), pp 278–282

Robert Bosch GmbH (2016) Bosch community-based parking. http://www.bosch-mobility-solutions.com/en/connected-mobility/community-based-parking/. Accessed 27 June 2016

Satonaka H, Okuda M, Hayasaka S, Endo T, Tanaka Y, Yoshida T (2006) Development of parking space detection using an ultrasonic sensor. In: Proceedings of the 13th ITS world congress

SFMTA (2014) SFpark: putting theory into practice. Pilot project summary and lessons learned. http://sfpark.org/resources/docs_pilotsummary/. Accessed 24 June 2016

Shoup D (2006) Cruising for parking. Transp. Policy 13(6):479–486

Teodorović D, Lučić P (2006) Intelligent parking systems. Eur J Oper Res 175(3):1666–1681

Xu B, Wolfson O, Yang J, Stenneth L, Yu PS, Nelson PC (2013) Real-time street parking availability estimation. In: 2013 IEEE 14th international conference on mobile data management (MDM), vol 1, pp 16–25

Yue Y, Zhuang Y, Li Q, Mao Q (2009) Mining time-dependent attractive areas and movement patterns from taxi trajectory data. In: 2009 17th international conference on geoinformatics, pp 1–6

A Big Geo Data Query Framework to Correlate Open Data with Social Network Geotagged Posts

Gloria Bordogna, Steven Capelli and Giuseppe Psaila

Abstract The objective of this paper is to fill in the gap existing between the need of business companies and analysts to spatially correlate open geo-data and social network geo-tagged information, and the lack of tools to enable this task in an easy way. To this end we propose a novel declarative query language named *J-CO* (*J*SON *Co*llections) to perform complex queries on heterogeneous collections of data stored within a NoSQL database as JSON objects.

Keywords Collections of JSON objects · Geo-tagged data sets · Query language for geographical analysis · Powerful spatial operators

1 Introduction

Nowadays, a potential powerful societal driver of the data economy is offered by both crowd-sourced information within social networks and open data published on the Web by public administrations. On the one side, due to the diffusion of smart devices, more and more messages posted within social networks are geo-tagged by the geographic coordinates of the places where messages were sent. Collections of geo-tagged posts constitute *Social Media Geographic Information* (SMGI), which are recognized as a valuable source, not only of geographic facts but even of perceptions, opinions and feelings of local communities in space and time. Therefore it is important to enable territory administrators, sociologists and psychologists to perform spatio-temporal analysis of SMGI.

G. Bordogna
CNR IREA, Via Bassini 15, 20133 Milano, Italy
e-mail: bordogna.g@irea.cnr.it

S. Capelli (✉) · G. Psaila
University of Bergamo—DIGIP, Viale Marconi 5, 24044 Dalmine BG, Italy
e-mail: steven.capelli@unibg.it

G. Psaila
e-mail: giuseppe.psaila@unibg.it

A. Bregt et al. (eds.), *Societal Geo-innovation*, Lecture Notes
in Geoinformation and Cartography, DOI 10.1007/978-3-319-56759-4_11

On the other side, the huge amount of geo-referenced information on territorial, social and economic resources in the form of *Open Geo-Data* can complement information coming from social networks with semantic information on the territory. In fact, correlating these two sources of information can help understanding habits and needs of citizens.

Currently, we are only at the starting line, w.r.t. the availability of tools that enable analysts and business companies to access and spatially correlate geo-tagged information coming from social networks with open geo-data. In fact, at present, GISs provide only a visualization framework, while they lack query facilities to easily perform complex analysis of geo-tagged data in databases, as both open geo-data and social network geo-tagged posts are.

The situation is further complicated by the fact that open geo-data are highly heterogeneous in genre, data schema, semantics and formats, due to missing regulations and policy agreements and directives: the result is that homogeneous data sets are published by different public administrations with different formats. As far as social networks are concerned, every system provides posts with its own format and structure. The only common aspects of both open data and social network data is *JSON* and *GeoJSON* Butler et al. (2016) exchange formats. The ability of JSON to flexibly represent formats with varying data structures (often perceived as a negative aspect) turns out to be a positive aspect when having to manage highly heterogeneous information, as in our context. For this reason, *NoSQL* databases Nayak et al. (2013) are becoming more and more popular, in particular those, like *MongoDB* Banker (2011), which natively stores collections of heterogeneous JSON objects.

The goal of this paper is to overcome the inability of GISs to flexibly query and transform heterogeneous data sets of geo-tagged JSON objects, in order to answer the needs of business companies and analysts wishing to correlate open geo-data and social network geo-tagged posts. To this end, we propose a novel declarative query language named *J-CO* (*J*SON *Co*llections), which allows specifying complex queries on data sets (e.g., open data and social network posts) stored in a NoSQL database as JSON objects, and visualized within a GIS. As Fig. 1 shows, the query language is thought to operate as a plug-in within a traditional GIS: the user writes the query, that is sent to the *J-CO Framework*, which takes input data set from the NoSQL database and stores the results into the same NoSQL database (namely, MongoDB); this way, the user can immediately visualize the query results on his/her maps, possibly integrated with other data sets (information layers).

J-CO is conceived to provides a pool of operators for various data transformation tasks on collections of JSON objects (possibly geo-tagged with GeoJSON). In particular, we defined specific spatial operators on the geometric attributes of JSON objects. By means of a running examples, we illustrate their application to study the semantic mobility patterns of tourists visiting a region.

The paper is organized as follows.[1] Section 2 introduces the basic concepts underlying the J-CO framework, such as the data model and the execution model. Section 3

[1] The work was partially supported by the FHfFC Project jointly funded by CNR and Regione Lombardia D.G.R. n.3866 17/07/2015.

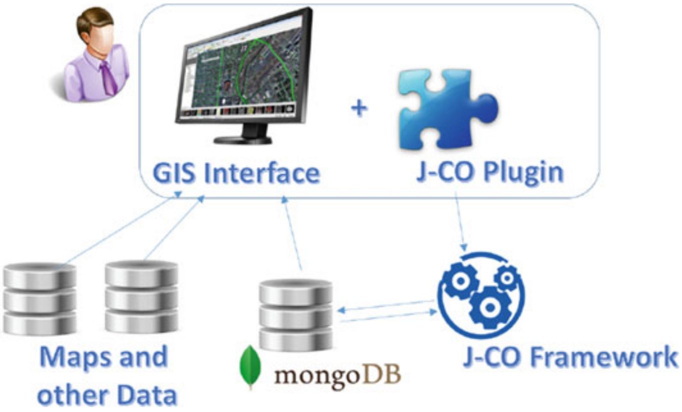


Fig. 1 Perspective application of the J-CO framework

introduces the basic J-CO operators, useful to perform those operations not strictly related with spatial properties. Then, Sect. 4 presents J-CO spatial operators, specifically designed to easily write complex analysis based on the spatial properties. Section 5 discusses relevant related work and, finally, Sect. 6 draws the conclusions.

2 J-CO Framework

The J-CO query language is part of a framework devised to provide advanced query capabilities on collections of JSON objects stored in NOSQL databases such as MongoDB. The *J-CO Framework* provides a *Data Model*, an *Execution Model* and a pool of operators which constitute the query language.

While the data model and the execution model will be presented in the rest of this section, Sect. 3 will introduce basic J-CO operators, while Sect. 4 will introduce spatial J-CO operators.

2.1 Data Model

The data model is based on the basic concept of *JSON* object. JSON (JavaScript Object Notation) is a de facto standard serialized representation for objects. Fields (object properties) can be simple (numbers or strings), complex (i.e., nested objects), vectors (of numbers, strings, objects).

As far as spatial representation is concerned, we rely on the GeoJSON standard Butler et al. (2016). In particular, we assume that the geometry is described by a field named `geometry`, defined as a *GeometryCollection* objects type in GeoJSON standard. The absence of this top-level field means that the object does not have an explicit geometry.

As an example, consider the collection (array) of objects named `Tourist Tweets` reported in Listing 1. This collection contains three JSON objects describing three tweets posted by travelers. The `geometry` field describes a point representing the position of the tourist in the world, when he/she posted the tweet (note that it is based on the *GeometryCollection* object type of GeoJSON). Consider this collection inside the database named `MyDB`.

Listing 1 Collection `TouristTweets`

```
[{"TweetId":"603755043112759296",
  "UserName":"'User1",
  "geometry":{"type":"GeometryCollection",
              "geometries":[
                 {
                  "type": "Point",
                  "coordinates":
                      [45.63120, 90.289]
                 }]
              },

  "Date": "2015-05-28",
  "Time": "04:49:54",
  "StartDate": "2015-05-28",
  "OriginAirport": "Malpensa"
},

{"TweetId":"603755043112759340",
 "UserName":"'User1",
 "geometry":{"type":"GeometryCollection",
             "geometries":[
                {
                 "type": "Point",
                 "coordinates":
                    [46.75813, 92.459]
                 }]
           },
 "Date": "2015-05-30",
 "Time": "10:50:43",
 "StartDate": "2015-05-28",
 "OriginAirport": "Malpensa"
},

{"TweetId":"603755043112759450",
 "UserName":"'User2",
 "geometry":{"type":"GeometryCollection",
             "geometries":[
                {
                 "type": "Point",
                 "coordinates":
                    [49.75421, 95.322]
                }]
             },
 "Date": "2015-07-30",
 "Time": "11:30:52",
 "StartDate": "2015-07-29",
 "OriginAirport": "Linate"
}]
```

The following definition formalizes the concepts of *collection* and *Database*.

Definition 1 (*Collections and Databases*) A *Database db* is a set of collections: $db = \{c_1, \dots, c_n\}$. Each collection c has a name *c.name* (unique in the database) and an instance $Instance(c) = [o_1, \dots, o_m]$ that is a vector of JSON objects o_i.

2.2 Execution Model

Queries transform collections stored in the databases, and generate new collections that will be stored again into these databases, for persistence. For simplicity we call such databases as *Persistent Databases*. Hereafter, we introduce the concept of query process and its execution model.

Definition 2 (*Query Process State*) A *state s* of a query process is a tuple $s = (tc, IR)$, where tc is a collection named *Temporary Collection.* while IR is a database named *Intermediate Results database*.

Definition 3 (*Operator Application*) Consider an operator op. Depending on the operator, it is parametric w.r.t. input collections (present in the persistent databases or in IR) and, possibly, an output collection, that can be saved either in the persistent databases or in IR.

The application of an operator op, denoted as $\overline{op}$, is defined as:

$$\overline{op} : s \rightarrow s'$$

where both domain and codomain are the sets of query process states. The operator application takes a state s as input, possibly works on the temporary collection $s.tc$, possibly takes some intermediate collection stored in $s.IR$; then, it generates a new query process state s', with a possibly new temporary collection $s'.tc$ and a possibly new version of the intermediate result database $s'.IR$.

The idea is that the application of an operator starts from a given query process state and generates a new query process state. The *temporary collection tc* is the result of the operator; alternatively, the operator could save a collection as *intermediate result* into the IR database, that could be taken as input by a subsequent operator application.

Definition 4 (*Query*) A query q is a non-empty sequence of operator applications, i.e., $q = \langle \overline{op}_1, \dots, \overline{op}_n \rangle$, with $n \geq 1$.

Thus, the query is a sequence of operator applications; each of them starts from a given query process state and generates a new query process state, as defined by the following definition.

Definition 5 (*Query Process*) Given a query $q = \langle \overline{op}_1, \dots, \overline{op}_n \rangle$, a query process QP is a sequence of query process states $QP = \langle s_0, s_1, \dots, s_n \rangle$, such that $s_0 = (tc : [\,], IR : \emptyset)$ and, for each $1 \leq i \leq n$, it is $\overline{op} : s_{i-1} \rightarrow s_i$

The query process starts from the empty temporary collection $s_0.tc$ and the empty intermediate results database $s_0.IR$. Thus, the J-CO query language provides operators able to start the computation, taking collections from the persistent databases, while other operators carry on the process, continuously transforming the temporary collection and possibly saving it into the persistent databases. Nevertheless, the query could be complex and composed by several subtasks, thus the temporary collection could be saved into the intermediate results database *IR*. At this point, a new subtask can be started by the same operators that can start the query, which can take collections either from persistent databases or from the intermediate result database as input, giving rise to a new subtask.

3 Basic J-CO Operators

In this section, we introduce the basic J-CO operators. These operators allow users to perform basic operations on collections, which are not related with the spatial dimension. Nevertheless, they are necessary to write queries and perform complex transformations.

Hereafter, we will make use of terms reported in Table 1 to introduce the syntax of operators. Regarding the notation for the syntax of operators, we will make use of the * symbol to denote 0 or more repetitions and of the + symbol to denote 1 or more repetitions; square brackets denote optionality; the vertical bar | separates alternatives.

Table 1 Meaning of terms

Terms	Meaning of terms
dbName	Name of a persistent database
collectionName	Name of a collection
fieldName	Name of a field
value	Value of a attribute
dbName.collectionName	Collection inside a persistent database
collectionName.fieldName	Field of a collection

3.1 GET COLLECTION

The GET COLLECTION operator takes a collection from a database (persistent or intermediate) and makes it the new temporary collection. The syntax of the operator is the following:

```
GET COLLECTION [dbName.]collectionName;
```

When the *dbname* is specified, the JSON collection named *collectionName* is retrieved from the persistent database named *dbname*; otherwise, it is retrieved from the intermediate result database *IR*.

Notice that the GET COLLECTION operator permits to start a query process on one single collection, by retrieving the collection to transform.

3.2 SAVE AS and SET INTERMEDIATE AS

Two operators are necessary to store the results. The SAVE AS operator saves the input temporary collection into a persistent database.

```
SAVE AS dbName.collectionName;
```

The name of the new collection stored into the persistent database *dbName* is *collectionName*. An example is shown in Example 2.

The SET INTERMEDIATE AS operator stores the input temporary collection into the intermediate results database *IR*. The syntax of operator is:

```
SET INTERMEDIATE AS collectionName;
```

Notice that *collectionName* is the name given to the new temporary collection into the intermediate results database.

3.3 DERIVE GEOMETRY

In this subsection we introduce two operators that are defined to deal with the geometric components of the JSON objects, i.e., GeoJSON objects. At this stage they only deal with the Point and MultiPoint geometry types. In the future we will extend these operators to manage the other more complex geometry types.

The DERIVE GEOMETRY operator adds the geometry field to each object o_i in the temporary collection, deriving it from other properties.

There are two version of DERIVE GEOMETRY operator. The first version has the following syntax:

```
DERIVE GEOMETRY POINT(latFieldName,lonFieldName);
```

In this first version of the operator, the values of the two specified fields play the role, respectively, of latitude and longitude. The new `geometry` field will contain a point (represented in GeoJSON format) with those coordinates.

The second version of the operator has the following syntax:

```
DERIVE GEOMETRY AGGREGATE (arrayName);
```

In this second version of the operator, for each object o_i in the temporary collection, the new field `geometry` is obtained by aggregating all the `geometry` fields of all objects appearing within the array field *arrayName* in o_i. In practice, it allows to derive a unitary geometrical representation of all objects grouped together after a `GROUP BY` operator (see Sect. 3.5).

3.4 FILTER

The `FILTER` operator permits to filter objects in the temporary collection, according to some selection conditions, and possibly changes the structure of selected objects. The operator is designed to deal with the heterogeneous nature of JSON collections; for this reason, the syntax is more articulated than other operators.

```
FILTER
  (CASE:
    (fieldName = value (,fieldName = value)* |
     WITH fieldName (,fieldName)* |
     WITHOUT fieldName (,fieldName)*)+
       [WHERE selectionCondition]
       [PROJECT fieldName (,fieldName)*])+
(KEEP OTHERS|DROP OTHERS);
```

Each `CASE` branch (at least one) specifies a subset of objects to select by means of a list of selectors. A selector could be: a `WITH` selector that asks for objects with the specified field name; a `WITHOUT` selector, that asks for objects without the specified field name; an equal conditions on fields. For any object in the input temporary collection that matches with the specified selectors, the `WHERE` clause, if specified, is evaluated and if it is false, the object is discarded, otherwise the object is kept and inserted into the output temporary collection. Finally, if the `PROJECT` clause is specified, the selected object is projected on the specified list of fields, in order to reduce the number of its fields.

`CASE` branches are evaluated in the order: an object is handled by the first branch that matches with it. If none of them matches, the object is kept into the output temporary collection if the `KEEP OTHERS` option is specified, while the object is discarded if the `DROP OTHERS` is specified.

Example 1 Consider collection *TouristTweets* shown in Listing 1, which represents some tweets posted by tourists. We are interested in tweets with origin airport equals to *Malpensa*. The query is hereafter.

```
GET COLLECTION MyDB.TouristTweets;
FILTER
  CASE: OriginAirport="Malpensa" WITH geometry
  PROJECT TweetId
  DROP OTHERS;
```

The GET COLLECTION operator retrieves the initial TouristTweets collection from the persistent database named MyDB; this collection becomes the new temporary collection. The subsequent FILTER operator works on this temporary collection.

Only one CASE branch is sufficient to our purpose, with two selectors: the first one is an equal condition on field OriginAirport; the second one is a WITH selector on field geometry: only objects having the geometry field and the OriginAirport field having value Malpensa are selected.

As a result, two objects in our sample TouristTweets collection will be selected, and then projected on the field TweetId. Other possibly not matching objects are dropped.

The new temporary collection produced by the operator is reported hereafter.

```
[{
  "TweetId":"603755043112759296"
 },
 {
  "TweetId":"603755043112759340"
 }]
```

Notice the very simple structure of the resulting objects.

3.5 GROUP BY

The GROUP BY operator groups objects in the input temporary collection based on a list of grouping fields. Here is its syntax.

```
GROUP BY fieldName (,fieldName)*
INTO fieldName
[SORTED BY fieldName (,fieldName)*];
```

The GROUP BY operator groups the objects in such a way a group contains all the objects $o_1, \dots, o_n$ having the same values for field names specified after the GROUP BY keywords.

For each group, an object g_k appears in the output temporary collection, such that it has all the grouping fields and a field vector containing objects $o_1, \dots, o_n$; the name of this field is specified in the clause INTO.

Finally, the optional clause SORTED BY specifies whether to sort objects into the vector fields. If so, the following field names are the sort keys.

Notice that the output temporary collection contains as many objects as the number of groups.

Example 2 Consider collection *TouristTweets* shown in Listing 1. We might be interested to know the tourist trips. The query is hereafter.

```
GET COLLECTION MyDB.TouristTweets;
GROUP BY UserName
  INTO Trip
  SORTED BY Date, Time;
DERIVE GEOMETRY AGGREGATE (Trip);
SAVE AS MyDB.AllTrips;
```

The `GET COLLECTION` operator retrieves the `TouristTweets` collection from the persistent database named `MyDB`, making it the new temporary collection, on which the `GROUP BY` operator works. Tourist tweets are grouped by attribute `UserName`; the name given to the vector field containing grouped objects is `Trip`. The vector field is sorted according to the fields `Date` and `Time`. The output will be the new temporary collection on which the `DERIVE GEOMETRY AGGREGATE` operator works. The `DERIVE GEOMETRY AGGREGATE` creates a field `geometry` representing the tourist trip, inside each grouped object, in order to create the user trip. Here is the output temporary collection that is made persistent by the `SAVE AS` operator with the name `AllTrips` into the persistent database named `MyDB`.

```
[{"UserName":"User1",
  "Trip":[
    {"TweetId":"603755043112759296",
     "UserName":"’User1",
     "geometry":{"type":"GeometryCollection",
                 "geometries":[
                    {
                     "type": "Point",
                     "coordinates": [45.63120, 90.289]
                    }]
                },
     "Date": "2015-05-28",
     "Time": "04:49:54",
     "StartDate": "2015-05-28",
     "OriginAirport": "Malpensa"
    },

    {"TweetId":"603755043112759340",
     "UserName":"’User1",
     "geometry":{"type":"GeometryCollection",
                 "geometries":[
                    {
                     "type": "Point",
                     "coordinates": [46.75813, 92.459]
                    }]
                },
     "Date": "2015-05-30",
     "Time": "10:50:43",
     "StartDate": "2015-05-28",
     "OriginAirport": "Malpensa"
    }],
    "geometry":{"type":"GeometryCollection",
                "geometries":[
                   {
                    "type": "LineString",
                    "coordinates": [[45.63120, 90.289],
                                    [46.75813, 92.459]]
                   }]
               }
  },
  {"UserName":"User2",
   "Trip":[
      {"TweetId":"603755043112759450",
       "UserName":"'User2",
       "geometry":{"type":"GeometryCollection",
                   "geometries":[
                      {
                       "type": "Point",
                       "coordinates": [49.75421, 95.322]
                      }]
                  },
   "Date": "2015-07-30",
   "Time": "11:30:52",
   "StartDate": "2015-07-29",
   "OriginAirport": "Linate"
 }],
 "geometry":{"type":"GeometryCollection",
             "geometries":[
                {
                 "type": "LineString",
                 "coordinates": [[49.75421, 95.322]]
                }]
            }
}]
```

The output contains an object for user *User1* (first object) and another object for user `User2` (second object). The first object contains a field `UserName` (grouping field) with value `User1`, a field `Trip` which contains all objects contained in the collection reported in Listing 1 with the field `UserName` equals to `User1` (grouped objects) and sorted according to the fields `Date` and `Time`; finally, field `geometry` represents the trip trajectory of user `User1`. The second object has the same structure but the grouping field is `User2`.

Example 3 As a final example of J-CO non-spatial operators, we show how to filter, from the `AllTrips` collection (obtained by the query in Example 2), only the trips that are longer than two legs, i.e., the objects in `AllTrips` containing more that two elements in the array field named `Trip`.

This is done by submitting the following J-CO query:

```
GET COLLECTION MyDB.AllTrips;
FILTER
 CASE:
   WITH UserName
     WHERE COUNT(Trip)>2
DROP OTHERS;
```

4 Spatial J-CO Operators

In order to correlate big geo-data, it is necessary to enrich the J-CO query language with spatial operators that permit to evaluate both metric and topological operations between the geometric fields of the JSON objects. Hereafter, we introduce two key operators.

4.1 SPATIAL JOIN

The first operator is the `SPATIAL JOIN` operator that makes the geo-spatial join between two input collections based on the truth of a metric or topological condition evaluated between the geometries of any pair of two objects from the two input collections. The metric conditions can be defined on the distance between two geometries, for example requiring that their distance is lower than a maximum threshold, or the intersection of their geometries that must be not empty or with an area greater than a specific value. The topological condition can be defined on the orientation of one geometry w.r.t. the other geometry, or that the two geometries share some part of their boundary, i.e., they meet, or that one geometry is covered (is included) or covers (includes) the second geometry.

The syntax of the `SPATIAL JOIN` operator is the following:

```
SPATIAL JOIN OF COLLECTIONS
[dbname.]collectionname1, [dbname.]collectionname2
[ON spatialJoinCondition]
[WHERE selectioncondition]
KEEP (INTERSECTION | RIGHT | LEFT | ALL);
```

The operator makes the join between two collections if a metric or topological condition defined on their geometries is satisfied and the optional selection condition in the `WHERE` clause on non geometric attributes is satisfied too. The result becomes the new temporary JSON collection. The input collections can come from two persistent, possibly distinct, databases or from the intermediate result database *IR*.

For each object l_i in the left collection (*dbname.collectionname1*) and an object r_j in the right collection (*dbname.collectionname2*), an object $o_{i,j}$ appears in the output collection if the geometries of l_i and r_j meet the *spatialJOinCondition* (expressed in the `ON` clause either on a metric relationship or on a topological relationship between the two geometries), and the optional *selectionCondition* (expressed in the `WHERE` clause on non-geometric fields) is met as well. The output object $o_{i,j}$ has three fields: one with the name of the left collection that contains object l_i, one with the name of the right collection that contains object r_j, a `geometry` field.

Metric and topological relationships are expressed by the following pre-defined properties:

- DISTANCE (*unit*) is the distance between the centroids of the two geometries, expressed based on the specified *unit*, that can be M (meters), KM (Kilometers), ML (miles).
- AREA (*unit*) is the area of the intersection of the two geometries expressed based on the specified *unit*, that can be M (square meters), KM (square Kilometers), ML (square miles).
- ORIENTATION (*from*) reports the cardinal orientation of l_i . geometry w.r.t. r_j . geometry; If the *from* parameter is LEFT (resp. RIGHT), it is the orientation of the spatial vector with origin in the centroid of l_i . geometry and directed to the centroid of r_j . geometry or vice versa; orientation values are strings obtained by composing the 4 letters N (for North), E (for East), S (for South) and W (for West): a single letter (e.g., "N" for North orientation), a pair (e.g., "NE" for North-East), a triple (e.g., "NNE" for North-North-East). These 16 orientation values identify distinct sectors partitioning the round angle having equal width (22 30'). Thus the evaluation of the cardinal orientation of a spatial vector is implemented as its inclusion into a sector of the round circle centered in the origin of the vector.
- INCLUDED (*side*) is a boolean property that denotes the inclusion of
  l_i. geometry w.r.t. r_j . geometry;
  if the *side* parameter is LEFT, INCLUDED(LEFT) is true if l_i . geometry is completely included in r_j . geometry;
  if the *side* parameter is RIGHT, INCLUDED(RIGHT) is true if r_j . geometry is completely included in l_i . geometry.
- MEET is a boolean property, that is true when the two geometries share a common part of their boundaries.
- INTERSECT is a boolean property that is true when the two geometries intersect.

The KEEP parameters allows specifying the geometry of the output object $o_{i,j}$: when KEEP INTERSECTION is specified, $o_{i,j}$. geometry is the intersection of the geometries of the joined objects; when KEEP RIGHT (resp. KEEP LEFT) is specified, $o_{i,j}$. geometry is the geometry of the right (resp., left) input object; when KEEP ALL is specified $o_{i,j}$. geometry is the union of the geometries of the input objects.

Let us make some examples of possible utility of the SPATIAL JOIN.

Example 4 Let us consider an open data collection named POIs (stored in the persistent database named MyDB), containing geo-referenced descriptions of territorial *Points Of Interest*, i.e., POIs, such as cultural and earth-heritage sites.

The analysis intent could be to semantically enrich the POIs description with social network messages (stored in collection TouristTweets, reported in Listing 1) that are sent in their immediate neighborhood, or within the boundary of their geo-reference, in order to associate, with each POI, messages of tourists that may have visited the same POI and possibly expressed some opinion about it. This operation can aid the analysis of tweets in the proximity of POIs, for understanding the appreciation of the POIs among their visitors.

This analysis can be easily performed by submitting the following Spatial JOIN query:

```
SPATIAL JOIN OF COLLECTIONS MyDB.POIs, MyDB.TouristTweets
  ON DISTANCE(KM) < 0.3
  KEEP ALL;
```

where a POI is joined with a tweet if the distance between the centroid of the POI and the geo-reference of the tweet is less than 0.3 km. Alternatively the following query joins a POI with a tweet if the geo-reference of the tweet is included within the area occupied by the POI (e.g., within a garden POI).

```
SPATIAL JOIN OF COLLECTIONS MyDB.POIs, MyDB.TouristTweets
  ON INCLUDED(RIGHT) = true
  KEEP ALL;
```

Example 5 As another example, we want to semantically enrich tweets in collection `TouristTweets` (reported in Listing 1) with the open data collection `OpenZIPs` collection of the zip code areas:

```
SPATIAL JOIN OF COLLECTIONS MyDB.OpenZIPs, MyDB.TouristTweets
  ON INCLUDED(RIGHT) = true
  KEEP ALL;
```

This way, all tweets are associated with the zip code of the area where they have been sent from.

Example 6 A final example of spatial join query allows joining tweets posted by the same user and sent from localities that denote a movement of the user from south to north. The following query could be formulated

```
GET COLLECTION MyDB.TouristTweets;
SET INTERMEDIATE AS Tweets;
SPATIAL JOIN OF COLLECTIONS  MyDB.TouristTweets, Tweets
  ON ORIENTATION(LEFT) = "N"
  WHERE TouristTweets.UserName = Tweets.UserName AND
        TouristTweets.time < Tweets.time
  KEEP  ALL;
SAVE AS MyDB.Northmovements;
```

This query exhibits a certain complexity, motivated by the need of joining collection `TouristTweets` with itself. For this reason, the `GET COLLECTION` operator retrieves the collection and the subsequent `SET INTERMEDIATE AS` operator stores it into the intermediate result database *IR* with name `Tweets`.

At this point, the `SPATIAL JOIN` operator can join the original collection `TouristTweets` with its copy named `Tweets`. The join is performed in such a way tweets in the left collection temporally precede tweets in the right collection and are sent by the same user (WHERE clause), as well as the orientation from the left tweet to the right tweet is to north (ON clause). Geometries of both joined tweets constitute the geometry of the output object (KEEP ALL option).

Finally, the output collection is stored with name Northmovements into the persistent database MyDB by the SAVE AS operator.

4.2 SPATIAL FILTER

Another useful spatial operator is the one that allows to spatially filter JSON objects based on the similarity of their geometries to a given target reference geometry. This is the SPATIAL FILTER operator and its syntax is reported hereafter.

```
SPATIAL FILTER
 WRT geometrySpec
   FOR (ATLEAST n | HALF | ALL)
 WHERE DISTANCE(unit) compOp value
   FOR (ATLEAST n | HALF | ALL);
```

This geo-spatial filtering evaluates the distance of the geometry of each object in the input collection w.r.t. the geometry *geometrySpec* (specified in the WRT clause); it creates an output collection containing an object for each object in the input collection if points (all or part of them) in its geometry satisfy a selection condition based on the DISTANCE of points from *geometrySpec*. *compOp* can be any comparison operator, i.e., <, <=, >, >=, = and <>, but we expect that conditions will be mostly based on < or <=.

The two FOR sub-clauses (in the WRT clause and in the WHERE clause) specify for how many points in the *geometrtySpec* and in the geometry of the object o_i the selection condition in the WHERE clause must be satisfied.

- If FOR ALL is specified, the condition must be satisfied by all points either in *geometrySpec* or in o_i. geometry;
- if ATLEAST n (with $n \geq 1$) is specified, the condition must be satisfied by at least n points either in *geometrySpec* or in o_i. geometry;
- if HALF is specified, the condition must be satisfied by at least half of points either in *geometrySpec* or in o_i. geometry;.

A possible use of the SPATIAL FILTER operator is to discover trajectories that mostly pass close to a given set of POIs. We show this in the following example.

Example 7 Consider three typical POIs in Lombardy, i.e., *Duomo di Milano* (coordinates (45.464105, 9.191916)), the city of Como (coordinates of the center are (45.809709, 9.084462)) and the city of Bergamo (coordinates of the old city center are (45.704106, 9.662772)). If we would like to discover which travelers visit at least two of these three places, the query could be the following one.

```
GET COLLECTION MyDB.AllTrips;
SPATIAL FILTER
 WRT POINTS( (45.464105, 9.191916),
             (45.809709, 9.084462),
             (45.704106, 9.662772))
                       FOR ATLEAST 2
 WHERE DISTANCE(KM) < 1
  FOR ATLEAST 2;
```

The query takes the collection `AllTrips` calculated by Example 2 and looks for those trips with at least two points which are less than 1 km distant from two of the three specified points. In other words, the query looks for trips that visited the three city centers reported in the `WRT` clause. Notice that, in order to be sure that at least two POIs are visited, the `FOR ATLEAST` 2 sub-clause must be specified twice.

5 Related Work

The proposal is strictly related with the so called *Social Media Geographic Information* (SMGI) analytics Campagna et al. (2015). It is well known that there is a lack of shared analytical frameworks to collect, manage and process SMGI for different purposes. As far as free and open tools for analyzing social network posts the plug-in for Microsoft excel named *NodeXl*, developed in association with the Social Media Research Foundation, has been conceived to provide easy-to-use analysis capabilities Hansen et al. (2010), Luo and MacEachren (2014). It permits to represent the graphs of social relationships between posts in a visual form, as well as to compute summary metrics of these relationships. Excel offers the possibility to correlate such data with imported CSV file (CSV files are commonly used to represent flat open data, in alternative to JSON). Nevertheless, it is not possible to analyze the geographic fields and perform spatial analysis. A common practice to overcome this limitation is to import the SMGI and the open data into a GIS, as separate vector layers and, then, to perform spatial analysis by using the spatial operations on vector layers provided by GISs. However, if one wants to perform complex spatial operations such as grouping SMGI based on the similarity of their geo-reference, trajectory filtering and matching, semantic enrichment, one needs to know very well the spatial analysis tools of the GIS, which need an expert user.

Our proposal is also related with geo-social visual analytics of SMGI. Visual analytics methods can be classified into three groups, namely, data exploration, decision-making, and predictive analysis. For a survey of the visual analytics methods see Luo and MacEachren (2014). Nevertheless, a SMGI exploration phase can be visually supported but cannot avoid to provide querying facilities to filter and correlate data.

NoSQL database management systems (see Han et al. 2011 for an survey) for JSON data (and MongoDB in particular) revealed to be adequate tools to support SMGI and open geo-data spatial analysis, since they allow to manage heterogeneous objects as GeoJSON objects. however, query languages of NoSQL databases are difficult to use by not experts. This further motivated the definition of the J-CO language.

We took inspiration by the work in Bordogna et al. (2006), Psaila (2011), where the authors addressed the problem of querying heterogeneous collections of complex spatial data. Those works propose a database model for dealing with heterogeneous collections of possibly nested spatial objects, based on the composition of more primitive spatial objects; furthermore, a relational-style algebra is defined, to query complex spatial data. On the contrary, J-CO is based on the JSON standard, as well as it relies on a more flexible and intuitive execution model.

As far as query languages for JSON objects are concerned, a close proposal is *Jaql* Nayak et al. (2013) that was designed to help Hadoop programmers writing complex transformations, avoiding low-level programming White (2012). Although *Jaql* relies on the concept of pipe, that recalls the J-CO execution model, it is still oriented to programmers while J-CO constructs are at a higher level and truly declarative.

Another interesting language is *Pig Latin* Olston et al. (2008), a query language developed by Yahoo for writing complex analysis tasks on nested (1-NF, first normal form) data sets on top of Hadoop; thus, JSON collections are implicitly included. Although Pig Latin's constructs are similar to J-CO constructs, it still relies on the concept of variable. In contrast, the J-CO execution model privileges the process flow and the language provides high level spatial operators, specifically designed for non programmer users.

A third proposal is SQL++ has been defined to query both JSON native stores and SQL databases. The SQL++ semi-structured data model is a superset of both JSON and the SQL data model. Yet, SQL++ is SQL backwards compatible and is generalized towards JSON by introducing only a small number of query language extensions to SQL Ong et al. (2014).

Also the industry is looking at the extension of SQL to query JSON objects. An example is N1QL (http://www.couchbase.com/n1ql) that is a declarative language extending SQL for JSON objects stored in NoSQL databases, specifically implemented for Couchbase 4.0, in order to handle semi-structured, nested data. It enables querying JSON documents without any limitations sort, filter, transform, group, and combine data with a single query from multiple documents with a JOIN. Nevertheless it does not provide operators to manipulate GeoJSON objects.

Finally, other declarative languages for JSON objects have been defined as extensions of structured languages for semi-structured documents, such as JSONiq (http://www.jsoniq.org/) that borrowed a large numbers of ideas from XQuery, like the functional aspect of the language, the semantics of comparisons in the face of data heterogeneity, the declarative, snapshot-based updates. However, unlike XQuery, JSONiq is not concerned with the peculiarities of XML, like mixed content, ordered children, or the complexities of XML Schema, and so on. Nevertheless, like XQuery it can be hardly used by unexperienced users.

6 Conclusions

The paper proposes an initial set of operators to define a declarative query language able to flexibly transforming and querying heterogeneous collections of georeferenced data represented as JSON objects. Such collections could be open geodata and social network information, that together could be the source of interesting and unexpected information about habits of citizens. Currently, GISs do not provide declarative query languages guaranteeing the independence of the data, so the users must be aware of the data structures, and specifically of the geometric dimension representation in order to perform the analysis. Our proposal is a first attempt to leverage the spatial analysis by providing not-experts with high level spatial operators that encapsulate methods working on complex geometric fields. In fact, the J-CO query language is thought as a plug-in into GISs, so as to extend typical multilayer visualization capabilities of GISs with the features provided by a declarative and high-level query language on complex and geo-tagged data sets.

In the future, the proposed language will be enriched with further operators, mainly for dealing with complex geometric GeoJSON objects such as multi lines and multi polygons, which can be chained into execution pipelines to perform complex spatial analysis.

References

Banker K (2011) MongoDB in action. Manning Publications Co

Bordogna G, Pagani M, Psaila G (2006) Database model and algebra for complex and heterogeneous spatial entities. In: Progress in spatial data handling. Springer, pp 79–97

Butler H, Daly M, Doyle A, Gillies S, Hagen S, Schaub T (2016) The geojson format. Technical report

Campagna M, Floris R, Massa P, Girsheva A, Ivanov K (2015) The role of social media geographic information (SMGI) in spatial planning. In: Planning Support systems and smart cities. Springer, pp 41–60

Han J, Haihong E, Le G, Du J (2011) Survey on NoSQL database. In: 2011 6th international conference on pervasive computing and applications (ICPCA). IEEE, pp 363–366

Hansen D, Shneiderman B, Smith MS (2010) Insights from a connected world. Analyzing social media networks with NodeXL. Morgan Kaufmann

Kumar S, Morstatter F, Liu H (2013) Twitter data analytics. Springer

Luo W, MacEachren AM (2014) Geo-social visual analytics. J Spat Inf Sci 2014(8):27–66

Nayak A, Poriya A, Poojary D (2013) Type of NOSQL databases and its comparison with relational databases. Int J Appl Inf Syst 5(4):16–19

Olston C, Reed B, Srivastava U, Kumar R, Tomkins A (2008) Pig latin: a not-so-foreign language for data processing. In: Proceedings of the 2008 ACM SIGMOD international conference on management of data. ACM, pp 1099–1110

Ong KW, Papakonstantinou Y, Vernoux R (2014) The SQL++ semi-structured data model and query language: a capabilities survey of sql-on-hadoop, nosql and newsql databases. CoRR. arXiv:1405.3631

Psaila G (2011) A database model for heterogeneous spatial collections: definition and algebra. In: 2011 international conference on data and knowledge engineering (ICDKE). IEEE, pp 30–35

Wakamiya S, Belouaer L, Brosset D, Lee R, Kawai Y, Sumiya K, Claramunt C (2015) Measuring crowd mood in city space through twitter. In: International symposium on web and wireless geographical information systems. Springer, pp 37–49

White T (2012) Hadoop: the definitive guide. O'Reilly Media, Inc

Towards Automatic Large-Scale 3D Building Reconstruction: Primitive Decomposition and Assembly

Hai Huang and Helmut Mayer

Abstract In this paper we propose a pipeline for highly automatic building reconstruction based on 3D point clouds. 3D building models are of great interest for many applications including city planning, navigation, emergency response and tourism and their reconstruction has been intensively studied. It is, however, still a challenge to minimize manual intervention and to achieve highly automated processing in practical applications. The main reason lies in the variability and complexity of urban scenes. We believe that one possible key to tackle this is a reliable primitive-based decomposition of urban scenes as well as their constituent buildings. It links scene interpretation with model reconstruction and, thus, naturally completes an automatic reconstruction pipeline. We propose an effective scheme for the decomposition of the whole scene straight into individual building components, i.e., primitives. The primitives are reconstructed via statistical generative modeling and assembled into individual watertight building models. An experiment has been performed on a dataset of a complete central European village demonstrating the potential of the proposed approach.

Keywords 3D reconstruction · Point cloud · Building · Statistical modeling

1 Introduction

The automatic generation of 3D building models from remote sensing data is of great interest for many applications including city planning, navigation, emergency response and tourism. Many approaches have been reported in the past decades. Overviews are given in Brenner (2005), Schnabel et al. (2008), Vosselman (2009). Current work includes (Sampath and Shan 2010), which segments and reconstructs

H. Huang (✉) · H. Mayer
Institute for Applied Computer Science, Bundeswehr University Munich,
Werner-Heisenberg-Weg 39, 85577 Neubiberg, Germany
e-mail: hai.huang@unibw.de

H. Mayer
e-mail: helmut.mayer@unibw.de

A. Bregt et al. (eds.), *Societal Geo-innovation*, Lecture Notes
in Geoinformation and Cartography, DOI 10.1007/978-3-319-56759-4_12

complicated buildings from airborne LIDAR point clouds based on polyhedral models. Starting from planar roof segments, Zhou and Neumann (2012) try to organize them using "global regularities" in the form of orientation and placement constraints. Lafarge et al. (2010) present building reconstruction from a Digital Surface Model (DSM) combining generic and parametric methods. For more sophisticated buildings, basic geometric primitives, e.g., planes, cylinders and cones, are combined with mesh-patches to present irregular roof forms (Lafarge and Mallet 2012). Huang et al. (2013a) propose a statistical approach for building model reconstruction from LIDAR data via generative models. Partovi et al. (2015) present an extension of a hybrid framework for data from stereo satellite imagery with ridge-line-based building mask decomposition.

Approaches for building footprint decomposition can be divided into two categories: With or without primitive overlap. Brenner and Haala (2000) propose a flexible decomposition scheme resulting in overlapping rectangular primitives. Lafarge et al. (2010) present a decomposition of given building footprints as adjacent primitives, which are not limited to rectangular shapes. Current work includes (Wang et al. 2015), which presents building decomposition based on the adjacency graph of detected planar roof patches and a primitive-based reconstruction.

In recent years, quality and availability of 3D point clouds from LIDAR and image matching have been significantly improved and some approaches reach a high level of automation (Lafarge and Mallet 2012; Huang et al. 2013a) and cover large suburban areas. There are, however, still several challenges towards fully automatic city model reconstruction. One of them is the parsing of complex scenes as well as buildings. This limits the reconstruction of larger urban areas and, thus, renders a pipeline to build a whole city model unreliable. We believe, that the key to tackle this deficit is a reasonable decomposition subdividing the whole scene as well as heterogeneous buildings into regular components.

As shown in Fig. 1, this paper extends and links our approaches to scene classification (Huang and Mayer 2015) and primitive-based reconstruction (Huang et al. 2011, 2013a) with a pre-processing—decomposition and a post-processing—assembly of primitives to complete a practical reconstruction pipeline. A reasonable and reliable decomposition of the whole scene as well as of complicated buildings into regular primitives precedes model reconstruction. The assembly of the primitives performs a true model merging in CAD (Computer Aided Design) style. The decomposition works as an intermediate step linking the preceding scene interpretation with the following model reconstruction and is, therefore, a crucial part to complete an automatic reconstruction pipeline. The building mask from scene classification is used as input along with the 3D point cloud derived from dense image matching. Scene decomposition splits individual buildings from the building mask while building decomposition further disassembles building complexes into simple building components. The latter are represented by predefined 3D primitives and reconstructed via statistical generative modeling. The primitives are subsequently assembled into individual watertight building models via CSG (Constructive Solid Geometry) modeling.

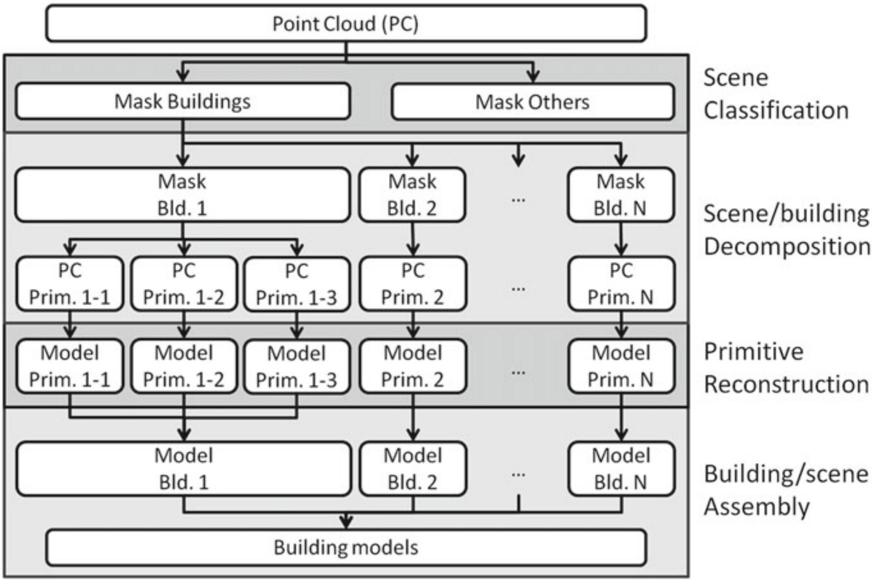


Fig. 1 Pipeline for automatic building reconstruction

This paper is organized as follows: Sects. 2 and 3 present scene and building decomposition, respectively. The statistical generative modeling of building primitives is given in Sect. 4. Section 5 describes the assembly of primitives into complete building models. Experimental results are given in Sect. 6. The paper ends up with the conclusion in Sect. 7.

2 Scene Decomposition

The goal of scene decomposition is to extract individual buildings from a (binary) building mask of the whole scene. The building mask (Fig. 2, left) is derived by previous scene classification, which may contain labeling errors often affecting the separation of buildings (middle). A mathematical morphological "opening" operation is conducted, as shown in Fig. 2 (right), to remove trivial areas and to better isolate the buildings. The radius of the disk-shaped structuring element is determined based on data quality as well as resolution. For the presented dataset with 0.2 m resolution, a radius of 1 m is employed for the structuring element.

Individual buildings are found via "blob" detection. As shown in Fig. 3, the input data are then correspondingly segmented into rectangular tiles. The segmentation is conducted with a buffer, so that certain classification errors can be tolerated. Overlapping between tiles is allowed to make sure buildings are completely included in

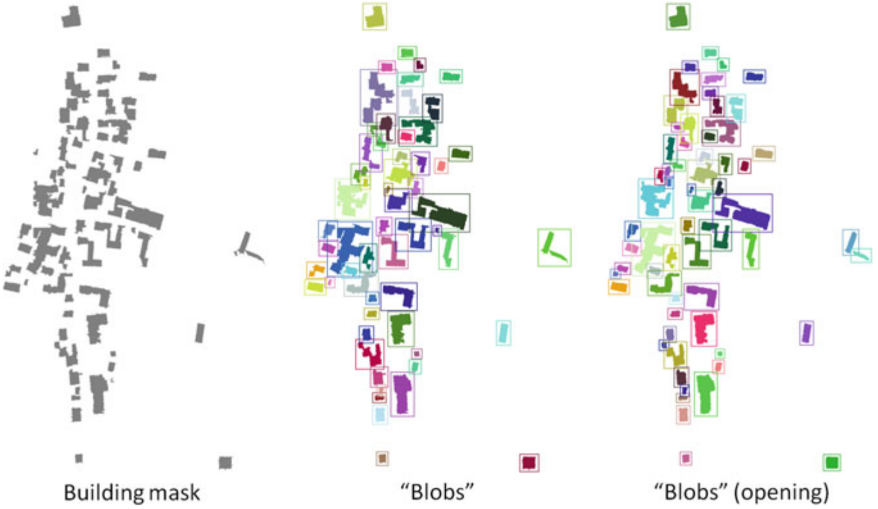


Fig. 2 Scene decomposition and individual building detection: comparison of building detection (as colorful "blobs") without (*middle*) and with (*right*) morphological opening on the input building mask (*left*)

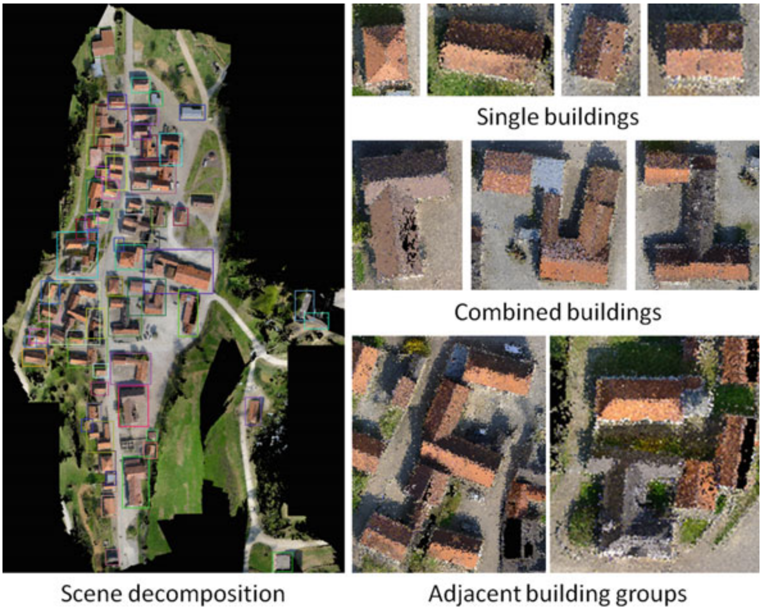


Fig. 3 Scene decomposition into tiles, which may contain individual or multiple buildings

the tiles. Global coordinates including the height (for undulating areas, cf. Sect. 6) are attributed to each tile for the final model assembly of the whole scene.

It, however, cannot be guaranteed that each "blob" contains exactly one individual building. This is due to the complexity of the scene and the imperfection of the classification i.e., after the decomposition, a data tile may contain a single building, but also a combined building (a building consisting of multiple building components), or multiple buildings which are closely adjacent to each other (cf. Fig. 3). The last case can often be found in densely inhabited areas. In this paper we call both, combined buildings and adjacent building groups, "building complexes". Further parsing of building complexes is described in the following section.

3 Building Decomposition

Generally, the goal of building decomposition is to divide building complexes into simple standard building components making the following model reconstruction easier and more efficient. This is especially true for primitive-based reconstruction methods, where the building components are directly represented by predefined building primitives.

Please note that building decomposition actually does not work on building models but directly on the input data, because the former does not exist yet. Conventionally, the decomposition is conducted bottom-up based on 2D footprints that are either already available (Lafarge et al. 2010; Kada and McKinley 2009) or extracted from the data (Partovi et al. 2015). This becomes, as shown in Fig. 4c, infeasible when 3D information has to be taken into account. Adjacent building components with different 3D geometry cannot be separated if they have similar width. This error will affect the following model reconstruction.

We propose a combined bottom-up and top-down scheme for the decomposition of building complexes. It works based on 3D geometry parsing with the support of

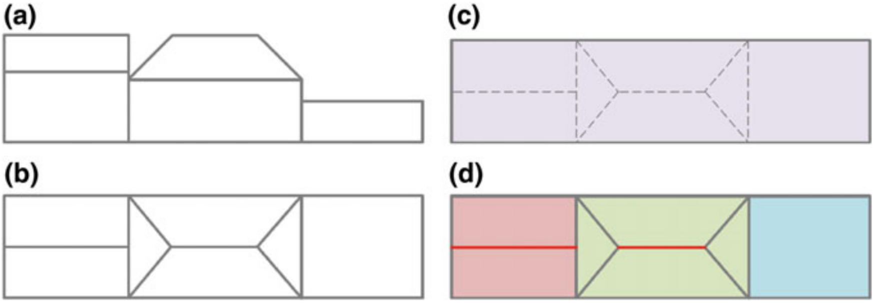


Fig. 4 Footprint- and ridge-based building decomposition: Side-view (**a**) and top-view (**b**) of a building complex and the decomposition based on footprint (**c**) and 3D geometry (**d**) using ridge (*red*) parsing and height differences

a predefined primitive library. As shown in Fig. 4d, the decomposition is improved by using 3D information including height differences and roof shapes.

We are aware that the strategies of the decomposition and the following model reconstruction have to match i.e., they should share the same construction concept for buildings and be adapted to each other, so that the pipeline works smoothly and the degree of automation is improved. There are two basic strategies for building (footprint) decomposition. The key difference is if the final building components are allowed to overlap (Brenner and Haala 2000; Huang et al. 2011) or not (Lafarge et al. 2010; Kada and McKinley 2009). Different primitive definitions are correspondingly employed. The concept allowing for overlaps fits better to the generative modeling process described in Huang et al (2013a). It is more flexible and has the potential to keep the model complete and regular with a reasonable size for the primitive library. No special primitives such as additional joint parts (Lafarge et al. 2010; Kada and McKinley 2009) are required. Besides, allowing primitive overlaps can tolerate a certain extent of uncertainty as result of the stochastic search during model reconstruction (cf. Sect. 4).

3.1 Ridge Lines

The ridge line is one of the key geometric features of many buildings. In approaches for roof interpretation ridge lines often play an important role, especially for those that use adjacency graph models (Elberink and Vosselman 2009; Huang and Brenner 2011). It has been demonstrated that the ridge line is not only the most significant, but also the most stable feature for 3D building roof model reconstruction.

Ridge lines are also used to improve the bottom-up decomposition performance (Arefi et al. 2010; Partovi 2015). The ridges are mostly treated as a building's central line and extracted from 2D contours or simple height information (line fitting to the points with a given height limit for ridges). Alternatively, "skeletons" derived from 2D footprints are often employed for building structure analysis (Haunert and Sester 2008, 2013b). Decomposition, however, becomes more difficult when the complexity of buildings increases.

We propose a top-down primitive-based decomposition scheme with emphasis on complicated building structures. The ridge lines are extracted and divided into straight line segments, which can be seen as ridges of individual primitives and, thus, guide the decomposition. In comparison with related work, the ridge lines are much more precisely extracted fully 3D by plane detection and intersection instead of an approximation by means of fitting lines to candidate ridge points (Arefi et al. 2010) or a morphological operation (Partovi et al. 2015). With the focus on the degree of automation, we concentrate on buildings with regularly structured components, which make up the majority of urban areas, instead of atypical or landmark buildings. The latter are rare and manual intervention in the reconstruction is mostly unavoidable anyway.

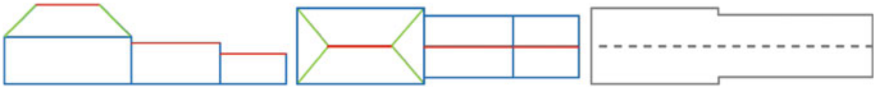


Fig. 5 Ridges and building skeleton—*Horizontal* ridge (*red*), diagonal ridges (*green*), eaves (*blue*), and skeletons (*gray dashed*) derived from 2D footprint (*gray*)

We define, as shown in Fig. 5, the following types of edges on the roof: (1) Horizontal ridge line (red), which connects two apexes of the roof, (2) diagonal ridge line (green) connecting one apex and one eave corner, and (3) eave line (blue), which links two eave corners. The roof contour consisting of eave lines can be used to approximate the building footprint when the overhang of the eave is ignored.

The ridge lines proposed in this paper have the following advantages as basis of building decomposition:

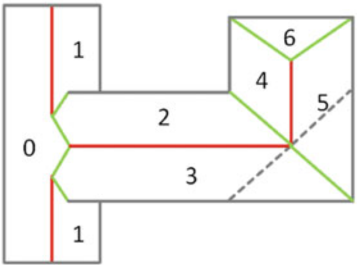
1. No additional footprint data is required. Ridge lines are derived directly from the input point cloud by means of plane intersection, which is much more accurate than conventional approximation methods.
2. With full 3D parsing more specific and accurate (e.g., asymmetric roofs, different roof heights and types) geometrical information is available.

For flat roofs, which do not have ridges, central lines are used instead. In comparison with building skeletons, central lines of roofs are still 3D i.e., as shown in Fig. 5, they have a height value and in the case of adjacent flat roofs they can differentiate multiple building components by means of their heights.

For non-flat roofs, ridge lines are found in the form of the intersection lines of the individual planes of the roofs. Planes of a building complex are detected by RANSAC. By this means, all intersection lines are actually determined via a consensus of all data points of both planes i.e., the intersection lines are much more reliable and precise.

To separate ridges from other kind of intersection lines, we employ the "relation matrix" (Huang and Brenner 2011) shown in Fig. 6. After the planes have been detected, the points that lie in the intersection area are counted and the numbers are entered in the relation matrix. The intersection area is defined based on the intersection line along with a buffer range, which is empirically determined proportionally to the point resolution.

"False" intersection lines are often found (cf. Fig. 6, dashed gray lines, planes 2–5 and 3–4), because the planes are actually infinite with no boundaries defined. An intersection line is verified by checking the normal directions of the planes on its both sides. Similar normal directions imply that the intersection line lies inside one slope of the roof and is,therefore, "false". The corresponding cell in the relation matrix is then set to null. The relation matrix is symmetric. For the purpose of illustration we, as shown in Fig. 6 (right), use the lower triangle (gray) to show the original numbers of intersection points while the upper triangle (colored) presents the verification of horizontal (red) and diagonal (green) ridges.



	0	1	2	3	4	5	6
0		*78*	*27*	*25*	*0*	*0*	*0*
1	78		*22*	*24*	*0*	*0*	*0*
2	27	22		*98*	*39*	*0*	*0*
3	25	24	98		*0*	*42*	*0*
4	0	0	39	35		*39*	*40*
5	0	0	35	42	39		*41*
6	0	0	0	0	40	41	

Fig. 6 Ridge line determination with relation matrix: *Horizontal* ridges (*red*) and diagonal ridges (*green*)

3.2 Primitive-Based Decomposition

A top-down building decomposition is proposed based on a predefined primitive library given in Huang et al. (2013b). Figure 7 presents the process of building decomposition guided by the ridges. The detected horizontal ridge lines (Fig. 7b, red) are decomposed into straight line segments (Fig. 7c, bold). The primitives have a rectangular contour as well as a single straight horizontal ridge line (except flat

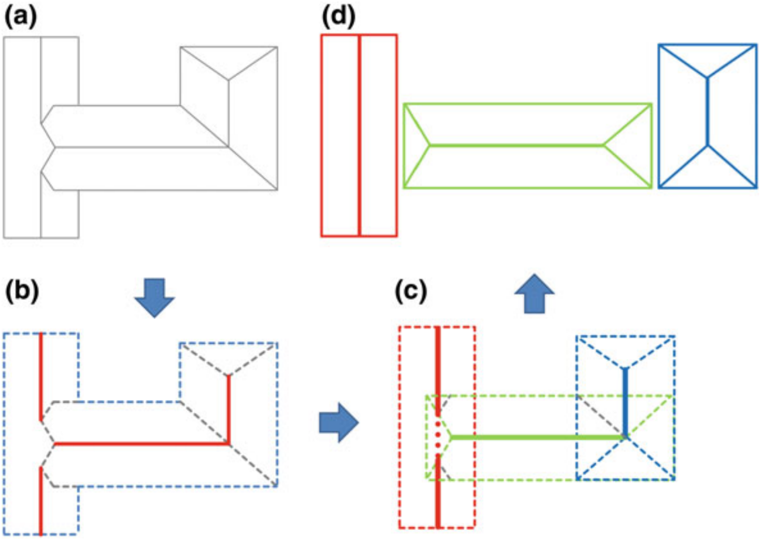


Fig. 7 Building decomposition: **a** Underlying model, **b** detected ridge lines (*red*), **c** decomposition based on ridge segments (*solid lines*) with completion using edges of the primitives (*dashed lines*), and **d** the decomposed primitives

roof and shed roof). The end points of the segments are determined by intersection with diagonal ridges or the boundary of the building mask.

Using the horizontal ridges as bases (cf. Fig. 7c), the most appropriate primitives are statistically selected from the library. The goal is a fit (1) to the already extracted diagonal ridges and (2) to the rest of the edges, without conflicts with the known planes and the boundary of the building mask. Again, this step decomposes no actual building model but underlying models, because the former does not yet exist. The goal is to define primitives that compose the building instead of modeling them i.e., the decomposition determines the number and types of primitives and the way of their combination (Fig. 7d). The concrete parameters of the primitives are calculated in the following primitive reconstruction (Sect. 4).

The primitives are, however, not the only information that can be derived from the building decomposition. The ridge line is a key component and plays an important role for the roof geometry. Initial values of the following primitive parameters can be derived solely from known horizontal ridges:

- Centroid coordinates (x and y) are defined by the center of the ridge line.
- Orientation (*azimuth*) is that of the ridge line.
- Ridge height (z_2) of the roof.
- Length is approximated proportionally to that of the ridge line.

Furthermore, the following initial values are obtained in combination with the known building mask. In the case that the building mask is not available, e.g., for building decomposition without previous scene decomposition, the diagonal ridges can be used instead:

- Area is approximately proportional to the mask area or the number of data points (for raster data).
- Width is derived from the area and length.
- Depth of hips (hip_{l1}, hip_{l2}, hip_{d2}, and hip_{d2}) are the longitudinal and radial distances from the end points of the horizontal ridge to the boundary of the building mask.

Since the ridge lines are precisely determined (cf. Sect. 3.1), the derived initial values are, to a certain extent, reliable and specific. They can, therefore, significantly improve the performance of the statistical search (cf. Sect. 4).

4 Primitive Reconstruction

We propose a generative primitive-based reconstruction, which extends the approach described in Huang et al. (2013a). The primitive parameters θ are defined as:

$$\theta \in \Theta; \Theta = \{\mathcal{P}, \mathcal{C}, \mathcal{S}\}, \tag{1}$$

where the parameter space Θ consists of position parameters $\mathcal{P} = \{x, y, azimuth\}$, contour parameters $\mathcal{C} = \{length, width\}$ (rectangle footprint), and $\mathcal{S}$ containing shape parameters, i.e., ridge/eave height and the depths of hips.

The goal of statistical reconstruction is to optimize the parameters fitting the primitive to the data. The Maximum A Posteriori (MAP) estimate of Θ can be expressed as

$$\widehat{\Theta}_{MAP} = \underset{\Theta}{argmax}\left\{ \frac{L(\mathcal{D}|\Theta)p(\Theta)}{P(\mathcal{D})} \right\} = \underset{\Theta}{argmax}\left\{ L(\mathcal{D}|\Theta)p(\Theta) \right\}, \tag{2}$$

where $L(\mathcal{D}|\Theta)$ is the likelihood function presenting the goodness of fit of the model to the data $\mathcal{D}$ and $p(\Theta)$ presents the prior for Θ, which is derived from empirical knowledge and incrementally improved during the reconstruction i.e., the parameter values of the already found building components or of adjacent buildings are used to update the priors. $P(\mathcal{D})$ is the marginal probability, which is regarded as a constant in the optimization as it does not depend on Θ.

The statistical optimization of the parameters is driven by reversible jump Markov Chain Monte Carlo with model selection in the transition kernel. Multiple hypothetic models ("candidates") are generated via statistical sampling of the primitive type as well as the corresponding parameters and evaluated based on the given 3D point cloud. The final model is the verified candidate model with the best goodness of fit to the data. Markov Chain Monte Carlo is employed for an efficient exploration of the high-dimensional (determined by the number of parameters) search space and the reversible jump mechanism is used for switching between different search spaces, i.e., different types of primitives. By these means, the statistical optimization including the change of primitive types is fully automatic. In comparison to Huang et al. (2013a), the search spaces are of the same size, but due to the more reliable initial values (cf. Sect. 3) the search entropy is much lower i.e., the computational effort is significantly reduced (cf. Sect. 6) by the proposed building decomposition before the reconstruction.

3D point clouds from image matching may contain data flaws such as false color and false positions of points. One important issue are gaps in the data, which often occur on surfaces with homogeneous color/texture (i.e., no matching points) and in case of occlusions. It leads to (cf. Fig. 8, left) missing points—holes on the roofs. Please note that such data gaps also exist in LIDAR data in areas where reflections, e.g., on glass surfaces and water bodies, occur. Conventional bottom-up methods may encounter difficulties in this case resulting in irregular and/or incomplete building components. Complete and watertight building models can, therefore, not be guaranteed. As shown in Fig. 8, the proposed method is robust despite such data flaws and ensures plausible results.

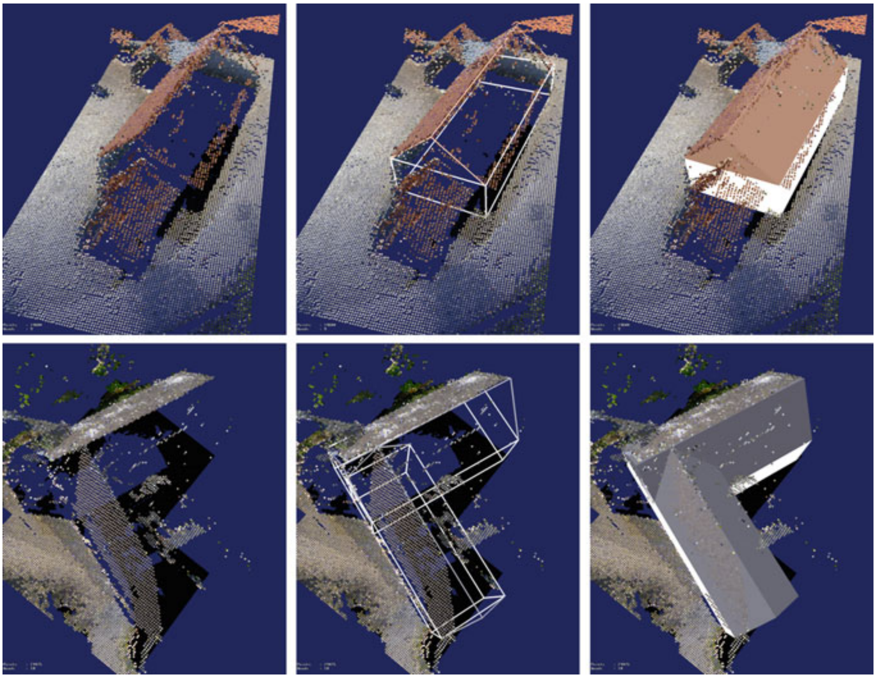


Fig. 8 Robust reconstruction despite data flaws: the input point clouds (*left*), detected primitives shown as wireframes (*middle*), and final building models (*right*)

5 Assembly of Buildings

The reconstructed primitives of a building complex are assembled into a single model. The modeling methods used for buildings can be categorized in two concepts (Brenner 2004): Surface modeling, also known as boundary representation (B-Rep) and solid body modeling, which is mainly based on CSG (Mäntylä 1987). In B-Rep modeling buildings are described by their bounding surfaces and their relationship of intersection. CSG is widely employed in CAD tools. Complicated models are represented by multiple volumetric primitives combined using Boolean operators. Both of them are employed in this paper for building assembly, which consists of two consecutive parts: (1) Joint parametric adjustment and (2) geometrical model merging.

Joint parametric adjustment helps to remove trivial conflicts between primitives and compensates for small deviations (cf. Fig. 9a), which may happen during the reconstruction driven by a stochastic process (cf. Sect. 4). The parameters of all building components are jointly adjusted using two rules. In the joint adjustment the change of each side of a primitive is proportional to its size, i.e., footprint area.

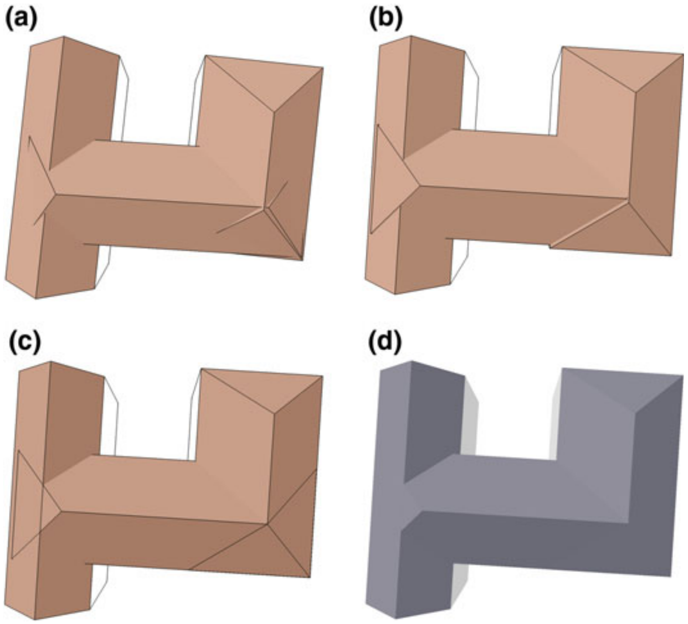


Fig. 9 Primitive adjustment: Reconstructed primitives (**a**), joint parametric adjustment (**b**), vertices-shifting adjustment (**c**), and rendered model (**d**)

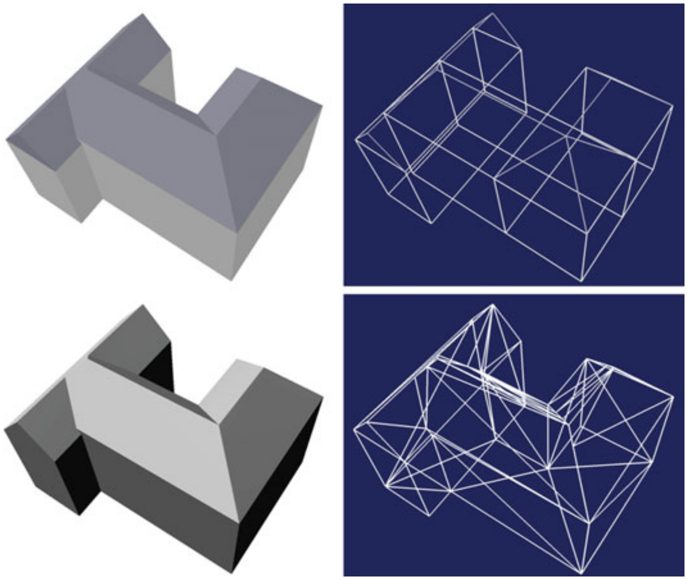


Fig. 10 Primitive merging: from multiple primitives (*top*) to single watertight model (*bottom*)

- Rule 1: The intersection angles of the primitives are jointly regularized to 0° or 90° if the deviation is less than a threshold of 5°. An exception exists for merging a flat roof with a roof with a ridge line (e.g., gable roof): The flat roof is aligned to the latter instead of adjusting both, as the orientation of a ridge is much more reliable than that of a flat roof.
- Rule 2: Heights of flat roofs or ridge- and eave-heights of other roofs are harmonized if the deviation is less than 0.2 m.

Figure 9b shows that the parameter adjustment cannot remove all mismatching positions of the primitives. The mismatching is the result of deviations caused by stochastic processes and data uncertainty, which in principle cannot be corrected in this step. Therefore, further geometrical adjustment is required.

Geometrical model merging generates the final single model of the building complex. Inspired by Huang et al. (2013a), we conduct a simple vertices-shifting (Fig. 9c) to correct the geometrical mismatching and all the primitives are merged into a watertight model. The primitives are originally generated as B-Rep models, as shown in Fig. 10 (top) and simply placed together as two separate models which overlap. Although in the rendered model (left) the intersected part is hidden and does not affect the appearance, the model is ontologically not a single "subject" and geometrically not watertight. Our model merging employs CSG modeling. As shown in Fig. 10, the B-Rep primitives are first converted into CSG models and merged with a "union" operation into a single solid body. The latter is then converted back to a single and watertight B-Rep model, i.e., the final model.

6 Experiments

The experiment is performed for a complete and typical central European village with a mixture of detached buildings and building complexes, a church, and a small castle on a hill. The 3D point cloud has been reconstructed by dense matching of UAS (unmanned aerial system) imagery taken in Bonnland, Germany (Kuhn et al. 2014). The data are generated from 822 images that cover about 0.12 km^2 of undulating terrain. We use a reduced and rasterized version with a resolution of 0.2 m (Fig. 11a, b). The building mask (Fig. 11c, red) is provided by a previous scene classification (Huang and Mayer 2015).

The building mask is decomposed into 62 data tiles (cf. Fig. 2), which are processed in parallel and the reconstructed building models are assembled in the global system. Bird's-eye views of input point cloud (top) and the reconstructed model (bottom) are given in Fig. 12. Along with the watertight building models a mesh model is generated from the non-building points to model the ground.

The total runtime of the presented scene with 33 single buildings and 29 building complexes, which is composed of 112 primitives, is about 14 min on a laptop with a 4 cores/8 threads CPU at 2.3 GHz. The overall reconstruction error (for each individual building) is defined as average deviation from the data points to the model surface

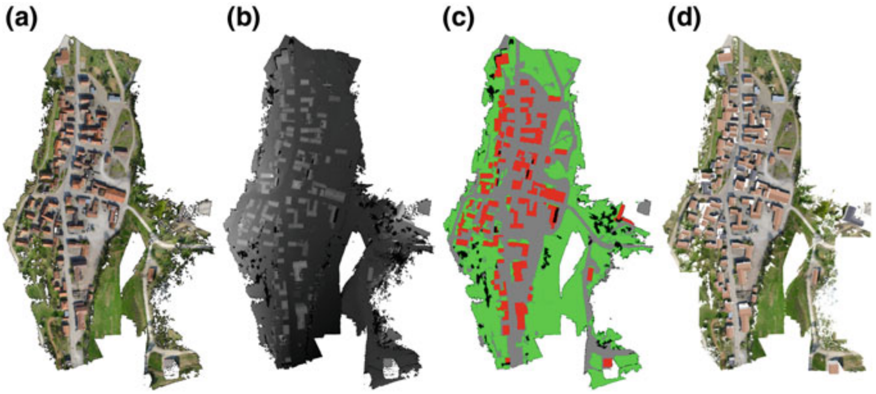


Fig. 11 Experiment on Bonnland data: **a** Input point cloud with color, **b** input point cloud with height presented as *gray* values, **c** the result of scene classification with the building mask in *red*, and **d** the final vector model of the whole scene

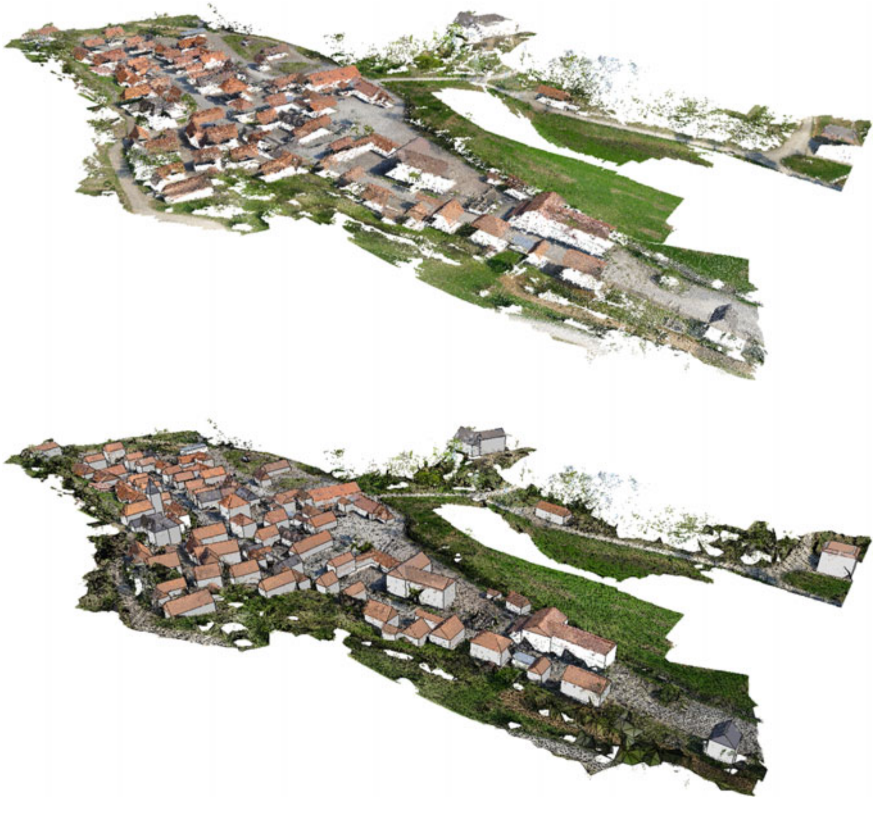


Fig. 12 Building reconstruction for Bonnland: input point cloud (*top*) and reconstructed models (*bottom*)

Table 1 Comparison of the runtimes (in seconds) with and without building decomposition

Methods	Prim. 1	Prim. 2	Prim. 3	Merging	Total	In %
w/o Decomp. (global + local)	104 + 36	89 + 28	65 + 32	25	379	100
with Decomp. (local only)	22	20	25	25	92	24.2

in nadir direction (cf. also Huang et al. 2013a). Except for buildings with large data flaws (cf. Fig. 8) or occlusion, where the average data deviation does not reflect the reconstruction accuracy, the reconstruction errors of the majority of the buildings are less than the half of the data resolution, i.e., 10 cm.

To demonstrate the performance improvement with building decomposition (cf. Sect. 3) in comparison to a direct reconstruction of a building complex we list in Table 1 the detailed runtime analysis for the example model presented in Fig. 10 with three primitives. With the initial values provides by building decomposition computational effort has been saved not only for the time-consuming global searching, but also for the local optimization of other parameters.

7 Conclusion

This paper presents an automatic pipeline for Level of Detail 2—LOD2 building model reconstruction focusing on a reliable scene as well as building decomposition into regular primitives and their subsequent assembly. We proposed:

1. Decomposition of the whole scene into data tiles containing individual buildings or building complexes
2. Decomposition of building complexes into standard primitives with fully 3D geometrical parsing
3. CSG primitive assembly into a single watertight model.

The primitive decomposition links the scene interpretation with the model reconstruction and, thus, completes the automatic modeling pipeline. Additionally, it significantly improves the reconstruction efficiency concerning the following aspects:

1. The time-consuming global search for buildings in a large scene is avoided.
2. The data tiles can be processed independently in parallel.
3. The initial values derived for building decomposition are more precise and reliable than that derived from a simple building mask making the parameter optimization of the primitives much more efficient (cf. Sect. 6).

Concerning future work, we consider to extend the proposed reconstruction pipeline to LOD3 building models. To this end, an approach to object detection on the facades and roofs will be included. CSG modeling should be advantageous for the consistent integration of windows, doors, balconies, chimneys, and dormers into the 3D model. Furthermore, the library of primitives will be improved with non-rectangular bases, e.g., ellipses or general polygons. The texturing of the buildings is of great interest for certain applications such as tourism. B-Rep models can be much more easily derived from CSG models than the other way around. We assume that a watertight B-Rep model with regular shapes, which we generate with our approach, is a good basis for texturing.

References

Arefi H, Hahn M, Reinartz P (2010) Ridge based decomposition of complex buildings for 3D model generation from high resolution digital surface models. In: The international archives of the photogrammetry, remote sensing and spatial information sciences, vol 38(1/W17)

Brenner C (2004) Modelling 3D objects using weak CSG primitives. In: international archives of photogrammetry and remote sensing, vol 35

Brenner C (2005) Building reconstruction from images and laser scanning. Int J Appl Earth Obs Geoinf, Theme Issue Data Qual Earth Obs Tech 6(3–4):187–198

Brenner C, Haala N (2000) Erfassung von 3D Stadtmodellen. Photogrammetrie - Fernerkundung - Geoinformation 2:109–117

Elberink SO, Vosselman G (2009) Building reconstruction by target based graph matching on incomplete laser data: analysis and limitations. Sensors 9(8):6101. doi:10.3390/s90806101, http://www.mdpi.com/1424-8220/9/8/6101

Haunert JH, Sester M (2008) Area collapse and road centerlines based on straight skeletons. GeoInformatica 12(2):169–191. doi:10.1007/s10707-007-0028-x

Huang H, Brenner C (2011) Rule-based roof plane detection and segmentation from laser point clouds. In: Joint urban remote sensing event (JURSE) 2011, 11–13 April. IEEE, Munich, Germany, pp 293–296. doi:10.1109/JURSE.2011.5764777

Huang H, Brenner C, Sester M (2011) 3D building roof reconstruction from point clouds via generative models. 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS), 1–4 November. ACM Press, Chicago, IL, USA, pp 16–24

Huang H, Mayer H (2015) Robust and efficient urban scene classification using relative features. In: 23rd ACM SIGSPATIAL international conference on advances in geographic information systems. ACM, New York, NY, USA, GIS'15, pp 81:1–81:4. doi:10.1145/2820783.2820872

Huang H, Brenner C, Sester M (2013a) A generative statistical approach to automatic 3D building roof reconstruction from laser scanning data. ISPRS J Photogrammetry Remote Sens 79:29–43. doi:10.1016/j.isprsjprs.2013.02.004, http://www.sciencedirect.com/science/article/pii/S0924271613000476

Huang H, Kieler B, Sester M (2013b) Urban building usage labeling by geometric and context analyses of the footprint data. In: 26th international cartographic conference (ICC)

Kada M, McKinley L (2009) 3D building reconstruction from LiDAR based on a cell decomposition approach. In: The international archives of the photogrammetry, remote sensing and spatial information sciences, vol 38(3/W4), pp 47–52

Kuhn A, Mayer H, Hirschmüller H, Scharstein D (2014) A TV prior for high-quality local multi-view stereo reconstruction. In: 2nd international conference on 3D vision (3DV), pp 65–72

Lafarge F, Mallet C (2012) Creating large-scale city models from 3d-point clouds: a robust approach with hybrid representation. Int J Comput Vision 99(1):69–85

Lafarge F, Descombes X, Zerubia J, Pierrot-Deseilligny M (2010) Structural approach for building reconstruction from a single DSM. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(1):135–147

Mäntylä M (1987) An introduction to solid modeling. Computer Science Press Inc, New York, NY, USA

Partovi T, Huang H, Krauß T, Mayer H, Reinartz P (2015) Statistical Building Roof Reconstruction from WORLDVIEW-2 Stereo Imagery. ISPRS—International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences pp 161–167. doi:10.5194/isprsarchives-XL-3-W2-161-2015

Sampath A, Shan J (2010) Segmentation and reconstruction of polyhedral building roofs from aerial lidar point clouds. IEEE Trans Geosci Remote Sens 48(3):1554–1567. doi:10.1109/TGRS.2009.2030180

Schnabel R, Wessel R, Wahl R, Klein R (2008) Shape recognition in 3D point-clouds. In: The 16th international conference in central europe on computer graphics, visualization and computer vision, 4–7 February, Plzen-Bory, Czech Republic

Vosselman G (2009) Advanced point cloud processing. In: Fritsch D (ed) Photogrammetric Week '09. Heidelberg, Germany, pp 137–146

Wang H, Zhang W, Chen Y, Chen M, Yan K (2015) Semantic decomposition and reconstruction of compound buildings with symmetric roofs from lidar data and aerial imagery. Remote Sens 7(10):13,945. doi:10.3390/rs71013945, http://www.mdpi.com/2072-4292/7/10/13945

Zhou QY, Neumann U (2012) 2.5D building modeling by discovering global regularities. In: The IEEE computer society conference on computer vision and pattern recognition, 16–21 June. IEEE Computer Society, Providence, RI, USA, pp 326–333

Dynamic Transfer Patterns for Fast Multi-modal Route Planning

Thomas Liebig, Sebastian Peter, Maciej Grzenda and Konstanty Junosza-Szaniawski

Abstract Route planning makes direct use of geographic data and provides beneficial recommendations to the public. In real-world the schedule of transit vehicles is dynamic and delays in the schedules occur. Incorporation of these dynamic schedule changes in multi-modal route computation is difficult and requires a lot of computational resources. Our approach extends the state-of-the-art for static transit schedules, Transfer Patterns, for the dynamic case. Therefore, we amend the patterns by additional edges that cover the dynamics. Our approach is implemented in the open-source routing framework OpenTripPlanner and compared to existing methods in the city of Warsaw. Our results are an order of magnitude faster then existing methods.

1 Introduction

In a changing world geo-spatial data is subject to dynamic changes and geo-information systems are required to incorporate real-time updates in their analysis and computations (Schnitzler et al. 2014). In this paper we focus particularly on route planning systems. While in a static world a bunch of algorithms exist to compute (shortest) paths from a starting location to a target location efficiently (compare Sect. 2), this problem becomes more difficult in case of multi-modal trip planning including public transport, as temporal constraints, e.g. transit times and departure times, need to be incorporated. In real world, these static schedules are not met, but delays occur (Mazimpaka and Timpf 2016) and deviations from the schedule could be observed. Incorporation of these dynamic information in route computation is

T. Liebig (✉) · S. Peter
TU Dortmund University, Dortmund, Germany
e-mail: thomas.liebig@tu-dortmund.de

M. Grzenda · K. Junosza-Szaniawski
Faculty of Mathematics and Information Science,
Warsaw University of Technology, Warsaw, Poland
e-mail: M.Grzenda@mini.pw.edu.pl

K. Junosza-Szaniawski
e-mail: K.Szaniawski@mini.pw.edu.pl

A. Bregt et al. (eds.), *Societal Geo-innovation*, Lecture Notes
in Geoinformation and Cartography, DOI 10.1007/978-3-319-56759-4_13

beneficial, as it allows to provide tractable travel recommendations to the public. The dynamic information on the delays can be achieved by monitoring positions of the vehicles and even by prediction of future delays. This enables pro-active trip computation.

In Liebig et al. (2014, 2017), we highlighted how vehicular traffic predictions can be incorporated into trip computations. The paper at-hand incorporates public transport delays (which could be the result of prediction) and focuses on the tractability of dynamic transit computation. Existing single source shortest path computation algorithms for the dynamic transit problem suffer from their long computation time, the very fast route planning algorithm for transit networks, Transfer Pattern, does not guarantee soundness in case of real-time delay information. Our approach, overcomes these shortcomings and introduces dynamic transfer patterns, a data structure that encodes which novel transit possibilities are enabled due to the delays.

In a comparison with existing dynamic transit routing schemes, in the city of Warsaw, we highlight the performance gain using our method. Our findings are implemented in the commonly used open source trip planning framework OpenTripPlanner and a pre-configured Virtual Machine is ready to use in industrial context.

The paper is structured as follows. Section 2 provides an introduction to routing algorithms and the transit routing problem. Section 3 presents our Dynamic Transfer Pattern method. Implementation details are provided afterwards in Sect. 4. Next, we analyse tram delays in the city of Warsaw, Poland, and we continue with performance evaluation in this city, Sect. 4. In the end, we discuss future research ideas, Sect. 7.

2 Related Work

In this paper we focus on the point-to-point shortest path problem (Bast et al. 2016), where in a graph $G = (V, E)$ a path between a source $s \in V$ and target $t \in V$ needs to be found such that the cumulative edge-wise cost $l(u, v), with(u, v) \in E \subseteq V \times V$ along the path is minimized.

2.1 Shortest Path Routing

Standard solution to the problem is using Dijkstra's algorithm (Dijkstra 1959). Given the graph $G = (V, E)$ and $s, t \in V$, it initializes a queue of nodes $Q = V$ and a distance function over $V \times V$ with $dist(s, s) = 0$ and $dist(s, v) = \infty, \forall v \neq s, v \in V$. Until the queue is empty the node u with the smallest distance $dist(s, u)$ is picked and removed from Q. For each neighboring node of u the distance is updated as follows: $dist(s, v) := dist(s, u) + l(u, v)$, if the latter is smaller than the former. Dijkstra's algorithm can be sped up by running it simultaneously from both s and t until a common node u is hit. In the slightly modified version of Dijkstra's algorithm A* (Hart et al. 1968) the order in the priority-queue for the traversal not only depends on the cumulated costs to reach a vertex in the graph but also on the expected costs to reach the goal from this vertex. Bound by Minkowski's inequality, whereas $||x + y||_p \leq ||x||_p + ||y||_p$ (known as triangle inequality for $p = 2$), A* prunes the search space in comparison to Dijkstra's Algorithm. A sound heuristic for the remaining cost estimation is the geographical distance that is always lower than the road-based distance.

In case of static cost functions Geisberger et al. propose a data structure called contraction hierarchies (Geisberger et al. 2008), which speed up the A* algorithm and enable trip calculation in large traffic networks at European scale. Instead of searching the shortest path directly within the traffic network, contraction hierarchies reduce the search space to the most important ones. In a preprocessing step these important segments are identified (based on the topology) and the network is extended by edges between these important links.

2.2 Shortest Path Routing in Transit Networks

In contrast to regular road networks, public transportation data enhances a spatial graph with temporal data by adding timetable information. A trip T serves a sequence of stops $stops(T) = (s_1, \ldots, s_n), s_i \in S$. Thus T connects two stops s_a and s_b if and only if $stop(T, s_a) < stop(T, s_b)$. If one or more trips contain the exact same sequence of stops, they form a line (Bast et al. 2013).

Common approach is to incorporate dynamic information into the graph G and then to apply Dijkstra's algorithm. This results in a time extended or time dependent model. In the time extended model every transit node is split into multiple vertices for each event (arrival, transit and departure). The time dependent model assigns every transit node one vertex and arcs encode temporal constraints.

A recent data structure and algorithm, *transfer patterns*, introduced by Bast et al. (2010) is considered state-of-the-art in public transport routing. Based on the assumption that during a day, there are only a few optimal routes from stop s_s to stop s_t that differ only in the time they take place. In a preprocessing phase, optimal routes are computed as a sequence of transfer stations, neglecting the time component as well as information about intermediate stations. For each origin and target destination a directed acyclic graph is saved, containing all routes starting with the destination and containing all intermediate stations until the origin is reached.

In a realistic route planning scenario, various delays occur amongst the public transport vehicles. In contrast to vehicular traffic, trams and trains can not overtake, and vehicles in transit networks wait for each others (e.g. connecting trains), this causes delays to propagate differently than vehicular traffic jams. In addition, two modes of transportation may share the same physical resource (e.g. buses or trams riding on vehicular street). Thus, two forms of delays in transit networks are distinguished in literature: (1) a vehicle is late due to own reasons, and (2) other vehicles are late caused by the former (Müller-Hannemann and Schnee 2009).

Several models for transit delays are reported in literature. The work in Dibbelt et al. (2013) assumes independence. In contrast, Goerigk et al. (2011) allow delays within their approach to cumulate. Sophisticated models incorporate dependencies among the vehicles into the delay (Higgins and Kozan 1998). In Mazimpaka and Timpf (2016) the delays are analyzed visually.

In a trip planning application real-time predictions of delay are a main benefit as future delays may influence the route choice. Thus, we highlight two recent works on delay prediction and delay recognition: Gal et al. (2015) applies queueing theory and assumes delays to aggregate, (Zygouras et al. 2015) detects delays and unexpected vehicle movement in real-time from the GPS traces.

In this work we do not focus on the prediction, but assume that we have information on delays of vehicles (in the commonly used GTFS realtime data format) either from vehicle observations or even predictions.

With such dynamics the trip computation becomes more difficult. Though states a previous paper (Bast et al. 2013) that transfer pattern are delay robust, but this only holds as long as no new transfers are enabled by the delay. In the likely case that novel transfers are enabled the existing transfer patterns do not represent this information and can not result in the optimal transit route.

3 Dynamic Transfer Pattern

Transfer Patterns were introduced in Bast et al. (2010). The method comprises a data structure and an algorithm for fast transit route computation. In a preprocessing step all possible routes are precomputed and stored in a compressed way. For each public transport line a table is stored denoting in the columns the stops along the line. In this way it holds the maximal possible route without changes. The rows of the table represent the actual trips of the line, Table 1 gives an example, compare also Cárdenas (2013).

In addition, for every station a list is stored with the passing lines and their position in the trips, see an example in Table 2.

Table 1 Transfer pattern example

Line L17	s_a	s_s	s_b	s_y	⋯
Trip 1	8:15	8:22 8:23	8:27 8:29	8:38 8:39	⋯
Trip 2	9:14	9:21 9:22	9:28 9:28	9:37 9:38	⋯

Table 2 Transfer pattern example (continued)

s_s:	(L8,4)	(L17,2)	(L34,5)	(L87,17)	⋯
s_y:	(L9,4)	(L13,5)	(L17,4)	(L55,16)	⋯

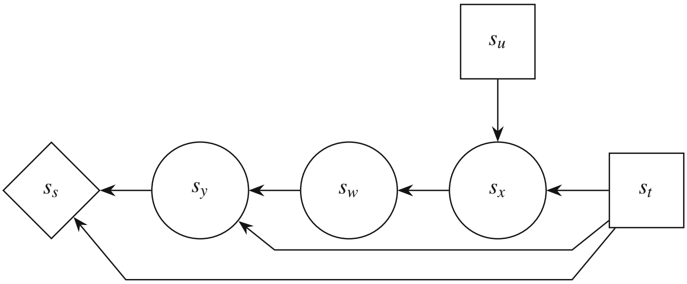


Fig. 1 DAG structure of the transfer pattern

With Transfer Pattern a route from s_{start} to s_{stop} at time t is calculated by the intersection of the lists for s_{start} and s_{stop}, the first connection after time t is the desired result. As an example, the route from s_s to s_y at 9:03 is computed by intersecting the lines in Table 2, we find that line 17 connects the stops on positions 2 and 4. According to Table 1 the earliest possible trip is departing 9:22 at s_s and arrives at s_y at 9:37. In the pre-processing phase the shortest paths amongst all stops (neglecting temporal information) are constructed and intermediate stops are stored in a directed acyclic graph (DAG). Figure 1 exemplifies this for s_s.

In case of a routing query from stop s_s to stop s_t the DAG for s_s is picked and all connecting paths among s_s to s_t are taken into account for routing.

With dynamic schedule information at hand, the time table information (Table 1 in our example) can be updated and the route computation could be performed as in the regular case (Bast et al. 2013). However, it might be that the optimal connection is not precomputed as new connections could be enabled by the delay itself, especially if multiple trips arrive belated. Incorporation of the delays in trip computation provides benefits to travelers, thus next section addresses our implementation of dynamic transfer pattern in the routing platform OpenTripPlanner.

Our approach is to include alternative patterns that emerge due to delays or cancellations in the data structure by simulating these very delays during precomputation. In order to achieve this, all lines of non-final trips of a Transfer Pattern originating in s_s are recorded during computation of static routes. In a next step, these lines are combined into delay scenarios.

As a simple example of such a scenario, consider a route from s_s to s_t by transferring at s_y. Furthermore, the first trip is delayed considerably, making an alternative route with a transfer at s_v favorable in the Pareto sense. Since the advantage of the second pattern depends on a certain scenario of realtime delays, it could not have been precomputed as a static Transfer Pattern.

Alternative Transfer Patterns are computed by applying to the graph by delaying according trips artificially, making transfers to the next respective trips in time infeasible. When running the routing algorithm now, it returns the next best routes considering the delay situation. All such alternative Transfer Patterns are eventually

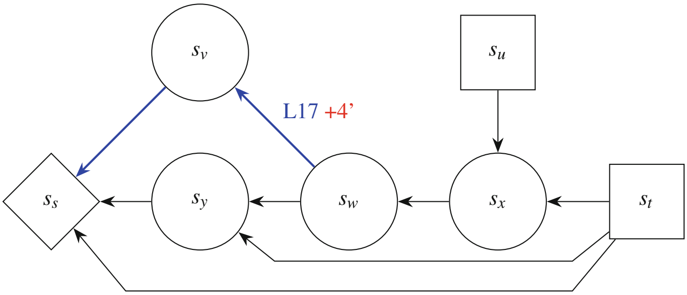


Fig. 2 DAG structure of the dynamic transfer pattern

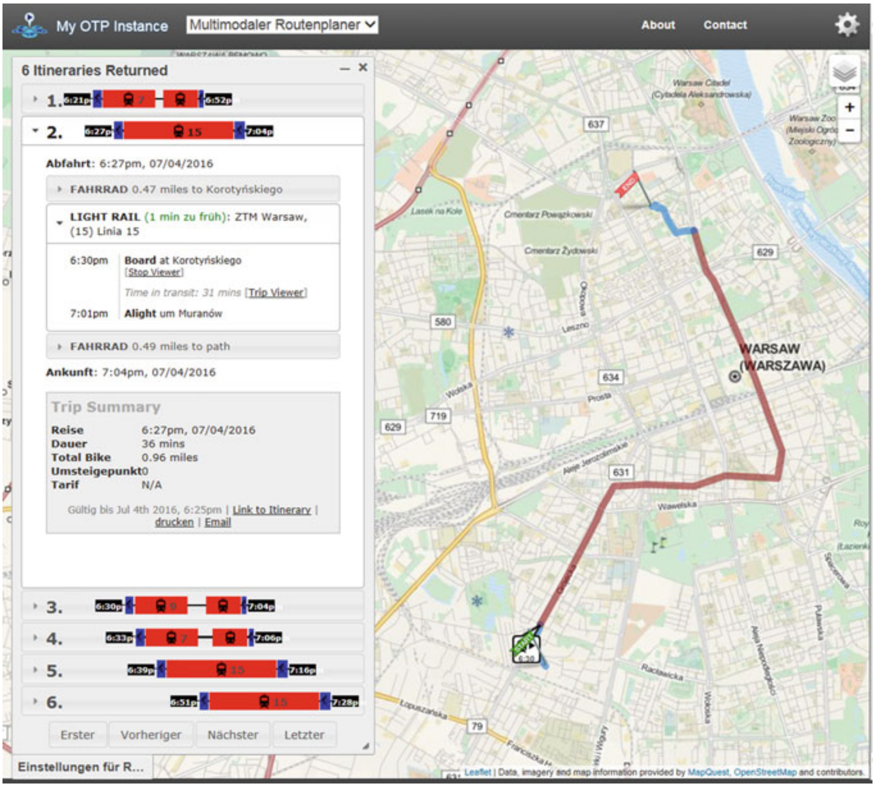
merged into the regular DAGs after having been classified in terms of lines that were artificially delayed and their respective amount of delay. See Fig. 2.

All lines used in regular Transfer Patterns can be combined into delay scenarios in numerous ways. Considering all combinations of n delayed lines producing 2^n scenarios implies enormous computational costs. Multiple ways of combining delayed trips were thus introduced, trying to cover many alternative routes while keeping the overall number of scenarios as low as possible.

A trivial approach is to incorporate delays of only a single line at a time. This method may already significantly increase computation costs for large graphs compared to merely computing static patterns. For this reason, picking a limited number of random lines per Transfer Pattern subgraph was introduced as another approach. Lastly, in pursuance of computing the most useful alternative routes, past data of lines with a high likelihood of delay can be utilized. This means picking often-delayed lines or combinations thereof more frequently when constructing a limited amount of delay scenarios.

When answering a routing query with dynamic traffic data, the corresponding query graph is fetched in a similar fashion regular Transfer Patterns are handled. When walking across the graph from target to source, the delay classification of each arc is checked. Arcs with no delay classification are always considered, in contrast to arcs with delay classification, which have to match the actual traffic situation. Realtime traffic information match a classification if and only if each trip of the classification is delayed by an amount bigger or equal to the amount specified.

Since real time delay information has been applied to departure and arrival times, it is possible that some patterns need a much longer travel time or are completely infeasible and thus no longer interesting to the user. These patterns are discarded either when direct connections are fetched for all trips or when they are dominated by other patterns in the Pareto sense.



6 Itineraries Returned

2\.

Abfahrt: 6:27pm, 07/04/2016

FAHRRAD 0.47 miles to Korotyńskiego

LIGHT RAIL (1 min zu früh): ZTM Warsaw, (15) Linia 15

6:30pm Board at Korotyńskiego [Stop Viewer]

Time in transit: 31 mins [Trip Viewer]

7:01pm Alight um Muranów

FAHRRAD 0.49 miles to path

Ankunft: 7:04pm, 07/04/2016

Trip Summary

Reise	6:27pm, 07/04/2016
Dauer	36 mins
Total Bike	0.96 miles
Umsteigepunkt	0
Tarif	N/A

Gültig bis Jul 4th 2016, 6:25pm | Link to Itinerary | drucken | Email

Fig. 3 OpenTripPlanner incorporating dynamic delay information

4 Integration in OpenTripPlanner

We implemented the hereby presented dynamic transfer pattern routing scheme in OpenTripPlanner (a commonly used open-source trip computation framework). Therefore we consume information on the transit network and schedules from a commonly used GTFS representation, and dynamic updates in GTFS-realtime format The latter could be retrieved either by an automatic vehicle location system, as in Mazimpaka and Timpf (2016) or by real-time predictions as in Zygouras et al. (2015). OpenTripPlanner uses the street network provided by OpenStreetMap. Our routing scheme is integrated as an optional routing algorithm in OTP, its sources are publicly available and a running setup is preconfigured as vagrant box.[1] A screenshot of the user-interface of the routing system in depicted in Fig. 3.

[1] https://bitbucket.org/tliebig/developvm/branch/transferpatterns.

5 Analysis of Tram Location Data

It is vital before delays are incorporated in route planning to understand some delay data. In our case, we study the delays in the city of Warsaw, Poland. These data are attained through the integration of location data retrieved in near real time manner from GPS sensors present in trams, tram stop point coordinates and schedule data. Every tram reports its GPS position together with its identifier, which is a combination of line number and brigade number. Such reports are produced every 30 s. These data are available via API to the public, except for the number of brigade, which was made available by the City of Warsaw for the project. The data are not clean, various types of problems appear: some records are missing, GPS positions are inaccurate, occasionally two trams report the same identifier. What should be emphasised here is that no imputation of missing data was applied. As an illustration of GPS data quality issues, see the approximation of the tram routes provided in Fig. 4.

Even if the GPS data were perfect, it would not be clear how to extract from GPS data precise times of arrival and departures of trams at tram stops. One of the reasons is that a tram stop is defined by point coordinates of the stop. The simplest method to compute the departure of a tram from a stop is to compute the time when a tram leaves a circle centered at stop point. However, it is not clear what radius of the circle is appropriate. If it is too small, then a tram can ‘miss’ the stop. The radius must be larger than the length of a tram, since two trams in a row can stop at the same time on

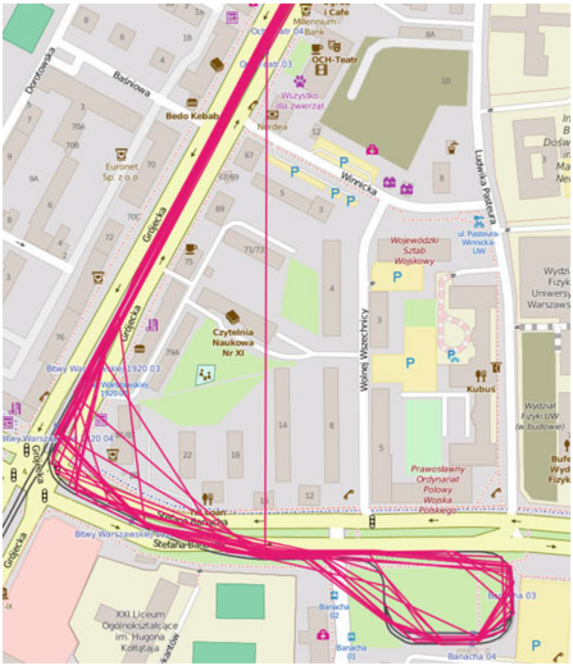


Fig. 4 Sample tram routes approximated from GPS data. Periodical location reporting and gaps in location reporting result in occasionally significant distance between consecutive tram coordinates. Best viewed in *color*

the same tram stop. On the other hand, if the radius is big then other problems appear. The tram might have left the stop but is waiting near the stop by the traffic lights, and we still consider it as present at the tram stop. In quite many cases, a tram stop is located immediately before traffic lights. In such cases, the fact that a tram does not leave the stop on time may mean delay caused by the tram itself or may be due to traffic lights suspending tram departure. All these factors contribute to the fact that precise calculation of arrival and departure times would not be possible even based on accurate tram coordinates i.e. coordinates not affected by GPS inaccuracies.

To address these problems, at least partially, for every tram stop we define a line going through the stop point. If a tram crosses this line, we consider the tram has left the tram stop. If the route of the tram near the tram stop is straight, the line is perpendicular to the route. If the tram stop is near the turn of the route, then the line is parallel to the bisector of the turn. These rules let us address the issue of estimating tram departure time based on location time series. Still, departure time estimates remain noisy for the reasons described above. Moreover, they can be affected by traffic lights. For these reasons it is inevitable for the trams to:

- arrive earlier than needed, because of varied traffic light conditions
- be considered late because of standing behind another tram (and tram stop line), while being already at tram stop and having its doors open
- be considered late because of waiting for traffic light change.

Finally, what is worth mentioning here is that temporary conditions such as events causing major traffic disruption may be not reflected in the schedules and are just announced through City of Warsaw Twitter and RSS channels. Taking into account all these aspects, what is considered as a delay or early arrival may mean on time departure or departure at a time not matching the schedules due to the circumstances beyond control of Warsaw Trams.

To illustrate the distribution of the differences between scheduled time and the most approximate observed departure time in time domain, the data for 23rd of March 2016 coming from Warsaw City tram system was used. In the preprocessing stage, it was limited to the time between 4:00 and 23:59. Moreover, differences between planned and observed departure exceeding 20 min were isolated for further investigation. Altogether events related to these two categories correspond to less than 4% of the data. Based on the remaining 96.4% of the data, the analysis described below was performed. First of all, Fig. 5 illustrating all differences has been developed. Not surprisingly, it is dominated by on-time arrivals. It is more interesting to look at early departures and delayed departures, provided in the left and right part of Fig. 5, respectively. In particular, what can be observed is quite a significant proportion of early departures. As stated above, based on the available data distinction between actual arrival and departure time may be problematic. Hence, the interpretation of the histograms necessitates particular attention being paid to the way the data has been collected and processed.

What is of particular interest for dynamic route planning is whether delays and early departures vary over the day. Figure 6 answers this question showing mean absolute delay for individual hours of the day. Quite surprisingly, lower values are

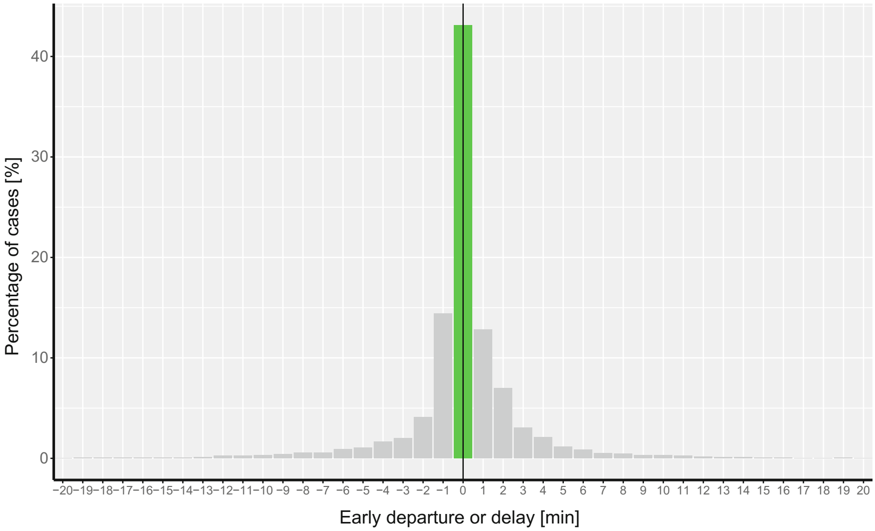


Fig. 5 Percentage of share of differences between planned and observed time of passing tram stop line; the bar in *green* shows the percentage of cases of less than 30 s difference. Based on the data for 23rd of March 2016 from Warsaw City trams, limited to the time between 4:00 and 23:59 and containing coordinate records from all 25 tram lines operated in this period. It can be observed that majority of trams pass tram stop lines in time. Still, relatively many early and delayed departures are observed. Among other reasons, the aforementioned limited ability to precisely determine arrival and departure time contributes to the problem. The figure is best viewed in *color*

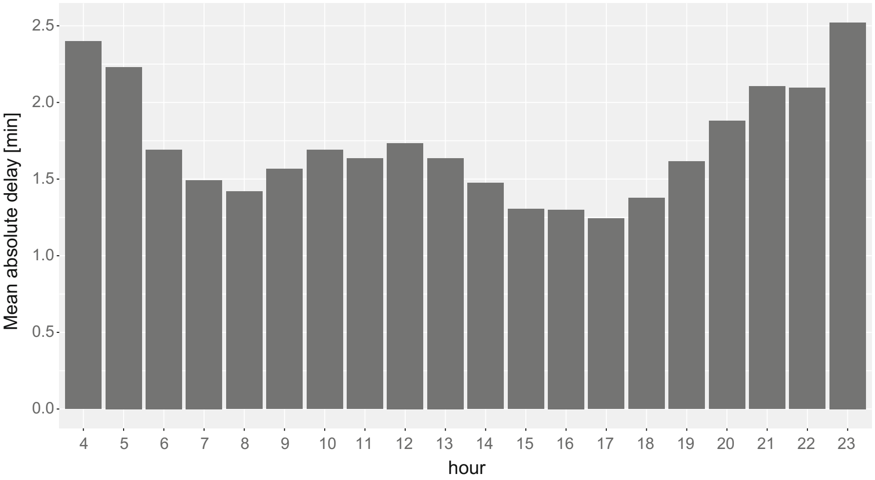


Fig. 6 Average mean absolute tram delay per an hour. Both early and late departures are considered. Significant variation of mean absolute difference between planned and observed departure takes place over the day. The differences observed are between approx. 1.3 and 2.5 min. Interestingly, the smallest differences meaning the most punctual trams are observed during peak traffic hours i.e. for $h \in \{7, 8, 9, 15, 16, 17, 18\}$

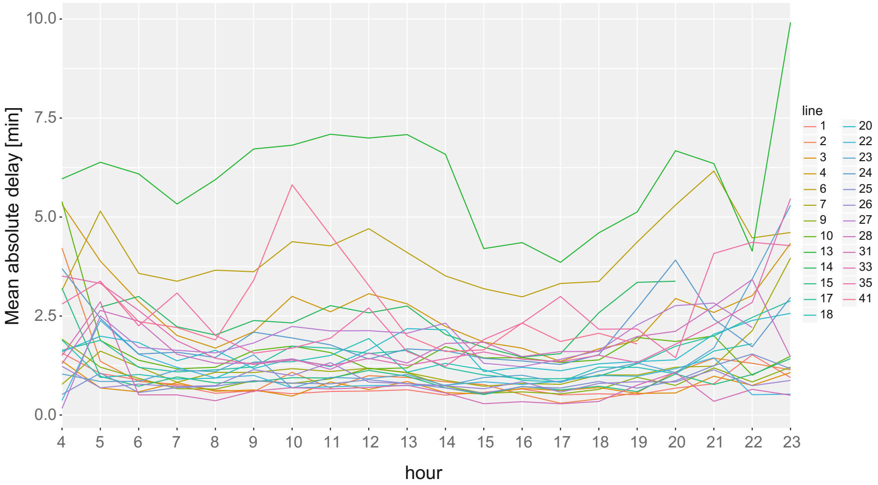


Fig. 7 Average mean absolute tram delay per an hour for individual tram lines. Both early and late departures are considered. Differences in the punctuality of individual tram lines are observed. These vary between 0.5 and 10 min depending on the line and time of the day. Best viewed in *color*

observed for peak hours 7–9 and 15–18. This may suggest the potential for both the development of prediction module and schedule improvement. Further analysis performed for individual tram lines separately is provided in Fig. 7. This reveals that two tram lines largely contribute to the overall mean absolute delays. Hence, some routes are far more susceptible to delays than others.

Another performance indicator to consider is the proportion of trams on time, departing early and late. We consider a tram to be on time, if it passes tram stop point at most 120 s before or 120 s after the scheduled time. The percentage of early, on time and late departures of trams throughout the day remains largely stable. It is surprising that the trams are relatively more punctual in the rush hours and less punctual just before the morning peak and just after the afternoon peak. In particular, minimum percentage of on time departures per an hour is 75.6% and occurs at $h = 21$. On the other hand, maximum percentage of on time departures per an hour is 85.2%, which is observed at $h = 17$ i.e. during peak afternoon period. The reasons of this are worth investigating in the future.

The preliminary analysis of tram location data compared with schedule data, reveals that:

- departure time not matching schedule time can be identified, but has to be analysed carefully, taking into account limited certainty of departure time estimation,
- still, noticeable number of early and late departure events can be observed in the data,
- tram delays and early departures significantly vary based on the time of the day and tram line.

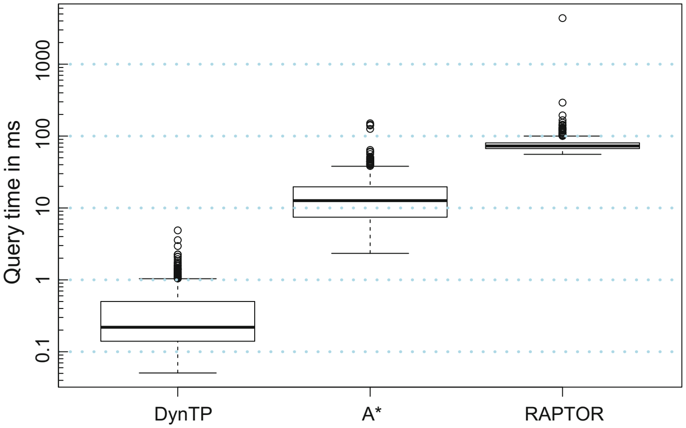


Fig. 8 Distribution of the computation time of public transport routing schemes (Dynamic Transfer Patterns, A*, RAPTOR) in milliseconds for 1000 randomly chosen source target locations in OpenTripPlanner under realistic conditions in Warsaw, Poland. Note logarithmic scaling

6 Performance Evaluation

Main goal of the paper hereby is to speed-up trip computation in case of dynamic delays. As we aim to apply the transit route computations in an industrial project, we decided to extend capabilities of existing open source platform OpenTripPlanner (OTP). Thus, we compare our dynamic transfer pattern with the transit routing schemes already available in OTP. In OTP the algorithms A* (Hart et al. 1968) and RAPTOR (Delling et al. 2012) are available. We test the routing performance in the city of Warsaw, Poland. On startup, we perform initial pre-processing of the transfer patterns based on a GTFS timetable information. Afterwards, we compute for approximately 1000 source destination pairs the public transport routes. Experiments are performed on a regular desktop machine, computation time is measured in milliseconds. The resulting distribution of computation times can be seen in Fig. 8, please note logarithmic scaling.

As can be seen in the Fig. 8 our algorithm is an order of magnitude faster than existing transit computation schemes. However, initial pre-processing is exhaustive and required about 8 h, this can easily be reduced by distribution of the initial preparations.

7 Discussion and Future Work

In this work, we focused on fast transit route computation in a changing world. We highlighted the shortcomings of existing algorithms, that are either very slow, or do not incorporate real-time transit information. We overcame these limitations by introduction of Dynamic Transfer Patterns. In this routing scheme, we applied the basic idea of Bast et al. (2010) but created additional links in the patterns for transit connections that occur due to the delays. Thus, the modified Transfer Patterns can be applied also in dynamically changing environment. The method was made publicly available as ready-to-use virtual machine and as source code integrated in the commonly used OpenTripPlanner. As input our implementation depends on the commonly used GTFS and GTFS Realtime data structures that encode transit information.

The performance of our implementation in comparison to existing methods was measured using real-world data, our method achieved computation times that are an order of magnitude faster than existing ones. However, precomputation is quite exhaustive. For this step we propose future research on biasing the precomputations to the most prominent ones. Visual inspection of the delays, as performed in Mazimpaka and Timpf (2016), can help to prioritize certain delay computations. Moreover the precomputation runs for every station separately, this step could probably benefit from parallelization. These two points are subject for future research.

Acknowledgements The authors received funding from the European Union Horizon 2020 Programme (Horizon2020/2014–2020), under grant agreement number 688380 “VaVeL: Variety, Veracity, VaLue: Handling the Multiplicity of Urban Sensors”.

References

Bast H, Delling D, Goldberg A, Müller-Hannemann M, Pajor T, Sanders P, Wagner D, Werneck RF (2016) Route planning in transportation networks. Springer International Publishing, Cham, pp 19–80

Bast H, Carlsson E, Eigenwillig A, Geisberger R, Harrelson C, Raychev V, Viger F (2010) Fast routing in very large public transportation networks using transfer patterns. In: European symposium on algorithms. Springer, pp 290–301

Bast H, Sternisko J, Storandt S (2013) Delay-robustness of transfer patterns in public transportation route planning. In: ATMOS-13th workshop on algorithmic approaches for transportation modelling, optimization, and systems-2013, vol 33 Schloss Dagstuhl Leibniz-Zentrum fuer Informatik pp 42–54

Cárdenas CJ (2013) Efficient multi-modal route planning with transfer Patterns. Master’s thesis Freiburg University

Delling D, Pajor T, Werneck R (2012) Round-based public transit routing. In: Proceedings of the 14th meeting on algorithm engineering and experiments (ALENEX’12). Society for Industrial and Applied Mathematics

Dibbelt J, Pajor T, Strasser B, Wagner D (2013) Intriguingly simple and fast transit routing. In: International symposium on experimental algorithms. Springer, pp 43–54

Dijkstra EW (1959) A note on two problems in connexion with graphs. Numerische mathematik 1(1):269–271

Gal A, Mandelbaum A, Schnitzler F, Senderovich A, Weidlich M (2015) Traveling time prediction in scheduled transportation with journey segments. Inf Syst

Geisberger R, Sanders P, Schultes D, Delling D (2008) Contraction hierarchies: faster and simpler hierarchical routing in road networks. In: International workshop on experimental and efficient algorithms. Springer, pp 319–333

Goerigk M, Knoth M, Müller-Hannemann M, Schmidt M, Schöbel A (2011) The price of robustness in timetable information. In: OASIcs-openaccess series in informatics. Volume 20 Schloss Dagstuhl-Leibniz-Zentrum für Informatik

Hart PE, Nilsson NJ, Raphael B (1968) A formal basis for the heuristic determination of minimum cost paths. IEEE Trans Syst Sci Cybern 4(2):100–107

Higgins A, Kozan E (1998) Modeling train delays in urban networks. Transp Sci 32(4):346–357

Liebig T, Piatkowski N, Bockermann C, Morik K (2014) Predictive trip planning—smart routing in smart cities. In: Proceedings of the workshops of the EDBT/ICDT 2014 joint conference (EDBT/ICDT 2014), Athens, Greece, March 28, 2014. Volume 1133 CEUR-WS.org, pp 331–338

Liebig T, Piatkowski N, Bockermann C, Morik K (2017) Dynamic route planning with real-time traffic predictions. Inf Syst 64:258–265

Mazimpaka JD, Timpf S (2016) A visual and computational analysis approach for exploring significant locations and time periods along a bus route. In: Proceedings of the 9th ACM SIGSPATIAL international workshop on computational transportation science. ACM, pp 43–48

Müller-Hannemann M, Schnee M (2009) Efficient timetable information in the presence of delays. In: Robust and online large-scale optimization. Springer, pp 249–272

Schnitzler F, Artikis A, Weidlich M, Boutsis I, Liebig T, Piatkowski N, Bockermann C, Morik K, Kalogeraki V, Marecek J, Gal A, Mannor S, Kinane D, Gunopulos D (2014) Heterogeneous stream processing and crowdsourcing for traffic monitoring: highlights. In: Machine learning and knowledge discovery in databases. Volume 8726 of Lecture notes in computer science. Springer, Berlin, pp 520–523

Zygouras N, Zacheilas N, Kalogeraki V, Kinane D, Gunopulos D (2015) Insights on a scalable and dynamic traffic management system. In: EDBT, pp 653–664

The Effect of Regional Variation and Resolution on Geosocial Thematic Signatures for Points of Interest

Grant McKenzie and Krzysztof Janowicz

Abstract Computational models of place are a key component of spatial information theory and play an increasing role in research ranging from spatial search to transportation studies. One method to arrive at such models is to extract knowledge from user-generated content e.g., from texts, tags, trajectories, pictures, and so forth. Over the last years, topic modeling techniques such as latent Dirichlet allocation (LDA) have been studied to reveal linguistic patterns that characterize places and their types. Intuitively, people are more likely to describe places such as Yosemite National Park in terms of *hiking*, *nature*, and *camping* than *cocktail* or *dancing*. The geo-indicativeness of non-georeferenced text does not only apply to place instances but also place *types*, e.g., state parks. While different parks will vary greatly with respect to their landscape and thus human descriptions, the distribution of topics common to all parks will differ significantly from other types of places, e.g., night clubs. This aggregation of topics to the type level creates thematic signatures that can be used for place categorization, data cleansing and conflation, semantic search, and so on. To make full use of these signatures, however, requires a better understanding of their intra-type variability as regional differences effect the predictive power of the signatures. Intuitively, the topic composition for place types such as *store* and *office* should be less effected by regional differences than the topic composition for types such as *monument* and *mountain*. In this work, we approach this regional variability hypothesis by attempting to prove that all place types are aspatial with respect to their thematic signatures. We reject this hypothesis by comparing the signature similarities of 316 place types between major cities in the U.S. We then select the most and least varying place types and compare them to thematic signatures from regions outside of the U.S. Finally, we explore the effects of LDA topic resolution on differences between and within place types.

G. McKenzie (✉)
Department of Geographical Sciences, University of Maryland,
College Park, USA
e-mail: gmck@umd.edu

K. Janowicz
STKO Lab, Department of Geography, University of California,
Santa Barbara, USA
e-mail: janowicz@ucsb.edu

A. Bregt et al. (eds.), *Societal Geo-innovation*, Lecture Notes
in Geoinformation and Cartography, DOI 10.1007/978-3-319-56759-4_14

Keywords Semantic signature · Point of interest · Place type · Geosocial · Topic modeling

1 Introduction

Selecting discriminative features[1] is a key prerequisite for the classification of places into categories, which, in turn, contribute to a wide variety of tasks such as data deduplication, cleansing, recommender systems, geographic information retrieval, and so on (Bao et al. 2015). One source for extracting features is user-generated content, e.g., tags, check-ins, ratings, and many other forms of structured or semi-structured data. To give a concrete example, previous work has shown that the time of the day and day of the week users of geo-social networks check in to places is highly indicative of the place type (Ye et al. 2011; Gao et al. 2013; Shaw et al. 2013). Simplifying, the underlying assumption is that if one is provided with a sufficiently large sample of places of a given type as well as user check-ins to these places, the extracted temporal features will follow our everyday intuition that *restaurants* are visited during lunch and dinner times and that *universities* show a strong decline in visiting intensity in the evenings and during weekends. Such temporal features can be expected to effectively discriminate between restaurants and universities but will more likely fail in telling apart high schools from universities. Consequently, multiple kinds of features are combined to improve classification accuracy. Examples include features extracted from spatial second-order analysis of place patterns by type (Mülligann et al. 2011), features based on the sequence of check-ins (Cheng et al. 2013), features based on user-centric commonalities (Scellato et al. 2011), to name a few.

The examples discussed so far, make use of 2 of the 3 components of geographic information, namely space and time. The third component, specifically *theme*, can be employed to derive features as well. In fact, a multitude of work has investigated user-contributed tags. More recently, the quest for additional features has lead researcher to explore unstructured data, e.g., full text user reviews and tips, as well. Latent Dirichlet allocation (LDA) (Blei et al. 2003) is a particularly popular technique that takes a *bag-of-words* approach to classifying documents based on the co-occurrence of words. An LDA approach produces distributions of topics that can be used to differentiate place types based on the terms and words used in describing these places. These *thematic signatures* offer an additional dimension through which place types can be understood. In many cases, thematic signatures uncover nuanced differences between places that spatial and temporal signatures cannot.

Extracting such statistical features for places raises the important question of whether these features are stable across space. Previous work with *temporal*

[1] The term *feature* has varying meanings across different communities. Here, following common practice in machine learning, we will use it to refer to the measurable characteristics of points of interest as extracted from geosocial data. We will refer to geographic features as *places*.

signatures has shown that some place of interest (POI) types vary regionally while others do not (McKenzie et al. 2015a, b). This is an important finding as it means that features extracted from global datasets will perform well for certain types such as *drugstores* but will be less effective for other types such as *theme parks*. In this work, we extract thematic signatures from use-generated content contributed to POI across the United States and internationally to study the effect of regional variation on place types. We assume that regional effects will be even more prominent across linguistic patterns compared to temporal patterns. **The research questions and contributions addressed in this work are as follows.**

RQ1 With regards to thematic similarity, **is there regional variation between POI types?** This question will be answered by testing a null hypothesis which states that any variation in the thematic properties of POI types can be explained by sampling variation and noise. The *Chi Squared Goodness of Fit* test will be used to test this hypothesis.

RQ2 Are regional variations in topics (themes) equally prevalent across all POI types? In other words, **are some POI types more influenced by a change in region than others?** To respond to this, we compare POI types from the top three largest cities in the United States (New York City, Los Angeles and Chicago). Two methods, namely *Jensen-Shannon Distance* and *Cosine Similarity*, are used to compute dissimilarity and their results are compared for concordance via *Kendall's W*. Second, does regional (in)variance transfer across feature type hierarchies? When exploring the parent types of lower-level POI types, are certain top-level parent POI types more or less regionally prevalent?

RQ3 Having first explored POI type regional variability on a national level, we must then ask, **is the measured thematic variability in POI type consistent internationally?** A subset of POI from Sydney, Australia and London, England are compared thematically with those from within the United States. Through this international comparison we will show that highly regional variant types within the United States continue to remain highly variant when compared to POI across national, cultural and physical borders. The opposite is also true for regionally invariant types.

RQ4 On the topic of resolution of thematic signatures, **does the number of selected LDA topics impact the similarity *within* and *between* place types?** By constructing a wide range of topic models, we show that the number of topics indeed influences place type similarity and that some place types are more susceptible to this influence than others.

2 Related Work

The complex relationship between language and place has been the subject of a considerable amount of research (Basso 1996; Cresswell 2014; Graham and Zook 2013; Tuan 1991; Kinsella et al. 2011; Stefanidis et al. 2013). While much of this work has focused on linguistic descriptions of place, a subset has discussed textual characteristics and the geo-indicativeness of terms and phrases. Along these lines, work by Hollenstein and Purves (2010) explored the ability of tags and textual content garnered from user-contributed geotagged Flickr photos to define local regions such as city centers. Cheng et al. (2010) build on the geo-indicativeness of language and terms to predict twitter users' rough locations based purely on the content of their tweets. Hecht and Gergle (2010) discuss the role of *localness* as it relates to contributing content to various online sources. The authors found that there are strong differences between the contributions of local users to different platforms such as Wikipedia or Flickr.

The recent rise in the use of topic models for describing and classifying documents has also played a role in the geospatial realm. Work by Adams and Janowicz (2012) has employed topic modeling to estimate the geo-indicativeness of non-georeferenced unstructured text. Additional work in this area (Adams et al. 2015) has shown that terms and phrases have probabilistic spatial extents. Combining a topic modeling approach of textual data with spatial clustering analysis and temporal check-in behavior has been successfully employed to reclassify existing POI types into unique and more semantically appropriate top-level types (McKenzie et al. 2015a, b).

The process of extracting unstructured text in the form of tips and reviews in order to compare places and place types has also been pursued in other research areas. While not specific to regional differences, Tanasescu et al. (2013) explored the personality of venues through analysis of the keywords and textual review data contributed about a place. The authors then referenced the *five-factor model of personality* as proposed in the psychology literature to assign personality traits to places. Similarly, Hu and Ester (2013) extract data from online social posts and review sites with the purpose of location prediction and place recommendation.

3 Data and Thematic Signatures

The data used for this research was accessed from the geosocial networking application *Foursquare*. Through the application programming interface (API), 938,031 POI[2] were accessed from three cities across the United States, 437,358 from New York City, New York, 213,279 from Los Angeles, California and 249,169 from Chicago, Illinois. These regions were selected based both on their geographic location (East Coast, West Coast, and Midwest, respectively) and the fact that they constitute the top three most populated urban areas in the United States as reported by the 2010 U.S. Census.

[2] Foursquare refers to Points of Interest (POI) as *venues*.

3.1 Place Types and Tips

Aside from geographic coordinates, attribute information attained from these POI included a *type*, which is assigned by the user contributing the POI, verified by other users, and confirmed by the application administrators. All POI types accessed via the API had been assigned types based on the Foursquare *Category Hierarchy*[3] consisting of 421 POI types. These range from types such as *Mexican Restaurant* to *Police Station* or *Mountain Top*. To ensure the validity of the thematic similarity method proposed in the remainder of this paper, only those *types* that consisted of 30 or more POI instances in each region were included. Given the specificity of some types (e.g., College Cricket Pitch) this reduced the number of POI types to 321.

Next, *tips* were accessed for each POI in the dataset. *Tips* consist of unstructured, textual comments and reviews of a POI contributed by users of Foursquare. To ensure a fair representation of the POI types, only those types to which 30 or more tips were contributed per region were included in our analysis. Otherwise some POI type may have been represented merely by a small number of terms. Cleaning the data in this way further reduced the number of POI types from 321 to 316.

3.2 Thematic Signatures

A topic modeling approach was taken to generate thematic signatures on which to measure the variability of POI types across regions. These signatures model the fact that people are likely to write about cocktails and loud music after visiting a nightclub, and contribute reviews about hiking routes, waterfalls, and camp grounds when visiting state parks. *Latent Dirichlet allocation (LDA)* (Blei et al. 2003) was employed through the use of the MALLET Toolkit (McCallum 2002) to generate a range of topics on which the regional similarity of POI types could be assessed. *LDA* is an unsupervised, generative topic model that takes a bag-of-words approach to organize content. In this case, all of the unstructured textual data (tips) are grouped together by POI type and region (316 types × 3 regions). The co-occurrence of words across these documents is examined, exposing latent topics within the data. A probability distribution of these topics is returned and can then be used to thematically define each type split by region. Since the topics remain the same across all types, the similarity of POI types and regions can be measured through calculating the (dis)similarity between probabilistic topic distributions.

[3] https://developer.foursquare.com/categorytree.

4 Regional Variation

In this section, methods for exploring regional variation between different POI types are presented. The first step involves determining whether some types are regionally *invariant* (aspatial) while others are regionally *variant*. Once this has been determined, the degree to which each type is influenced by regional variations is studied as well as the sensitivity of the type to regional nuances. Last, regional type variability is abstracted to the top level of the POI type hierarchy with the goal of determining whether or not regional variability transcends hierarchy levels. The thematic signatures on which the analysis is based are constructed from 65 LDA topics. Further discussion on the resolution of thematic signatures is given in Sect. 7.

4.1 Significance of Regional Variations

The first task, as outline in **RQ1**, is to investigate the possibility that regional variations in thematic signatures are simply a sampling artifact and merely the result of random variation. To test this, the null hypothesis is defined stating that all POI are regionally invariant in term of their thematic signatures. Using the χ^2 *Goodness of Fit Test* (Bentler and Bonett 1980), this hypothesis can be tested. The χ^2 Goodness of Fit Test involves comparing two distribution samples (thematic signatures) and determines the amount by which the two are statistically different. Equation 1 shows the comparison of two thematic signature distributions, P and Q, divided by the variance, σ^2, of the observation.

$$\chi^2(P, Q) = \sum \frac{(P - Q)^2}{\sigma^2} \tag{1}$$

Using thematic signatures (here topic distributions) of the same POI type from two different regions, variability of said type can be modeled via the p-value reported from the χ^2 test. Table 1 shows the percentage of the 316 POI types that are regionally variant split by regional pair and level of significance (0.1, 0.05, 0.01).

The highest percentage of regionally variant POI types exists between Los Angeles and Chicago at roughly 48% followed by New York and Los Angeles at around

Table 1 Percentage of POI types that are statistically different between regions as determined by the *Chi Squared Goodness of Fit Test*. The results for three p-values (0.01, 0.05, 0.1) are reported

	0.01	0.05	0.1
NY and LA (%)	29.7	32.9	33.0
NY and CHI (%)	22.5	25.0	25.6
LA and CHI (%)	46.5	47.5	48.7

Table 2 Agreement between regions on variant (0) and invariant (1) POI types, shown with percentage of overall POI types. Based on a χ^2 p-value of 0.05

NY and LA	NY and CHI	LA and CHI	Percentage (%)
1	1	1	11.1
0	1	1	6.3
1	0	1	9.5
0	0	1	20.3
1	1	0	4.4
0	1	0	2.8
1	0	0	6.6
0	0	0	38.6

33%. A discussion on possible explanations for these regional differences is presented in Sect. 5. These statistically variant regional results can be broken down further to look at the agreement between regions. Focusing specifically on the 0.05 χ^2 level (Table 2), we find that close to 50% of POI types are either regionally invariant or variant across all the three U.S. cities (11.1% invariant and 38.6% variant). The complimentary 50% is split between some combination of agreement between one or two of the regional pairs.

These results confirm our intuition and reject the null hypothesis stated in **RQ1**. There is regional variation in POI types and it is unlikely caused by random fluctuations, but rather by statistical differences in the thematic signatures. This finding points to a difference in words and language choices between regions which is in accordance with numerous existing studies (Tuan 1991; Johnstone 2004; Cheng et al. 2010; Graham and Zook 2013). Interestingly though, not all POI types are shown to vary by region implying that the linguistic characteristics of certain POI types are less regionally specific. This is important as it implies that features extracted from global datasets will not perform well for certain types thereby affecting tasks such as classification. Given that we have shown that some types do vary by region, but others do not, the next question asks *which* POI types are regionally (in)variant and by what amount?

4.2 Variability Between POI Types

RQ2 follows up on the previous section by examining the ways in which POI types differ between regions and the amount by which some types vary regionally while others remain invariant. Two different methods are used to assess the (dis)similarity between POI type using their thematic signatures.

4.2.1 Jensen-Shannon Divergence

The Jensen-Shannon Divergence (JSD) is a method for measuring the dissimilarity between two probability distributions *(P, Q)* (Lin 1991). This method takes a one-to-one bin matching approach to comparing discrete datasets. The *distance metric* calculated here is computed by taking the square root of the divergence shown in Eq. 2. The metric is finite, bounded between 0 and 1. *KLD* represents the *Kullback-Leibler Divergence* and is specified in Eq. 3.

$$JSD(P \parallel Q) = \frac{1}{2}KLD(P \parallel M) + \frac{1}{2}KLD(Q \parallel M) \tag{2}$$

$$KLD(P\|Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)} \tag{3}$$

Table 3 shows the top five most dissimilar POI types as well as the top five most similar types split by regional pairs. Note that while there are difference between regions, a number of types are found across regions. *Monument/Landmark* is listed in the top five dissimilar POI types for all region pairs and *Nail Salon* is in all similar POI groups. Using Kendall's Coefficient of Concordance *W* (Kendall and Smith 1939) to calculate the agreement between the *JSD* dissimilarity values across the three pairs of regions resulted in a value of 0.9 ($p < 0.01$) indicating that there is strong agreement in dissimilarity in terms of POI types between region pairs.

Table 3 Top five and bottom five dissimilar POI types based on normalized Jensen-Shannon Divergence and split by region pairs

NYC and LA	NYC and CHI	CHI and LA
Dissimilar POI types		
Football Stadium (0.90)	Vineyard (0.90)	Football Stadium (1.00)
Rest Area (0.77)	Meeting Room (0.84)	Meeting Room (0.93)
Plaza (0.76)	Independent Theater (0.78)	Vineyard (0.90)
Monument/Landmark (0.63)	Assisted Living (0.75)	Mountain Top (0.89)
Mountain Top (0.62)	Monument/Landmark (0.73)	Monument/Landmark (0.78)
Similar POI types		
Airport Lounge (0.00)	Nail Salon (0.02)	Mobile Phone Shop (0.00)
Nail Salon (0.05)	Pet Store (0.04)	Office Supply Store (0.03)
Barbershop (0.07)	Automotive Shop (0.06)	Nail Salon (0.05)
Dentist's Office (0.10)	Airport Lounge (0.06)	Electronics Store (0.07)
Automotive Shop (0.10)	Frozen Yogurt (0.08)	Pet Store (0.07)

4.2.2 Cosine Similarity

Cosine similarity is a measure that reports the similarity between two vectors, or distributions in this case (Eq. 4). Using the Euclidean dot product, the cosine of two vectors (P and Q) is computed. Provided non-negative values for P and Q, the resulting similarity value is bounded between 0 and 1.

$$CosSim(P, Q) = cos(\theta) = \frac{\sum_{i=1}^{n} P_i \times Q_i}{\sqrt{\sum_{i=1}^{n} (P_i)^2} \times \sqrt{\sum_{i=1}^{n} (Q_i)^2}} \tag{4}$$

Table 4 again shows the top five most dissimilar types as well as the top five most similar types split by regional pairs. Note that cosine similarity is a *similarity measure* and the values reported in the table are actually $1 - CosSim$ in order to mirror the dissimilarity values reported by the JSD. While there are clear differences between regions, a number of POI types are found between region pairs. In line with the types reported using JSD (Table 3), *outdoor* types such as *Scenic Lookout* and *Monument/Landmark* appear in the top dissimilar POI types. Comparatively, types related to *shopping* and *service* type activities such as *Optical Shop* or *Nail Salon* appear in the top similar types list. Applying Kendall's W again, the agreement between the cosine similarity values across the three pairs of regions resulted in a value of 0.5 ($p < 0.01$). While not as strong as the coefficient reported by *JSD*, this value still indicates agreement in similarity across types between region pairs.

Table 4 Top five and bottom five dissimilar POI types based on normalized cosine similarity and split by region pairs. Note that to align with JSD, all values reported in this table are computed as $1 - CosSim$

NY and LA	NY and CHI	LA and CHI
Dissimilar POI types		
Park (0.99)	Greek Restaurant (1.00)	Military Base (0.98)
Platform (0.98)	City Hall (0.97)	Platform (0.98)
Football Stadium (0.91)	Scenic Lookout (0.97)	Mid. East. Restaurant (0.95)
Gay Bar (0.91)	Monument/Landmark (0.92)	Conference Room (0.93)
Baseball Stadium (0.85)	Beach (0.90)	Water Park (0.87)
Similar POI types		
Car Wash (0.00)	Synagogue (0.03)	Airport Lounge (0.01)
Dessert Shop (0.03)	Women's Store (0.05)	Motel (0.02)
Accessories Store (0.12)	Shoe Store (0.19)	Cemetery (0.20)
Trade School (0.15)	Credit Union (0.24)	Women's Store (0.27)
Optical Shop (0.18)	Optical Shop (0.26)	Nail Salon (0.28)

Table 5 Kendall's coefficients of concordance W between Jensen-Shannon Distance and Cosine Similarity for pairs of regions

Region-pair	Kendall's-W	P-value
New York and Los Angeles	0.59	<0.02
New York and Chicago	0.56	<0.07
Los Angeles and Chicago	0.57	<0.05

4.3 Concordance Between Dissimilarity Measures

While JSD and CosSim are both highly popular in the LDA literature, they originate from different families of similarity measures and thus reflect different views on what *similarity* means. Hence, the two (dis)similarity measures produce individual results for inter-signature comparison. The value of these approaches is shown through their agreement. As was done between region pairs within each method, Kendall's coefficient of concordance *W* is used. Each of the three regions is compared to each other region using Jensen-Shannon Distance and Cosine Similarity. These methods produce single similarity (or dissimilarity for JSD) values for each region pair and for each POI type. *Kendall's W* is then used to calculate the concordance between these measures across all POI types.

A Kendall's W value of 1 indicates complete concordance where a value of 0 represents no concordance at all. As shown in Table 5, all *W* values are greater than random. This indicates a significant level of agreement between dissimilarity measures, thus reducing the possibility that the discovered similarities are simply artifacts of choosing one specific measure. Given these findings, we focus on JSD for the remaining analysis.

4.4 Hierarchy Homogeneity

Many Point of Interest type vocabularies are constructed as hierarchies[4] with subsumption relationships between subtypes and supertypes. The Foursquare POI type vocabulary follows this model by mapping every type to one of three distinct levels. The top-level into which all subtypes are assigned consists of 9 *supertypes*. For example, *Mexican Restaurant* is a subtype of *Food* and *Scenic Lookout* is a subtype of *Outdoors and Recreation*. These subsumption relationships are essential for many aspects of knowledge organization and play an important role in understanding POI type variability. Here we discuss how the regional variability of types traverses levels in this hierarchy.

As shown in Tables 3 and 4, the *most* and *least* regionally variant types are seen to be different. The most regionally variant POI types are primarily related to outdoor

[4] e.g., Schema.org or the place hierarchy used by the Ordnance Survey.

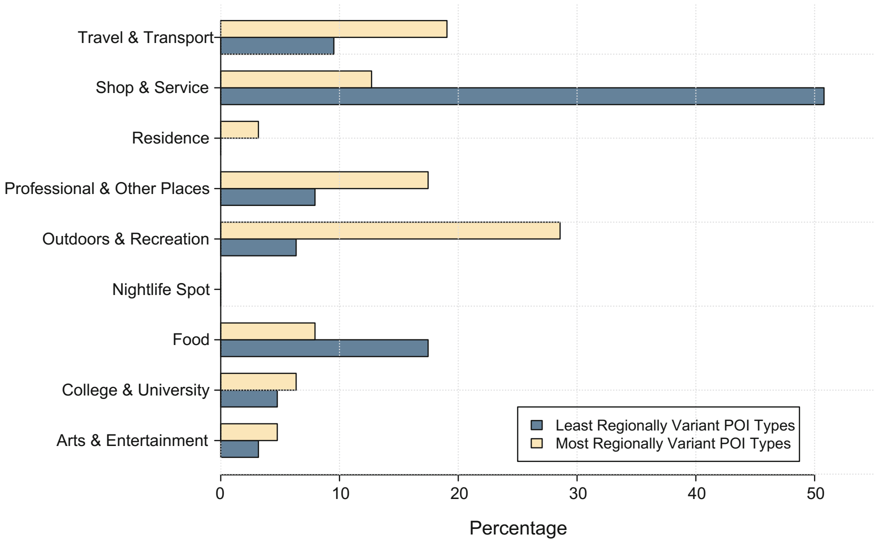


Fig. 1 Top level class for the top 20% (63) regionally dissimilar POI types as well as the top 20% (63) regionally similar POI types

activities while the most regionally invariant types can be typically categorized as shops, stores or service-related places. Here we extracted the top 20% most regionally variant and 20% least regionally variant POI types (as determined by *JSD*) and grouped them by their supertypes. Figure 1 shows the distribution of these supertypes grouped by their level of regional variability. The most invariant types clearly consist of primarily *Shop and Service* types while the most regionally variant POI types have a higher *information entropy*, distributed more evenly across the parent supertypes with peaks in *Outdoors and Recreation* and *Travel and Transport*.

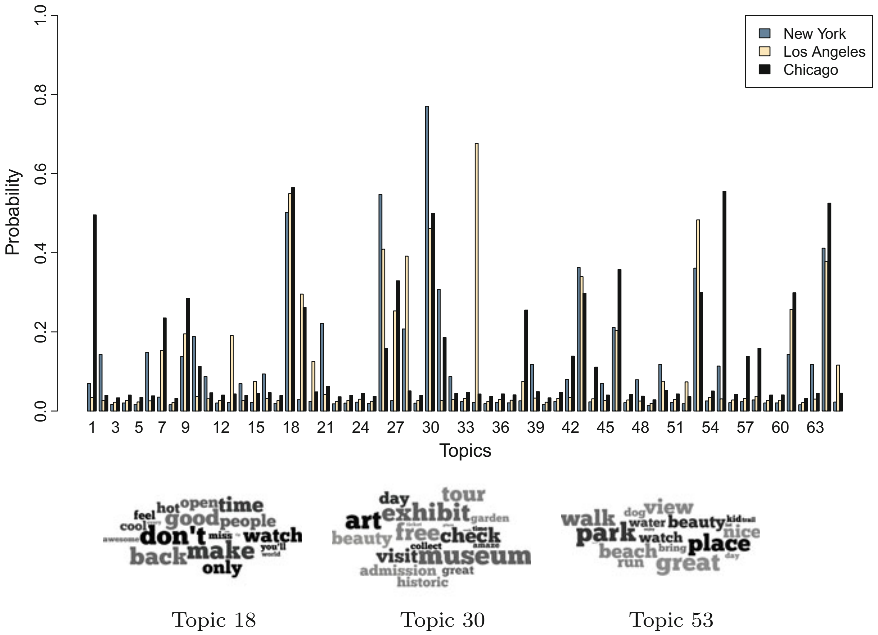
In summary and with respect to the second part of **RQ2**, POI hierarchies, such as Foursquare's are not completely homogeneous with regards to the regional (in)variability of POI types. There are clear differences in the top-level types that are regionally variant and those that are not. This is important as it implies that there are larger sets of types that do not vary significantly across space and thus can be learned from global datasets, while POI related to travel and outdoor themes need to be approached from a more local perspective.

5 Exemplary Investigation of Differences in Thematic Signatures

In this section, a subset of the POI types are examined further with the aim of better explaining the regional variability, or lack thereof, within and between types.

To gain a better understanding of the regional variability results presented in Table 1, one must understand the regions of interest. Relative to the layout of the United States, the cities of New York, Los Angles and Chicago vary greatly in their geographic locations. The climate deviates significantly between these cities with Los Angeles in a Mediterranean climate and New York and Chicago consisting of humid/continental climates. These climates allow for very different types of activities, specifically those *outdoor* activities that take place at POI types shown to have high variability (see Fig. 1). In addition to climate, there are important cultural differences between these regions. New York City, the "Gateway to America," is often viewed as a *melting-pot* of diverse cultures from around the world. Los Angeles, while still a multi-cultural city, is composed of a large Latino immigrant population with very different traditions and cultural backgrounds.

The POI type that consistently shows the highest level of thematic dissimilarity (regional variance) across all regions according to the *JSD* metric is *Monument/Landmark*. Figure 2 show the topic distribution for this type across the three U.S. cities using 65 LDA topics. Note that for visualization purposes the cube root of the distribution values are shown in order to make the very low topic values visible. Three of the most prominent topics are displayed as *word clouds* below the distribution. Not surprisingly, words such as *exhibit*, *visit* and *admission* appear to



Topic 18 Topic 30 Topic 53

Fig. 2 The POI type *Monument/Landmark* depicted as a probabilistic distribution of topics, split by region. The top words contributing to the most prevalent topics are shown below the graph. Note for visibility, as most values are close to zero, the cube root of the data is shown

contribute significantly to thematically defining this POI type. In comparing the topic distributions by region, we see that there is considerable disagreement between the regions as to which topics contribute to the type. Topics 1 and 55 contribute strongly to this type in Chicago while Topics 13 and 34 are more important, and unique, for defining *Monuments* in Los Angeles.

A *Monument or Landmark* is, almost by definition, something that is unique to the region in which it exists. Hence, the terms and language used to describe a landmark such as *The Hollywood Sign* in Los Angeles, CA are quite different than those used to describe the *Cloud Gate* in Chicago, IL or the *Grand Central Terminal Clock* in New York City, NY. Not only does this type invite the use of regionally specific terms and linguistic characteristics from locals, but POI of type *Monument/Landmark* are also the focus of many tourists or visitors to a region. These non-locals often visit these places from locations outside of the United States and bring with them their own unique descriptive terms and phrases.

An alternative POI type example is one that is highly regionally invariant. The type *Nail Salon* fits this description and is shown to have one of the lowest regional variation values reported by both *JSD* and *CosSim*. As opposed to the type *Monument/Landmark*, Fig. 3 visually depicts a higher topic probability agreement between regions. In most cases, the three U.S. regions agree on the prominent topics contributing to thematically defining *Nail Salons*. As shown in the word clouds below the distribution graph, topics 25, 43 and 61 use words such as *nail*, *hair*, *service* and

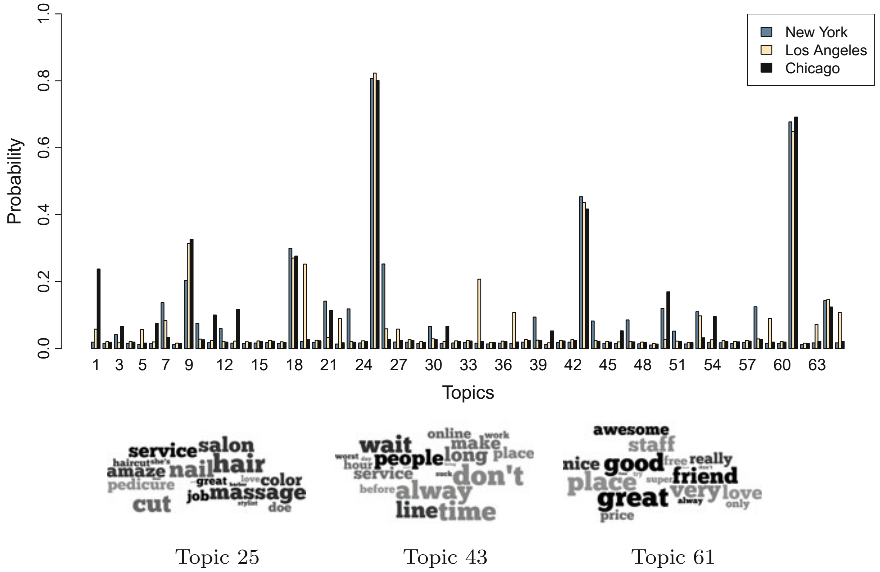


Fig. 3 The POI type *Nail Salon* depicted as a probabilistic distribution of topics, split by region. The top words contributing to the most prevalent topics are shown below the graph

staff to define this type. Given the agreement in topic probability across regions it is not surprising that this type is considered highly regionally invariant.

Nail Salons fit into what is often referred to as the *Service Industry*. Many service industry types appear in the regionally invariant group and most of them are based on activities that occur with regular frequency. The type *Nail Salon* is prototypical of this group as it involves an activity that occurs semi-regularly and traditionally involves a specific customers group, e.g., defined by gender and socio-economic status (Kang 2010). In addition, the focus of the services conducted at this POI type are highly specific to a part of the human body. In contrast to *Monument/Landmark*, this implies that the terms and words used to describe the type are more focused on certain theme (of nails) and less influenced by the geographic region in which the POI exists.

6 International Regional Variability

Up to this point, the focus of this research has been on understanding the regional variability of textual descriptions of POI types *within* the United States. The next step is to examine regional variation as it pertains to places outside of the United States. While the previous sections have shown that there are differences within a single country, **RQ3** questions the level of (dis)similar of POI types between countries. To answer this question, data from two cities outside of the United States were accessed. *Sydney, Australia* (SYD) and *London, England* (LDN) were selected based on the facts that (1) English is the primary language spoken and written in both regions, (2) technologically, both regions are quite similar to the United States wrt the use of location-based services, and (3) geographically, both regions are far away from the U.S. cities in this study.

The data accessed from Foursquare consists of 5,717 venues, 49,711 tips (SYD) and 8,179 venues, 132,193 tips (LDN). With the limitation that analysis requires a minimum of 20 venues and 20 tips per POI type, due to lighter usage of Foursquare outside the U.S., this reduced the number of types on which regional similarity could be calculated to 96 (SYD) and 172 (LDN). For this reason, these regions were not included in the original analysis, but instead are the focus of supplementary analysis. This subsequent work involved running a separate topic model with a reduced set of POI types in order to compare these new regions with those in the U.S.

6.1 POI Type Similarity Comparisons

Based on the JSD values reported for the three U.S. cities, 20 of the highest regionally variant types were selected as well as 20 of the highest regionally invariant types. POI types such as *Nail Salon*, *Pet Store* and *Doctor's Office* are examples of high regionally invariant types while *Monument/Landmark*, *Scenic Lookout* and *Skate*

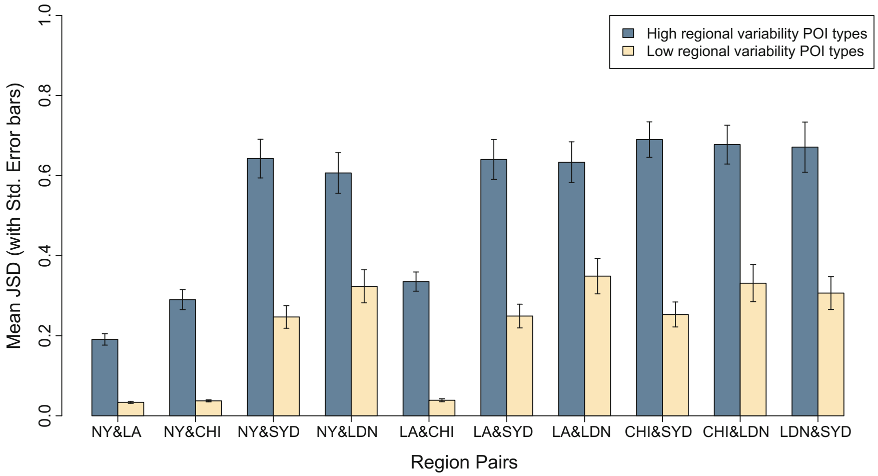


Fig. 4 The mean *JSD* of the top 20 regionally dissimilar POI types and top 20 regionally similar POI types along with standard error bar split by region pairs

Park contributed to the group of high regionally variant types. Not all types that appeared in the top intra-U.S. JSD values were chosen as, for example, types such as *New American Restaurant* had insufficient data.

The Jensen-Shannon distance was calculated between POI of the same type across all five pairs of regions (New York, Los Angeles, Chicago, Sydney and London). The JSD values were then averaged across the 20 regionally variant and the 20 regionally invariant types independently. Figure 4 shows these *mean JSD* values for each group of types split by region pair. The JSD dissimilarity values between pairs of U.S. cities is notably less than the international comparisons, both for the high regionally variable types as well as the high regionally invariant. One should also note that those POI types that are influenced by region changes within the US are also highly regionally influenced when compared internationally, and vice versa.

In response to **RQ3**, this demonstrates that regional (in)variability with respect to POI types exists both within the U.S. and internationally. Furthermore, highly variable types within the U.S. are also highly variable when compared to regions outside of the U.S. and this trend holds for regionally invariant types as well. These results have strong implications for the development of global POI thematic signatures suggesting that at least a subset of types can be sufficiently described by such global signatures even at an international level.

7 Thematic Resolution

One of the more difficult aspects with taking an unsupervised topic modeling approach to exploring thematic similarity is choosing and justifying the number of topics used in such analysis. Existing work in this area (Arun et al. 2010; Griffiths and Steyvers 2004) has presented approaches for choosing the number of topics to use in LDA topic modeling. The finding in both of these cases is that determining the number of "correct" topics is often complex, hard to validate, and dependent on the underlying dataset.

In this section, we take an exploratory analysis perspective to understand the affect of changing the number of topics on the similarity of POI types. Specifically we are interested in examining the process through which a corpus of words contributed to a type, such as *Park*, begins to merge with, and possible become indistinguishable from all other types in a sample dataset. Furthermore, how do types differ from one another as the number of topics increase (**RQ4**)? To conduct this analysis, the text from all tips across the three U.S. regions were collected at the *instance* level and tagged by their POI type. A series of LDA topic models were constructed from this data ranging in topic number from 10 to 500 in increments of 10. The resulting place type topic distributions were then analyzed in two different ways. First, Kullback-Leibler Divergence (KLD) similarity was calculated *within* a specific type (e.g., Park). Similarity was measured from each instance of the type to each other instance of the same type. Second, KLD similarity was measured from each instance of a given type (Park, in this example) to all other instances of *other* types (e.g., Donut Shop, University). Computing these two sets of KLD values for each type and for each topic number allowed us to examine how the number of topics influences *within* type similarity and *between* type similarity. Figure 5 shows a sample of five of these POI type KLD densities split by the number of topics.

These densities show a number of interesting phenomena. First, we find that regardless of the POI type, very low numbers of topics show a relatively high peak of similarity within the given type (peak around 0 on the X-axis). As the number of topics increases, the dominant peak at 0 for *within* place types decreases, in all cases. While we see a small decrease in peakedness of the *between* POI types as the number of topics increases, it is less pronounced indicating that *between* POI type similarity, at least for these example cases, is less influenced by a change in the number of topics.

There is a notable difference when comparing POI types to one another. All types show some percentage of overlap between the *within* and *between* KLD densities but the amount of overlap and magnitude differs across POI types. *Great Outdoors* for example shows a high amount of overlap at all number of topics with extremely high overlap in the high number of topics (e.g., 300 and 500). *Yoga Studio* on the other hand shows a relatively low amount of overlap. These overlap values are shown in greater detail in Fig. 6.

This figure shows all of the POI type overlap ratios as gray lines with our select five types highlighted. It was constructed by calculating the amount of overlap in the

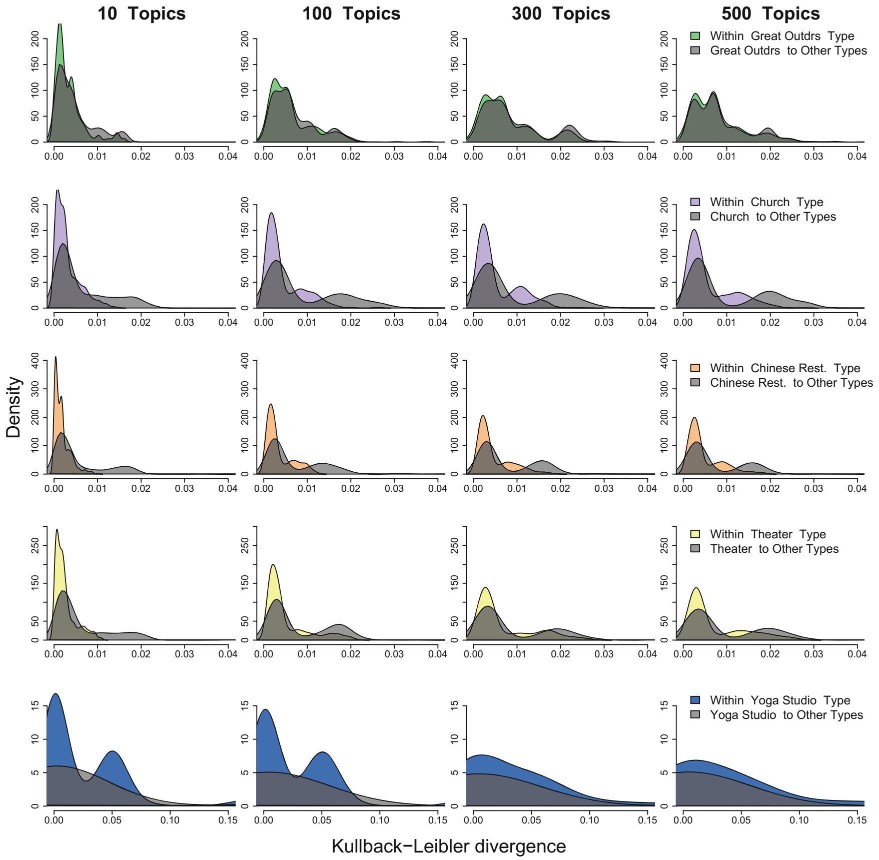


Fig. 5 Kullback-Leiber Divergence density graphs for five types shown as the number of LDA topics increase. The *colored graphs* show density graphs for KLD values for POI instances within the type. The *gray density graphs* show KLD values for POI instances within the listed type to all place instances not associated with the given POI type. Note that the Y-axis is different for each row in this figure

within and *between* KLD density plots. The integral, or the area under the density curve, was computed for both the between and within KLD densities for all topics between 10 and 500 at 100 topic intervals. The within and between density integrals were merged for the total area and the *ratios* were calculated as the overlap divided by this merged value. As was shown in Fig. 5, we see a range of density overlaps across our sample types. *Great Outdoors* shows a consistently high ratio of overlap across all topics. This reflects the *catch-all* nature of this place type. In the Foursquare category hierarchy, *Great Outdoors* is a higher-level category. It is therefore not surprising that the terms used within the category would be very similar to terms used outside of the category. In contrast, we find the type *Yoga Studio* to have a low ratio of overlap. A *Yoga Studio* is a very unique type of place which focuses on a specific activity and

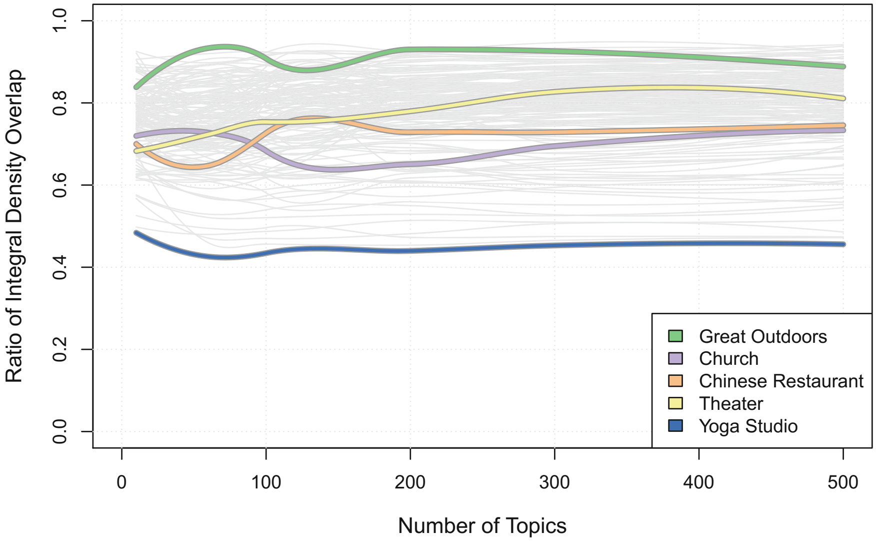


Fig. 6 Ratio of integral overlap between *within* type and *between* type densities shown varying by the number of topics. The *gray lines* show all POI types while the *colored lines* show a selection of five types

textual content about yoga studios are likely contributed from a narrow demographic of people. It is therefore not surprising that the overlap of words related to this type with terms from other types is relatively small.

Figure 6 also depicts a larger amount of overlap variance with smaller topic numbers than with larger ones. There is notable change in all types between 10 and 200 topics with an increase in the topic number after 200 showing much more limited influence. With respect to **RQ4**, we find that the influence of the number of topics, i.e., the resolution, on thematic similarity between POI types is dependent on the POI type itself and our findings confirm there is no hard and fast rule for determining the correct number of topics for place type similarity analysis.

8 Conclusions and Future Work

The terms used to describe places of interest play an important role in differentiating them from one another, thereby forming a prerequisite for classification, retrieval, and recommender tasks. Intuitively, reviews, social media postings, news, and so forth, about state parks are more likely to use terms such as trail, hike, and landscape, than descriptions of other places, say universities. This raises the important question how regional such type-based key characteristics are, i.e., whether we can learn a single embedding for POI types from global datsets or not. By constructing *thematic* *signatures* through topic models build on unstructured, user-generated text, we show that some types of places vary regionally, e.g., Monuments/Landmarks, while others do not, e.g., Nail Salons. Not only does this hold true for regions within the US, but also for many other English speaking regions. These findings are important as they speak to linguistic and cultural differences and similarities between people and the ways in which they interact with their environment. Furthermore, this work shows that regional variability does traverse levels in a place type hierarchy, e.g. generally, outdoor place types are more regionally variant, and that the resolution of the topics does impact the differentiation of place types.

Future work in this area will focus on exploring additional methods for constructing thematic signatures to better understand the robustness of the findings. Latent Dirichlet allocation has been shown to be a reliable model for studying similarities between places, but further research in this area will benefit from the incorporation of additional language-based similarity models. A method that combines these thematic signatures with previously constructed temporal signatures is currently underway with the goal of producing a robust set of *semantic signatures* for use in place-based activity behavior research.

References

Adams B, Janowicz K (2012). On the geo-indicativeness of non-georeferenced text. In: ICWSM, pp 375–378

Adams B, McKenzie G, Gahegan M (2015) Frankenplace: interactive thematic mapping for ad hoc exploratory search. In: 24th international world wide web conference, IW3C2

Arun R, Suresh V, Madhavan CV, Murthy MN (2010) On finding the natural number of topics with latent Dirichlet allocation: some observations. In: Advances in knowledge discovery and data mining. Springer, pp 391–402

Bao J, Zheng Y, Wilkie D, Mokbel M (2015) Recommendations in location-based social networks: a survey. GeoInformatica 19(3):525–565

Basso KH (1996) Wisdom sits in places: landscape and language among the Western Apache. UNM Press

Bentler PM, Bonett DG (1980) Significance tests and goodness of fit in the analysis of covariance structures. Psychol Bull 88(3):588

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022

Cheng C, Yang H, Lyu MR, King I (2013) Where you like to go next: successive point-of-interest recommendation. In: IJCAI, vol 13, pp 2605–2611

Cheng Z, Caverlee J, Lee K (2010) You are where you tweet: a content-based approach to geo-locating twitter users. In: Proceedings of the 19th ACM international conference on information and knowledge management. ACM, pp 759–768

Cresswell T (2014) Place: an introduction. Wiley

Gao H, Tang J, Hu X, Liu H (2013) Exploring temporal effects for location recommendation on location-based social networks. In: Proceedings of the 7th ACM conference on recommender systems. ACM, pp 93–100

Graham M, Zook M (2013) Augmented realities and uneven geographies: exploring the geolinguistic contours of the web. Environ Plan A 45(1):77–99

Griffiths TL, Steyvers M (2004) Finding scientific topics. Proc Natl Acad Sci 101(suppl 1):5228–5235

Hecht BJ, Gergle D (2010) On the localness of user-generated content. In: Proceedings of the 2010 ACM conference on computer supported cooperative work. ACM, pp 229–232

Hollenstein L, Purves R (2010) Exploring place through user-generated content: using flickr tags to describe city cores. J Spat Inf Sci 1(1):21–48

Hu B, Ester M (2013) Spatial topic modeling in online social media for location recommendation. In: Proceedings of the 7th ACM conference on recommender systems. ACM, pp 25–32

Johnstone B (2004) Place, globalization, and linguistic variation. Sociolinguist Var Crit Reflect 65–83

Kang M (2010) The managed hand: race, gender, and the body in beauty service work. University of California Press

Kendall MG, Smith BB (1939) The problem of m rankings. Ann Math Stat 10(3):275–287

Kinsella S, Murdock V, O'Hare N (2011) I'm eating a sandwich in glasgow: modeling locations with tweets. In: Proceedings of the 3rd international workshop on search and mining user-generated contents. ACM, pp 61–68

Lin J (1991) Divergence measures based on the shannon entropy. IEEE Trans Inf Theory 37(1):145–151

McCallum AK (2002) Mallet: a machine learning for language toolkit

McKenzie G, Janowicz K, Gao S, Gong L (2015a) How where is when? on the regional variability and resolution of geosocial temporal signatures for points of interest. Comput Environ Urban Syst 54:336–346

McKenzie G, Janowicz K, Gao S, Yang J-A, Hu Y (2015b) POI pulse: a multi-granular, semantic signatures-based information observatory for the interactive visualization of big geosocial data. Cartogr Int J Geogr Inf Geovis 50:71–85

Mülligann C, Janowicz K, Ye M, Lee W-C (2011) Analyzing the spatial-semantic interaction of points of interest in volunteered geographic information. In: Spatial information theory. Springer, Berlin, pp 350–370

Scellato S, Noulas A, Mascolo C (2011) Exploiting place features in link prediction on location-based social networks. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp 1046–1054

Shaw B, Shea J, Sinha S, Hogue A (2013) Learning to rank for spatiotemporal search. In: Proceedings of the sixth ACM international conference on web search and data mining. ACM, pp 717–726

Stefanidis A, Crooks A, Radzikowski J (2013) Harvesting ambient geospatial information from social media feeds. GeoJournal 78(2):319–338

Tanasescu V, Jones CB, Colombo G, Chorley MJ, Allen SM, Whitaker RM (2013) The personality of venues: places and the five-factors ('big five') model of personality. In: 2013 fourth international conference on computing for geospatial research and application (COM. Geo). IEEE, pp 76–81

Tuan Y-F (1991) Language and the making of place: a narrative-descriptive approach. Ann Assoc Am Geogr 81(4):684–696

Ye M, Janowicz K, Mülligann C, Lee W-C (2011) What you are is when you are: the temporal dimension of feature types in location-based social networks. In: Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems. ACM, pp 102–111

Distance Measure Based on Spatiotemporal Coexistence of Residents and Its Application to Urban Analysis

Toshihiro Osaragi and Ayaka Murakami

Abstract Distance measures, such as the Euclidean distance, network distance, and travel time, have been used in various urban models, such as models for transportation, land use, and human behavior. In this paper, we propose a novel distance measure based on the spatiotemporal distribution of residents, which is called the "Distance Measure based on Spatiotemporal Coexistence of Residents (*DM_SCORE*)". It is defined as the square root of the Jensen-Shannon Divergence, and can be calculated using person-trip survey data. We discuss the difference between the *DM_SCORE* and other distance measures, and apply it to urban models for the population density function, the spatial interaction model, and the spatial cluster analysis. The results of numerical analyses using actual data from Tokyo metropolitan area demonstrate that the *DM_SCORE* can be of great value in urban models.

Keywords Distance measure · Spatiotemporal coexistence · Person trip survey data · Jensen-Shannon divergence

1 Introduction

1.1 Research Background and Purpose

Conventionally, the concept of "distance to city center" has been mainly used in the huge number of theoretical urban models. This was because the urban activity at each location was empirically known to have a very strong relationship with the distance to city center. As urban transportation infrastructure has become more sophisticated, traffic network distance or time distance, rather than Euclidean

T. Osaragi (✉) · A. Murakami
Department of Architecture and Building Engineering, School of Environment and Society, Tokyo Institute of Technology, 2-12-1-M1-25 Ookayama, Meguro, Tokyo 152-8552, Japan
e-mail: osaragi.t.aa@m.titech.ac.jp

A. Bregt et al. (eds.), *Societal Geo-innovation*, Lecture Notes
in Geoinformation and Cartography, DOI 10.1007/978-3-319-56759-4_15

distance, have come to be used as an index to more precisely describe the distance to city center. However, in mainly numerical models, the attempts have been made to replace these by Euclidean distance due to the limitation of available data and the lack of computer resources.

Given these background, much research has been conducted on distance measures in urban space, and showed that there is a proportional relationship between the Euclidean distance and network distance between nodes in municipalities (Koshizuka and Kobayashi 1983; Gonçalvesa et al. 2014). Also, Tamura et al. (2000) proposed a method for calculating the correlation coefficient and regression coefficient for the Euclidean distance and network distance. Koto (1995) attempted to visualize adjacency relations by constructing a road network with edges whose length was proportional to the time required, as an indicator of the convenience of travel. Furthermore, Koshizuka and Kurita (1991) and Barhum et al. (2007) discussed methods for approximating the mean distance between pairs of points in a high-dimensional Euclidean space.

The relationships among the above three distance measures (Euclidean distance, network distance and time distance) have been discussed. The general consensus in the literature is that although network and time distance measures are superior when describing the movement of people and objects between two points, researchers have had to use the Euclidean distance due to limitations in the available data.

Recently, the spread of spatiotemporal data and the reduction in the cost of software that handles such data has made it possible to use network and time distance measures. However, due to a lack of transport modes or a low traffic density, there may be little flow of people or objects between places that are close in terms of distance. In other words, network and time distance cannot always sufficiently describe accessibility between points. The movement of people and objects between points is intricately related, not only to the state of transport infrastructure, but to a variety of other factors, and so discussing accessibility between points is not easy.

The spread of social networking services (SNS) in recent years has increased human interaction via the Internet. Online interaction is a factor that also influences the real movement of people and objects, and so there have been attempts to construct a new concept of distance based on online spatiotemporal simultaneity. For example, there is research that attempts to define distance between people (social distance) based on online social activity (people's social connections). Pham et al. (2011) defined social distance between any two people based on co-occurrence (being in the same place at the same time). Crandall et al. (2010) similarly divided space into cells, and estimated the probability of being in the same cell at the same time using Bayes' rule, to show that people who are socially connected are highly likely to make interrelated movements. There have also been attempts to represent accessibility between areas based on people's activity online, such as the research by Wakamiya et al. (2013), in which accessibility between areas was visualized on the basis of SNS data.

Against this background, this paper measures distance between areas based on the movement of people and the influence of accessibility. In other words, we define a new distance measure that takes into account the research findings of Pham et al. (2011) and Crandall et al. (2010) that people who spend time in the same area at the same time have a closer relationship with each other than those who do not. Specifically, focusing on the degree of opportunity for people living in different places to be present in the same area at the same time (hereafter, "spatiotemporal coexistence"), we propose the "Distance Measure based on Spatiotemporal Coexistence of Residents (*DM_SCORE*)", where the higher the spatiotemporal coexistence, the shorter the distance, and the lower the spatiotemporal coexistence, the longer the distance. We analyze urban space using the *DM_SCORE*, and examine its potential as a new distance measure.

1.2 Structure of This Study

Section 2 formulates the *DM_SCORE*, and describes a specific method for calculating *DM_SCORE* using the available data. Section 3 considers the relationship between the *DM_SCORE* and conventional distance measures, as well as clarifying the characteristics of the *DM_SCORE* based on the results of basic numerical calculations. Section 4 applies the *DM_SCORE* to classic urban models (population density function, spatial interaction model), and conducts comparative analyses, primarily with the Euclidean distance, to demonstrate the usefulness of the *DM_SCORE*. Section 5 performs spatial divisions using the *DM_SCORE*, and considers the characteristics of the area boundaries obtained. It also demonstrates that clearer aggregate results can be obtained by aggregating the results of social statistical surveys according to this spatial division. Section 6 summarizes this study and presents our conclusions.

2 Definition of New Inter-area Distance Measure

2.1 Formulation of the DM_SCORE

Figure 1 shows that the spatial distributions of people living in different places sometimes overlap and they are present in the same places at the same time during specific time period of a day. It is highly likely that such people are engaged in activities that are mutually related, and their social relationship is considered to be relatively close (Pham et al. 2011; Crandall et al. 2010). Therefore, we define a new inter-area distance measure based on the overlapping of the spatiotemporal distributions of residents, as described below.

The number of residents of an area i is represented by N_i. The number of these residents who stay in area k during time t is represented by $n_{ik}(t)$. In this case, the probability distribution for residents of area i at time t and area k, $p_{ik}(t)$, is obtained

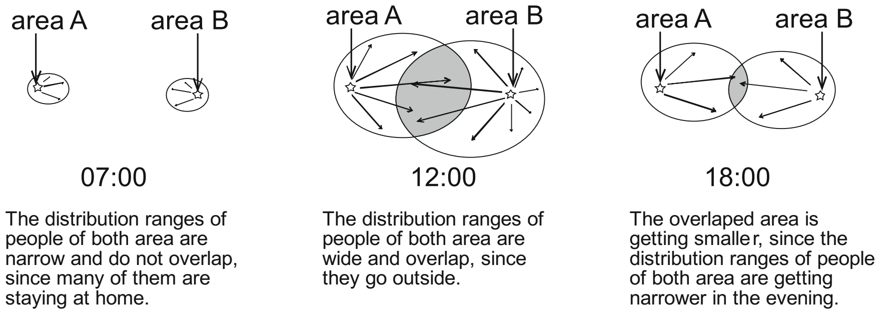

area A
area B
07:00
The distribution ranges of people of both area are narrow and do not overlap, since many of them are staying at home.
area A
area B
12:00
The distribution ranges of people of both area are wide and overlap, since they go outside.
area A
area B
18:00
The overlaped area is getting smalle r, since the distribution ranges of people of both area are getting narrower in the evening.

Fig. 1 Overlapping of the spatiotemporal distributions of residents

using the following equation. Area *j* can be defined similarly, and is written as $q_{jk}(t)$ to make it clearly distinguishable from area *i*.

$$p_{ik}(t) = \frac{n_{ik}(t)}{N_i}, \quad q_{jk}(t) = \frac{n_{jk}(t)}{N_j}. \tag{1}$$

These probability distributions can be regarded as the spatiotemporal distributions of residents living in areas *i* and *j*. In this case, the degree of opportunity for residents of areas *i* and *j* to encounter each other at the same time *t* in the same area *k* is equivalent to the similarity of the probability distributions at time *t* and area *k*, $p_{ik}(t)$ and $q_{jk}(t)$. Therefore, we consider the Jensen-Shannon Divergence (JSD), which is a measure of the distance between two probability distributions (Majtey et al. 2005; Lamberti et al. 2008).

In generally, the distance measure, D_{ij} which indicates distance between two areas *i* and *j*, is required to satisfy the following conditions:

$$\begin{aligned} &D_1 : D_{ij} \geq 0 \ (non - negativity), \\ &D_2 : D_{ij} = 0 \Leftrightarrow i = j \ (identity\ of\ indiscernibles), \\ &D_3 : D_{ij} = D_{ji} \ (symmetry), \\ &D_4 : D_{ij} \leq D_{ik} + D_{kj} \ (subadditivity/triangle\ inequality). \end{aligned} \tag{2}$$

The JSD does not satisfy the triangle inequality (D_4), one of the axioms of distance shown in Eq. (2). However, the square root of the JSD is known to satisfy D_4 (Endres and Schindelin 2003). Therefore, using the square root of the JSD and normalizing by taking time averages, the *DM_SCORE* is defined by the following equation (Briet and Harremoes 2009).

$$D_{ij} = \frac{1}{T} \sum_{t=1}^{T} \sqrt{JSD_{ij}(t)}, \tag{3-1}$$

here,

$$JSD_{ij}(t) = \frac{1}{2}\sum_{k=1}^{n}\left[p_{ik}(t)\log_2\frac{2p_{ik}(t)}{p_{ik}(t)+q_{jk}(t)} + q_{jk}(t)\log_2\frac{2q_{jk}(t)}{p_{ik}(t)+q_{jk}(t)}\right]. \tag{3-2}$$

Time t is naturally a continuous quantity, but here, for simplicity, it is approximated using discrete values (t = 1, 2, …, T). Also, $p_{ik}(t) = 0$ or $q_{jk}(t) = 0$ can be handled using the following equation.

$$\lim_{p\to 0} p \log p = 0. \tag{4}$$

As a result of the above, the *DM_SCORE* between areas i and j, D_{ij}, satisfies all of the axioms of distance D_1–D_4 shown in Eq. (2), and is therefore a distance measure.

2.2 Calculation Method for DM_SCORE Using Real Data

The number of residents who are stationary and the number who are traveling are investigated using time and location data (PT data) from the Tokyo Metropolitan Person Trip Survey (2008). Person Trip Survey aims at investigating the travel behavior in large cities of Japan, and has been carried out basically every ten years by the Ministry of Land, Infrastructure, Transport and Tourism of Japan. “Trips” defined in PT data are illustrated in Fig. 2, as an example (Osaragi 2016).

The principal information provided in PT data is shown in Table 1. They include personal attributes (age, gender, occupation), the position and time information of

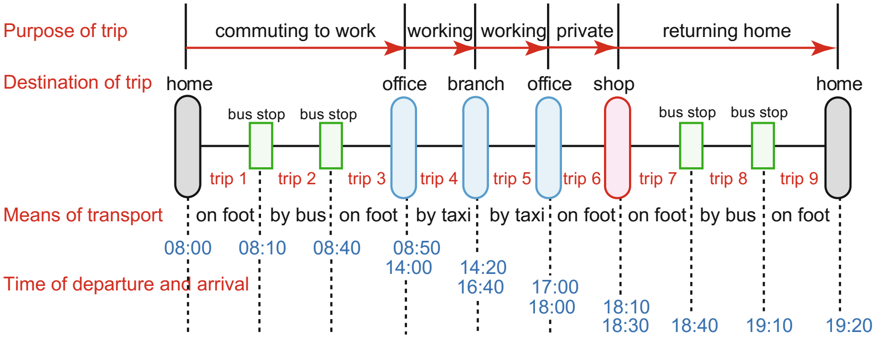


Fig. 2 Example of trips in Person Trip Survey data

Table 1 The information provided in Person Trip Survey data

Item	Contents
Regions subject to survey	Tokyo, Kanagawa, Saitama, Chiba and Southern Ibaragi Prefectures
Survey time and day	24 h on weekdays in October, 2008, excluding Monday and Friday
Object of survey	Persons aged 5+ living in the above region
Sampling	Random sampling based on census data (34,618,738 persons)
Data	733,873 samples (mean weighting coefficient is approximately 47.2)
Content of data	Personal attribute, position and time of departure/arrival, purpose of trip
Purpose of trip	Purpose of each trip (18 purposes: e.g., commuting to work/school, shopping, eating, etc.)
Means of trip	Means of trip (5 means: on foot, bicycle, car, bus, train, ship, airplane)

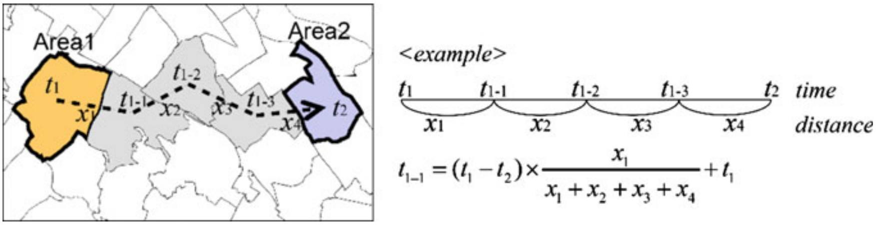

Area1
Area2
t_1
$t_{1\text{-}1}$
$t_{1\text{-}2}$
$t_{1\text{-}3}$
t_2
x_1
x_2
x_3
x_4
<example>
t_1 $t_{1\text{-}1}$ $t_{1\text{-}2}$ $t_{1\text{-}3}$ t_2 time
x_1 x_2 x_3 x_4 distance
$t_{1-1} = (t_1 - t_2) \times \frac{x_1}{x_1 + x_2 + x_3 + x_4} + t_1$

Fig. 3 Preprocessing of PT data for calculating spatiotemporal distributions of moving people

departure and arrival, purpose of trip (18 purposes: e.g., commuting to work/school, shopping, eating), and means of transportation (5 means: on foot, bicycle, car, bus, train, ship, airplane), etc. The area for this survey covers a range of 70 km radius centered on the Tokyo. About 1.2 million persons were selected from around 34 million residents by random sampling. The number of valid samples is 733,873.

Specifically, time *t* covers the period from 7:00 to 19:00, when people's movement increases. Area *i* is taken as a small zone (with a size of about 10 small administrative units in the city center) in the PT data. The spatiotemporal distribution of stationary people is found using time and location data during the period between arrival from a certain trip and departure on the next trip. Also, the spatiotemporal distribution of moving people is found by (1) creating a network connecting representative points in adjacent areas with straight lines, then (2) estimating trip paths (shortest paths) using this network, and (3) assuming that people move on these paths at a certain speed (Fig. 3).

The spatiotemporal distributions of stationary and moving people are found through the above procedure, and the *DM_SCORE* between areas *i* and *j*, D_{ij}, is calculated based on Eqs. (3-1) and (3-2).

3 Basic Analysis Using the *DM_SCORE*

3.1 Relationship with Conventional Distance Measures

We investigated the relationship between the *DM_SCORE* and conventional distance measures (Euclidean distance, network distance, time distance) (Fig. 4).

A weak positive correlation was found between the *DM_SCORE* and conventional distance measures. However, the value of the *DM_SCORE* may be large for neighboring areas, and conversely, may be small for areas that are far apart. In other words, despite having properties similar to conventional distance measures, the *DM_SCORE* is a distance measure that fluctuates according to the degree of human interaction.

3.2 Spatial Distribution of the DM_SCORE

3.2.1 Spatial Distribution from a Specific Area

We calculated the *DM_SCORE* from a specific area to another area, and found the spatial distribution (Fig. 5). The *DM_SCORE* for an area that includes Shinjuku

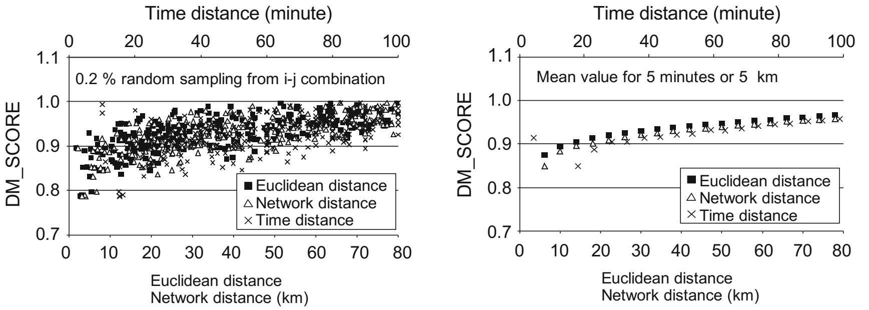


Fig. 4 Relationship between conventional distance measures and *DM_SCORE*

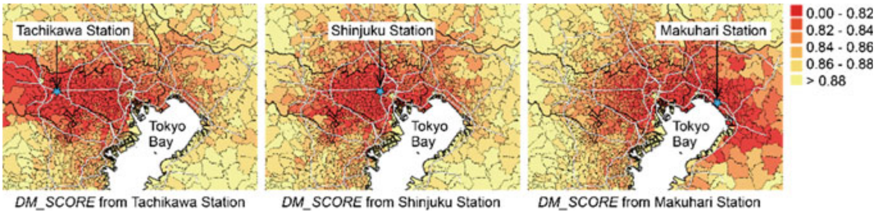


Fig. 5 Spatial distribution of the values of *DM_SCORE* from the specific areas

Station is distributed in a concentric pattern that is somewhat biased towards the west of Tokyo. The *DM_SCORE* for an area that includes Tachikawa Station fluctuates along the main railroad line. The distribution of the *DM_SCORE* values for an area that includes Makuhari Station is similarly affected by "railroad catchment areas". In other words, the *DM_SCORE* is closely linked to the direction of spatial migration of residents by modes of high-speed transport.

3.2.2 Mean *DM_SCORE*

For each area, we calculated the mean *DM_SCORE* to another area and then determined the spatial distribution (Fig. 6). These mean values spread out in order of increasing mean value in a concentric pattern centered on Tokyo's 23 wards. The areas with low mean values are all located in central Tokyo. A low mean *DM_SCORE* means that the *DM_SCORE* to many other areas is low. In other words, the mean *DM_SCORE* can be considered to correspond to "distance to the city center". The clusters obtained by estimating accessibility through a more conventional measure (e.g. Euclidean distance) would always be high in the peripheral areas and low in the center of the study area. In contrast, the clusters obtained by *DM_SCORE* are not dependent on the location of city center.

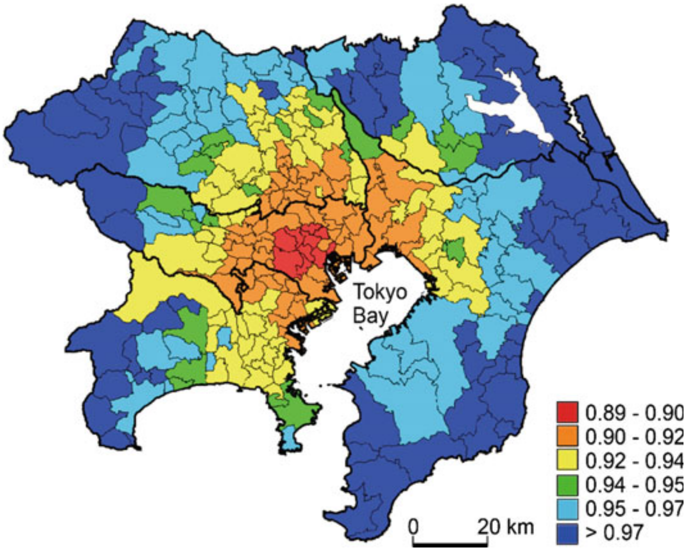


Fig. 6 Spatial distribution of mean *DM_SCORE*

4 Application to Urban Models

4.1 Application to Population Density Function

Below is a model equation describing population density ρ (people/km^2) using an exponential function of distance from the center, r (km), which is known as the Clark formula (A and b are unknown parameters) (Clark 1951).

$$\rho(r) = Ae^{-br}. \tag{5}$$

We applied the Euclidean distance from the Imperial Palace in Tokyo, which is often assumed as the center of Tokyo, and the mean *DM_SCORE* D_i as the distance in Eq. (5). Comparing the estimated results and real values (Fig. 7), the goodness-of-fit was higher when the mean *DM_SCORE* D_i was applied. In other words, the mean *DM_SCORE* D_i could potentially be used as a new index for expressing distance from the city center (accessibility to the city center).

4.2 Application to a Spatial Interaction Model

Here, we examine a case in which Euclidean distance is used and a case in which the *DM_SCORE* is used for the inter-area distance in a spatial interaction model. First, the spatial interaction model is described using the following equation.

$$T_{ij} = A_i B_j O_i D_j \exp[-\beta C_{ij}], \tag{6}$$

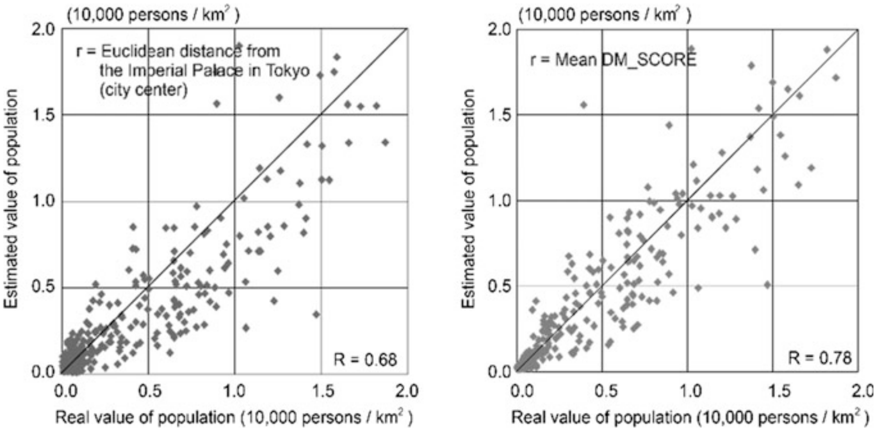


Fig. 7 Comparison of estimated results of population density using Euclidean distance and mean *DM_SCORE*

here, $C_{ij} = \begin{cases} C_{ij} \ (i \neq j) \\ C_o \ (i = j) \end{cases}$,

T_{ij} Number of trips between areas *i* and *j,*
C_{ij} Distance between areas *i* and *j,*
C_0 Distance within own area,
O_i Number of trips starting from area *I,*
D_j Number of trips arriving in area *j*

Next, the following three representative indices are used as goodness-of-fit indices for the estimation model, where *n* is number of areas.

(1) Root mean squared error (*RMSE*)

$$RMSE = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} \left(T_{ij} - \hat{T}_{ij}\right)^2 / m}, \quad where\, m = n \times n \tag{7}$$

(2) Dissimilarity index (*D*)

$$D = 50 \times \sum_{i=1}^{n} \sum_{j=1}^{n} \left| \frac{\hat{T}_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{n} \hat{T}_{ij}} - \frac{T_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{n} T_{ij}} \right|. \tag{8}$$

(3) Correlation coefficient (*R*)

$$R = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \left(T_{ij} - \bar{T}\right)\left(\hat{T}_{ij} - \bar{T}\right)}{\sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} \left(T_{ij} - \bar{T}\right)^2} \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} \left(\hat{T}_{ij} - \bar{T}\right)^2}}. \tag{9}$$

Here, it is noted that a good model cannot be estimated when the distance within own area C_{ii} is taken as 0 (Bharat and Larsen 2011; Kordi and Kaiser 2012). Therefore, the spatial interaction model is estimated by describing the model using the unknown parameter C_0, without taking the value of distance within own area C_{ii} to be 0. Specifically, parameters A_i, B_j, and β are found by a convergence calculation using Hyman's method (Wilson 1971), and the value of C_0 is estimated to minimize the value of the dissimilarity index *D* in Eq. (8).

4.2.1 Comparison of Estimated Results Using Differences Between Distance Measures

We compared a case in which Euclidean distance X_{ij} is used and a case in which the *DM_SCORE* D_{ij} is used as the inter-area distance C_{ij}. Looking at the goodness-of-fit indices and the relationship between the estimated trip values and real trip values, compatibility is higher when the *DM_SCORE* is used (Fig. 8).

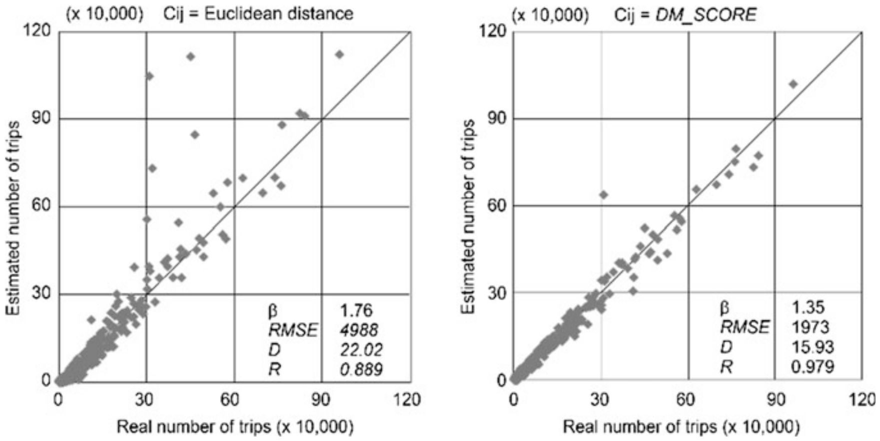


Fig. 8 Application of *DM_SCORE* to spatial interaction model

Table 2 Estimated results for trips by purpose

Purpose	C_{ij} = Euclidean distance				C_{ij} = *DM_SCORE*			
	β	*RMSE*	*D*	*R*	β	*RMSE*	*D*	*R*
1. Commuting to work	2.00	374	23.0	0.959	1.42	334	21.2	0.983
2. Shopping	3.73	236	13.4	0.986	1.92	233	12.7	0.992
3. Commuting to school	1.60	125	18.7	0.984	1.47	112	17.1	0.994
4. Going out on other private	3.30	115	14.7	0.984	1.87	109	13.4	0.994
5. Eating out/entertainment	3.28	126	15.3	0.986	1.74	134	14.7	0.990
6. Picking up/dropping off	3.91	55	13.1	0.919	2.06	61	13.2	0.993
7. Hospital visit	3.31	54	15.8	0.988	1.87	56	15.0	0.992
8. Meeting or conference	2.69	99	29.6	0.938	1.71	115	27.7	0.948
9. Going out on business	2.40	57	24.3	0.986	1.56	52	22.6	0.982
10. Sales/purchasing	3.13	42	29.7	0.978	1.77	49	31.3	0.959
11. Tourism/leisure	2.36	30	36.1	0.982	1.49	32	36.0	0.967
12. Work/repair	2.32	25	37.6	0.941	1.55	27	37.3	0.957
13. Rural/fishery work	3.85	5	10.2	0.924	2.35	5	9.5	0.994

Next, we estimated trips by purpose in a similar way. Table 2 shows the results. Each model describes the interaction between areas well. The value of β, which represents the distance resistance, exhibits characteristics of the trip purposes. For example, β is low in models in which the trip purpose is commuting to work or school, and people's movements are found to extend over a wide region. Meanwhile, β is high in models in which the trip purpose is shopping or picking up/dropping off people, and most trips are short distances centered on the home. For any of the evaluation indicators, the model using the *DM_SCORE* performs better than the model using the Euclidean distance.

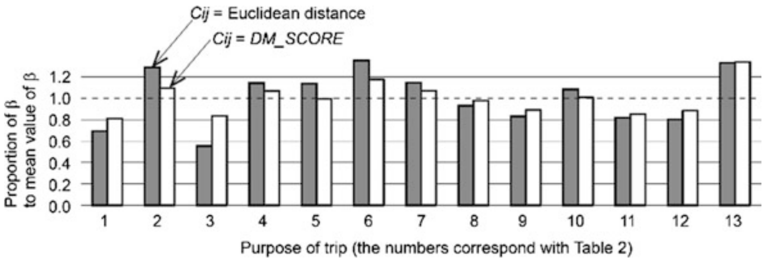


Fig. 9 Comparison of estimated parameters for trips by purpose

Furthermore, comparing β in the estimated models using Euclidean distance and the *DM_SCORE* (value of β compared to mean), β is higher in the model using the *DM_SCORE* than the model using Euclidean distance when the trip purpose is commuting to school (Fig. 9). This shows that, although commuting to school involves moving long distances similar to commuting to work, people move to areas with a relatively low *DM_SCORE*. Meanwhile, for shopping and picking up/dropping off people, etc., β is higher in the model using Euclidean distance than the model using *DM_SCORE*, which shows that this kind of travel is more strongly influenced by Euclidean distance. In this way, the influence of distance differs according to the purpose of travel, and this is of great interest.

4.2.2 Relationship with Distance Obtained from the Spatial Interaction Model

The following logit model can be constructed by taking the probability of residents of area i choosing area j as P_{ij}, the utility obtained by doing so as U_j, and the disutility required to move between areas i–j as C_{ij}.

$$P_{ij} = \frac{\exp[U_{ij}]}{\sum_k \exp[U_{ik}]}, \tag{10}$$

$$U_{ij} = U_j - C_{ij}. \tag{11}$$

By implementing a simple expansion of the expression, it is evident that this logit model is equivalent to the spatial interaction model (Aoki and Osaragi 1993).

$$\begin{aligned} T_{ij} &= O_i P_{ij} \\ &= \frac{1}{\sum_k \exp[U_k]\exp[-C_{ik}]} \cdot O_i \cdot \frac{\exp[U_j]}{D_j} \cdot D_j \exp[-C_{ij}] \\ &= A_i O_i B_j D_j \exp[-C_{ij}]. \end{aligned} \tag{12}$$

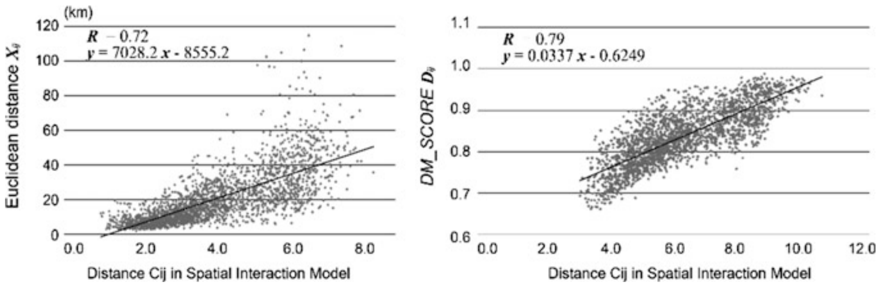


Fig. 10 Relationship between the distances estimated from spatial interaction model and Euclidean distance/*DM_SCORE*

Also, by finding the choice probability ratio Q_{ij} (odds ratio), and assuming that $C_{ii} = C_{jj} = C_0$, $C_{ij} = C_{ji}$, the distance C_{ij} in the spatial interaction model can be inversely estimated using the probability P_{ij}.

$$\begin{aligned} Q_{ij} &= \frac{P_{ij}P_{ji}}{P_{ii}P_{jj}} \\ &= \exp[(C_{ii} - C_{ij}) - (C_{jj} - C_{ji})] \\ &= \exp[2C_0 - 2C_{ij}], \end{aligned} \tag{13}$$

$$C_{ij} = \frac{1}{2}\log Q_{ij} + C_0. \tag{14}$$

Looking at the relationship between the distance C_{ij}, inversely estimated from the spatial interaction, with the Euclidean distance X_{ij} and the *DM_SCORE* D_{ij}, it is found that *DM_SCORE* D_{ij} has a higher correlation than the Euclidean distance X_{ij} (Fig. 10). It is difficult to theoretically relate the distance C_{ij} and D_{ij} directly (to derive one from the other) by expanding the expression, but when calculated numerically, it is evident that the two are positively correlated. One possible reason that the *DM_SCORE* is a better proxy for accessibility than Euclidean distance is that the *DM_SCORE* includes the individual trip data. We will confirm this point using other kinds of datasets in future.

5 Space Division Using the *DM_SCORE*

5.1 Example of Space Division

Spatial division of an urban space that extends continuously is often carried out by focusing on a certain property and taking areas where that property is similar as one group. Conventionally, this kind of spatial division has been frequently performed based on static data, such as nighttime population and number of offices. Below, we consider the possibility of dividing space using the *DM_SCORE* and applying the *DM_SCORE* to spatial division.

5.2 Method and Results for Spatial Division

We performed a cluster analysis (Ward's method) (Ward 1963) of the entire Greater Tokyo Area (275 municipalities) and Tokyo's 23 wards (265 small zones) using the *DM_SCORE* as a similarity index, and divided each into 15 areas (spatial clusters). Figure 11 shows the results.

5.2.1 Division Results for the Entire Greater Tokyo Area

Although the *DM_SCORE* contains no information about administrative boundaries, the boundaries of the spatial clusters obtained were consistent with administrative boundaries, such as boundaries between prefectures and wards (Fig. 11, left panel). Geographical barriers such as rivers existing on administrative boundaries define people's spatial migration, and as a result, human interaction is influenced by administrative boundaries, which is thought to be reflected in the *DM_SCORE*. It is very interesting that, although people move within cities without being aware of the administrative boundaries, the boundaries of the spatial clusters obtained clearly reflect the administrative boundaries. One possible reason for this interesting result is that residents use a ward office, a public junior high school, a public high school, and so on, which are located in the administrative district where

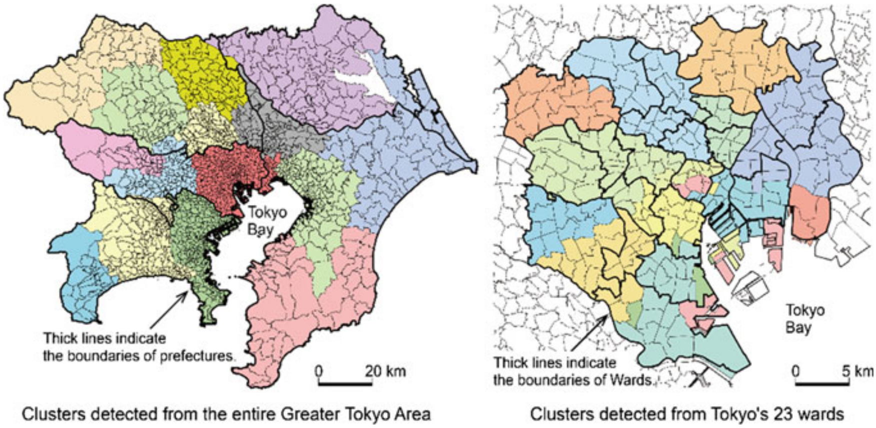

Tokyo Bay
Thick lines indicate the boundaries of prefectures.
0 20 km
Clusters detected from the entire Greater Tokyo Area
Tokyo Bay
Thick lines indicate the boundaries of Wards.
0 5 km
Clusters detected from Tokyo's 23 wards

Fig. 11 Space division using the *DM_SCORE* and administrative boundaries of prefectures or wards

they live. Namely, a part of their spatial movement is done inside the administrative boundary. This fact may lead to that *DM_SCORE* between the places included in the same administrative ward are shortened and they have been classified into the same cluster.

5.2.2 Division Results for Tokyo's 23 Wards

The division results for Tokyo's 23 wards (Fig. 11, right panel) similarly show that the boundaries of many spatial clusters are consistent with the ward boundaries. However, looking more closely, in some cases an individual ward forms one spatial cluster, or several wards form one spatial cluster, while in other cases, one ward is divided into multiple spatial clusters. These results show that "railroad catchment areas" are influential as a factor in the formation of spatial clusters. In other words, the movement of people within cities is deeply dependent on railroads, and the *DM_SCORE* between areas located in a region along the same railroad line is low, which means that areas in the same "railroad catchment area" form one spatial cluster.

5.3 Usefulness of Spatial Division in Social Statistical Surveys

Spatial clusters based on the *DM_SCORE* show the range of areas where human interaction is relatively active. In other words, the consciousness of residents living in the same spatial cluster could be relatively similar. In order to verify this, we tallied the results of questionnaire surveys in a spatial division using the *DM_SCORE* and a spatial division commonly used by local governments, and compared the results.

5.3.1 Summary of Social Statistical Survey and Spatial Distribution Application

The analysis was carried out using the results of a social statistical survey (2003 Housing Demand Survey). From a total of 96 items, we focused on 27 items relating to the consciousness of residents (Table 3). The consciousness of residents was surveyed using four options, which were 1: Satisfied, 2: Quite satisfied, 3: Somewhat dissatisfied, 4: Very dissatisfied.

We tallied the survey results based on a spatial division using the *DM_SCORE* and a spatial division established by local governments (Tokyo, Saitama Prefecture, Chiba Prefecture, Kanagawa Prefecture), and for each question, found and compared the entropy and the entropy of the number of residents.

Table 3 Question items of social statistics survey

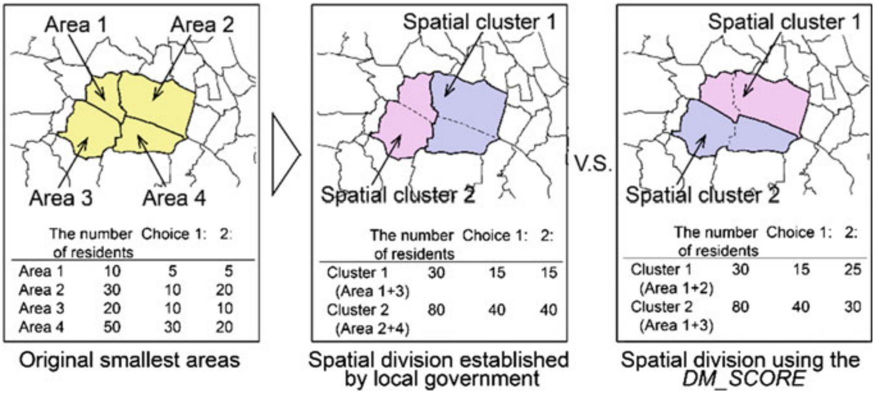
1	Comprehensive evaluation on the housing	15	Comprehensive evaluation of the living environment
2	Size, floor plan of the house	16	Safety against such as fire, earthquake, flood damage
3	Storage space	17	Safety of the road at the time of walking
4	Safety at the time of earthquake/typhoon	18	Security and prevention of crime
5	Safety of evacuation in case of a fire	19	Noise and air pollution
6	Security of the housing	20	Convenience of commuters
7	Damage of housing	21	Convenience of everyday shopping, medical care, welfare and cultural facilities
8	Ease of housing maintenance and management	22	Playgrounds and parks for children
9	Thermal insulation and air-tightness of the housing	23	Contact with nature as green, waterside
10	Energy- and cost-saving by cooling/heating	24	Allowance of the size of space, sunshine, fresh air
11	Care for the elderly, etc.	25	Landscape of the city
12	Ventilation performance	26	Relations with neighboring people and the community
13	Lighting of the main living room	27	Comprehensive evaluation of housing and living environment
14	Sound insulation against such noise from the outside		

Specifically, first, (1) data on the number of residents and the number of respondents per option was tallied for each spatial cluster (Fig. 12).

Next, (2) the entropy of each classification and each question option was found (Shannon and Weaver 1949) as follows. When the original data ($i = 1, 2, \ldots, n$) are divided into the m spatial cluster j ($j = 1, 2, \ldots, m$) for each choice k ($k = 1, 2, \ldots, l$), the average information content I_k of respondents' distribution of choice k is expressed by the following equation.

$$I_k = -\sum_{j=1}^{m} q_j^k \log_2 q_j^k, \text{ where } q_j^k = \frac{\sum_{i \in G_j} x_i^k}{\sum_{i=1}^{n} x_i^k}, \tag{15}$$

I_k The average information content of respondents' distribution of choice k,
G_j Set of i included in the space cluster j,
q_j^k Choice probability of choice k in the space cluster j,
x_i^k The number of respondents of choice k in area i



The number of residents		Choice 1:	2:
Area 1	10	5	5
Area 2	30	10	20
Area 3	20	10	10
Area 4	50	30	20

Original smallest areas

The number of residents		Choice 1:	2:
Cluster 1 (Area 1+3)	30	15	15
Cluster 2 (Area 2+4)	80	40	40

Spatial division established by local government

The number of residents		Choice 1:	2:
Cluster 1 (Area 1+2)	30	15	25
Cluster 2 (Area 1+3)	80	40	30

Spatial division using the *DM_SCORE*

Fig. 12 Example of the number of residents and the number of respondents per option in each spatial cluster

That is, the average information content of the distribution of respondents' distribution, *I,* is given by the following equation.

$$I = -\sum_{k=1}^{l} I_k. \tag{16}$$

Finally, (3) the entropy of questions was compared to the entropy of the number of residents. Percentage, R_c and R_g, of the average information content of each question item to the average information content of the number of residents are defined as follows (Osaragi 2002).

$$R_c = \frac{I_c}{I_c^*}, \quad R_g = \frac{I_g}{I_g^*}, \tag{17}$$

where

c The case of spatial division by *DM_SCORE,*
g The case of the spatial division of local government,
R_c, R_g Percentage of the average information content,
I_c, I_g The average information content of each question item,
I_c^*, I_g^* The average information content of the number of residents

The smaller the entropy of the number of residents, the more uniformly residents are distributed in the spatial division, and this is a desirable quality of a spatial unit in a questionnaire survey. On the other hand, the larger the entropy of questions, the clearer the spatial distribution of consciousness of residents (difference by area) can be understood, and this is a desirable quality. Therefore, we compared the value of the latter divided by the former (entropy ratio). Here, the entropy ratio based on the spatial division using the *DM_SCORE* is represented by R_c, and the entropy ratio based on the spatial division established by local governments is represented by R_g. When R_c is larger than R_g, the spatial division using the *DM_SCORE* can be said to be superior as a spatial unit for tallying social statistical surveys (surveys that identify the spatial distribution of consciousness of residents) compared to the spatial division established by local governments. In other words, for the items that satisfy the following formula, the spatial clusters based on *DM_SCORE* are suitable to understand the spatial distribution of the resident consciousness.

$$R_c > R_g. \tag{18}$$

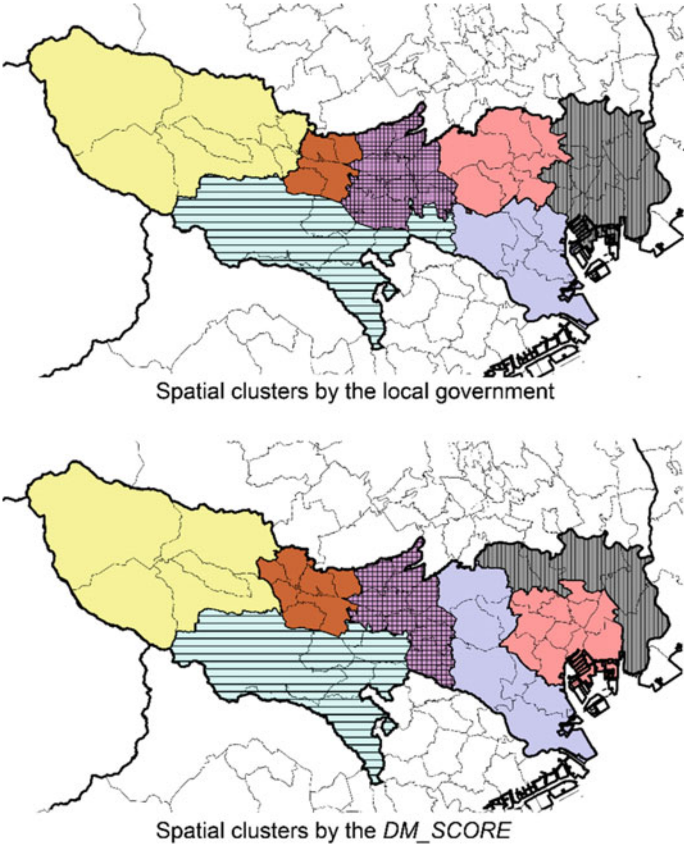
5.3.2 Usefulness of Spatial Division Using the *DM_SCORE*

Table 4 shows the numbers of municipalities and spatial clusters established by local governments. In order to compare similar cluster numbers, we divided the space using the *DM_SCORE* into the same number of spatial clusters as the number of clusters established by local governments. Figure 13 shows the results of the spatial division in Tokyo. Next, we performed a comparison of the entropy of questions and number of residents in each spatial cluster using the method explained above, and found the number of questions that satisfied the condition of $R_c > R_g$ [see Eq. (18)]. In other words, we calculated the percentage of questions in which the spatial division using the *DM_SCORE* was superior. The results showed that R_c was greater than R_g for approximately 90% of "satisfied" answers in Tokyo and the three prefectures (Table 4). This means that the spatial division using the *DM_SCORE* [Fig. 13 (bottom)] is better than the spatial division established by local governments [Fig. 13 (top)] when extracting regional biases in positive consciousness of residents relating to housing. However, the results for the negative answer of "very dissatisfied" were varied, and it is not possible to clearly determine which spatial division is superior.

The above results suggest that, when conducting social statistical surveys, there is a risk of losing sight of the characteristics of areas if the survey results are tallied by establishing the aggregate unit (space division) from an experiential perspective based on simple spatial adjacency relationships, administrative relevance, population size, industry dynamics, etc. In other words, the results suggest that movement within cities based on transport infrastructure, etc., not only increases opportunities

Table 4 Percentage of clusters which satisfy the conditions in Eq. (18) in the all question items

	Tokyo Metropolitan	Kanagawa Prefecture	Chiba Prefecture	Saitama Prefecture
The number of municipalities	53	56	61	79
The number of divisions by local government	7	8	11	10
1: Satisfied	92.59	92.59	85.19	88.89
4: Very dissatisfied	7.41	77.78	96.30	37.04



Spatial clusters by the local government

Spatial clusters by the *DM_SCORE*

Fig. 13 Spatial clusters by local government and by *DM_SCORE*

for encounters between residents, but also might influence the formation of various values and consciousness as well as solidarity. This is a sort of hypothesis for future research, but may be a subject that should be considered when conducting social statistical surveys.

6 Summary and Conclusions

In this paper, we defined a new inter-area distance measure, the *DM_SCORE*, based on overlapping of spatiotemporal distributions of residents associated with spatial migration. As well as examining the basic properties of the *DM_SCORE*, we attempted to analyze urban spaces using real data. Specifically, (1) we showed from an analysis using a population density function that the mean *DM_SCORE* is an excellent indicator of accessibility to the city center, (2) we applied the

DM_SCORE to a spatial interaction model and showed that it is an excellent distance measure for describing interactions, (3) we performed a spatial division using cluster analysis based on the *DM_SCORE*, and acquired the interesting result that the boundaries of the spatial clusters obtained are consistent with administrative boundaries, such as boundaries between prefectures and wards, and "railroad catchment areas", (4) we showed from an analysis using social statistical survey data that spatial division using the *DM_SCORE* is effective when extracting positive consciousness relating to housing, and demonstrated the possibility that the *DM_SCORE* is related to the similarity of values and consciousness, as well as solidarity, of residents.

The use of the *DM_SCORE* could make it possible to interpret urban spatial structures from a different perspective than that of conventional distance measures.

Acknowledgements The authors would like to acknowledge the valuable comments and useful suggestions from reviewers of AGILE Scientific Program Committee.

References

Aoki Y, Osaragi T (1993) Spatial influence model linking to logit model for the distribution of residential activities, spatial influence model for categorical explaining variable. J Archit Plan Environ Eng 444:97–103 (in Japanese)

Barhum K, Goldreich O, Shraibman A (2007) On approximating the average distance between points, APPROX and RANDOM 2007. In: Charikar M et al (eds) LNCS vol 4627, pp 296–310

Bharat PB, Larsen OI (2011) Are intrazonal trips ignorable? Transp Policy 18:13–22

Briet J, Harremoes P (2009) Properties of classical and quantum Jensen-Shannon divergence. Phys Rev A 79(052311):1–13

Clark C (1951) Urban population density. J R Stat Soc Ser A (General) 114(4):490–496

Crandall D, Backstrom L, Cosley D, Suri S, Huttenlocher D, Kleinberg J (2010) Inferring social ties from geographic coincidences. Proc Natl Acad Sci 107(52):22436–22441

Endres DM, Schindelin JE (2003) A new metric for probability distributions. IEEE Trans Inf Theory 49(7):1858–1860

Gonçalvesa DNS et al (2014) Analysis of the difference between the Euclidean distance and the actual road distance in Brazil. Transp Res Procedia 3(2014):876–885

Kordi M, Kaiser C (2012) A possible solution for the centroid-to-centroid and intra-zonal trip length problems. In: Proceedings of the 15th AGILE conference on geographic information science, Avignon, France

Koshizuka T, Kobayashi J (1983) Road distance and Euclidean distance. City Plan Rev 18:43–48 (in Japanese)

Koshizuka T, Kurita O (1991) Approximate formulas of average distances associated with regions and their applications to location problems. Math Program 52:99–123

Koto H (1995) Visualization of multi-city structure using time-distance network. Papers City Plan 30:553–558 (in Japanese)

Lamberti PW, Majtey AP, Borras A, Casas M, Plastino A (2008) Metric character of the quantum Jensen-Shannon divergence. Phys Rev A 77(052311):1–6

Majtey AP, Lamberti PW, Prato DP (2005) Jensen-Shannon divergence as a measure of distinguishability between mixed quantum states. Phys Rev A 72(052310):1–6

Osaragi T (2002) Classification methods for spatial data representation (CASA Working Papers 40). Centre for Advanced Spatial Analysis (UCL), London, UK

Osaragi T (2016) Estimation of transient occupants on weekdays and holidays for risk exposure analysis. In: 13th international conference on information systems for crisis response and management (ISCRAM 2016). In: Tapia et al (eds) Proceedings of the ISCRAM 2016 conference. ISBN 978-84-608-7984-8

Pham H, Hu L, Shahabi C (2011) Towards integrating real-world spatiotemporal data with social networks. In: Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems, pp 453–457

Shannon CE, Weaver W (1949) The mathematical theory of communication. University of Illinois Press, Urbana and Chicago

Tamura K, Koshizuka T, Ohsawa Y (2000) Relationship between road distance and euclidean distance on road networks. Oper Res Soc Jpn 234–235 (in Japanese)

Wakamiya S, Lee R, Sumiya K (2013) Urban proximity analysis with crowd movements based on location-based social networks. Inf Process Soc Jpn 6(3):159–176 (in Japanese)

Ward Jr JH (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58:236–244

Wilson AG (1971) A family of spatial interaction models, and associated developments. Environ Plan 3:1–32

When Granules Are not Enough in a Theory of Granularities

Ricardo Almeida Silva, João Moura Pires and Maribel Yasmina Santos

Abstract Several approaches have been proposed to model spatiotemporal phenomena at multiple LoDs, in particularly, under the granular computing research area, where a granularities-based model was proposed. Such model stands out from the related literature, but has two major limitations. On one hand, it has difficulties for describing regions, intervals of time, among others complex descriptions, and on the other hand, the generalization process is the same whether it is generalizing spatial, temporal or other features of a phenomenon. These problems reduce its applicability. To overcome such limitations, this paper extends the granularities-based model by introducing the granular term concept. We apply this concept to represent time instants and intervals as well as cells and raster regions. For each granular term, generalization rules are defined so that a phenomenon can be expressed from one LoD to a coarser one in an automatic way. Changing a phenomenon's LoD can simplify granular terms, transforming for instance a time interval into a time instant or a raster region into a cell. Our contributions are shown based on a real dataset about tornadoes in the USA. The results obtained show an enhancement of application scenarios from the extended granularities-based model to its ability of providing different phenomenon's representations in each LoD, while keeping its original strengths.

Keywords Granule · Granularity · Granular computing

R.A. Silva (✉) · J.M. Pires
NOVA LINCS, DI, FCT, Universidade NOVA de Lisboa, Lisbon, Portugal
e-mail: ricardofcsasilva@gmail.com

J.M. Pires
e-mail: jmp@di.fct.unl.pt

M.Y. Santos
ALGORITMI Research Centre, University of Minho, Braga, Portugal
e-mail: maribel.santos@algoritmi.uminho.pt

A. Bregt et al. (eds.), *Societal Geo-innovation*, Lecture Notes
in Geoinformation and Cartography, DOI 10.1007/978-3-319-56759-4_16

1 Context and Motivation

Many phenomena like crimes, storms, fires are being logged as a collection of spatiotemporal events at high levels of detail (LoDs). A spatiotemporal event can be modeled as *event(space, time*, $A_1, \ldots, A_N$), where *space* describes the spatial location of the event, *time* specifies the time moment, and $A_1, \ldots, A_N$ are attributes detailing what has happened. For example, an event involving victims can be described as *event*(*space, time, victims*). This way, *crime* (coords, 09/05/2015 20:45, 1) stands for a crime that occurred at some earth coordinates at 20:45 leading to one victim.

Spatiotemporal events can be expressed at different LoDs. Particularly, *space* can be described using cells with different sizes, cities, counties or states; *time* can be specified with a detail of minutes, hours, months or years.

The LoD plays a crucial role during the analytical process. From one LoD to another some patterns can become easily perceived or different patterns may be detected (Keim et al. 2008; Andrienko et al. 2010). For this reason, approaches allowing users to study and explore phenomena across multiple LoDs have been developed (Sips et al. 2012; Goodwin et al. 2016; Silva et al. 2016).

Granular computing approaches have been used to model phenomena at several LoDs based on granularities (Yao et al. 2013). Roughly, a granularity defines a division of a domain in a set of granules disjoint from each other (Keet 2008; Pires et al. 2014). Granules might represent familiar concepts for us (as humans) or not, i.e., are just a portion of the granulated domain. For example, *Counties*, *States* are common examples of spatial granularities; $Raster\left(2\,\mathrm{km}^2\right)$ represents a granularity where granules have equal square sized extents of 2 km^2, i.e., raster granularity; and *Minutes*, *Hours*, *Days* are examples of temporal granularities.

Under a general theory of granularities (Pires et al. 2014; Silva et al. 2015a), a granular computing approach was devised to model spatiotemporal phenomena at multiple LoDs labeled as the granularities-based model (Silva et al. 2015b) where a phenomenon is modeled through a collection of statements. Granules are used in the statements' arguments. For example, we can model a crime event through the statement: *crime*(Oakland, 03/01/2015 18 h, 1) where the granules used come from the following granularities {(space, *Counties*), (time, *Hours*), (victims, *Natural Numbers*)}. Notice that, the crime event did not occur in the entire extent of Oakland and did not occur from 18 h until 18:59. Instead, the crime occurred in some part of the Oakland county at some point between 18 h and 18:59. Therefore, we follow a *weak* interpretation of granules (Bravo and Rodríguez 2014).

Following the proposed approach in (Silva et al. 2015b), statements can be generalized to coarser LoDs automatically based on the relationship *coarsening* that occurs between granules: a granule g_1 is a coarsening of another granule g_2 if the extent of g_1 contains the extent of g_2. For example, the previous crime event can be generalized to *crime* (California, 03/01/2015, 1) where the granules used come from the following granularities {(*space*, *States*), (*time*, *Days*), (*victims*, *Natural Numbers*)}. Notice that, the granule California is coarsening of the granule Oakland, and the same property is verified in the remaining arguments.

By definition, a statement must have assigned one and only one granule to each argument (Silva et al. 2015b). This constraint is reducing the applicability of the granularities-based model. Let's consider that we aim to model the following tornado event: "*a tornado occurred on May 9th, 2015 between 15:45 and 16:10 pm. It affected a particular area, resulting 20 victims*". Following the formalization of the granularities-based model we are unable to model the tornado event like for example *tornado(RasterRegion({cells}), Interval(09/05/2015 15:45, 09/05/2015 16:10), 20)* where the granules would come from {(*space*, *Grid* $(2\,\mathrm{km}^2)$), (*time*, *Minutes*), (*victims*, *Natural Numbers*)}. In other words, it is not possible to assign complex descriptions to statements' arguments. This issue emerges because the granules are being used only as the domain of discourse instead of being used also to build the domain of discourse. For example, we can use a granule from the granularity *Counties* to mention a particular county in which an event occurred. However, in several scenarios it is desirable to use granules to express particular concepts which are not captured by the granules themselves. For example, based on the granules of *Grid* $\left(2\,\mathrm{km}^2\right)$ we would like to describe the region where the mentioned tornado moved as *RasterRegion({cells})*, and also its duration as *Interval (09/05/2015 15:45, 09/05/2015 16:10)* using the granules from *Minutes*.

To meet this need, our contributions are the following. Firstly, we propose the *granular term* concept. Different types of granular terms can be defined. This paper defines temporal (*Instant* and *Interval*) and spatial types (*Cell*, *RasterRegion*) of granular terms. Then, we extend the granularities-based model in order to handle granular terms. By introducing the granular term concept, the generalization is revisited. We propose to define the generalization process for each type of granular term in order to take a granular term from one LoD to a coarser one. This enhances the ability of generalizing a phenomenon from one LoD to a coarser one. Instead of having the same generalization process independently of the statement's argument (e.g., space, time) as originally, specific rules can now be defined for the generalization of a particular type of granular term.

Extending the granularities-based model with the proposed granular terms represents a very important enhancement in the applicability scenarios, while enhancing its original capabilities to represent phenomena at different LoD as shown in the demonstration case presented later in this paper.

The remaining paper is organized as follows. Section 2 introduces the granularities-based model and its limitations. In Sect. 3, our contributions are presented namely the extension of the granularities-based model. Section 4 presents and discusses the related work and their limitations. Section 5 presents a demonstration case based on a real dataset about tornadoes in the USA. Section 6 concludes with some remarks about the work undertaken and guidelines for future work.

2 Background

Pires et al. (2014) proposed a theory of granularities. This theory is founded on the general concept of granularity applicable to any domain where granules are disjoint from each other. Each granule g_α is composed by its extent $E(g_\alpha)$, i.e., a subset of the granulated domain and its index value $Ind(g_\alpha)$.

A fundamental relation between granularities is the relationship *finer than*, i.e., a granularity G is finer than H ($G \preccurlyeq H$) if and only if each extent of a granule of G is contained in one and only one extent of a granule of H (Bettini et al. 2000; Keet 2008; Silva et al. 2015a). For example, *Counties* is finer than *States* (*Counties* $\preccurlyeq$ *States*) and *Hours* is finer than *Days* (*Hours* $\preccurlyeq$ *Days*). As such, the granules can be arranged in a hierarchical alike structure, which allows the same phenomenon be perceived at different LoDs.

Furthermore, the authors provide a formal approach to bring relations defined in the original domain to the granules. For example, let g_a and g_b be two granules of a granularity defined over the time domain; let *before* be a relation such that an element t_i is before another element t_j if and only if $t_i < t_j$ ($t_i, t_j \in \mathbb{R}$). Following the (Pires et al. 2014) approach, granules can be related g_a and g_b using the following induced relationships formally defined in (Pires et al. 2014): (*i*) g_a is completely before of g_b (see Fig. 1a) denoted as $g_a before^C g_b$; (*ii*) g_a is partially before of g_b (see Fig. 1b); (*iii*) g_a weakly before of g_b (see Fig. 1c); g_a existentially before of g_b(see Fig. 1d).

Under the theory of granularities proposed in (Pires et al. 2014; Silva et al. 2015a, b), a granularities-based model was devised to model spatiotemporal phenomena at different LoDs. This model is composed by a collection of statements that are made using granules. These come from a set of granularities that are defined a priori. As introduced in Sect. 1, we found two major issues, which are detailed here. Both issues emerge because one and only one granule must be assigned to each statement's argument.

Recall the example of tornado event that we would like to model as *tornado (RasterRegion({cells}), Interval(09/05/2015 15:45, 09/05/2015 16:10), 20)*. Often, the temporal granularities available to describe time are *Minutes*, *Hours*, *Days*, *Weeks* and so forth. For example, if we intend to model the mentioned tornado event using the granularity *Minutes*, an issue emerges because we can only assign a single granule to each statement's argument. This way, we would need to define a

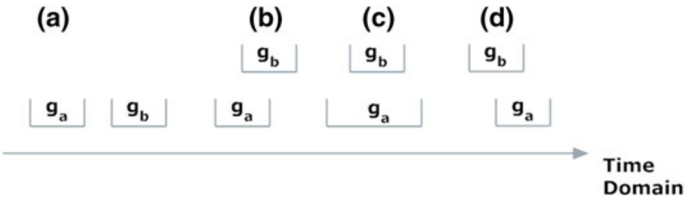


Fig. 1 Illustration of the induced relations (Silva et al. 2015a)

temporal granularity containing one granule with the extent: May 9th, 2015 between 15:45 and 16:10 pm.

Let's consider another tornado event that occurred at May 9th, 2015 between 15:55 and 16:05. Again, the granule needed to describe when the tornado occurred is not available in the granularities *Minutes*. This way, each time we need a particular granule that it does not exist in the available granularities, a new granularity needs to be defined. This issue becomes even worse if we look at the argument *space*, where is common the need to represent "arbitrary" spatial features like regions or trajectories. In the end, creating granularities as they are needed is not suitable because more granularities will be required as more and more statements are added to a granularities-based model.

Another major drawback is about the automated approach to generalize a phenomenon from one LoD to a coarser one. Once statements have one granule assigned to each argument, the generalization process defined in (Silva et al. 2015b) is the same for all the arguments of a statement. Recall that, *crime* (Oakland, 03/01/2015 18 h, 0) can be generalized to *crime* (California), 03/01/2015, 0) because $\underline{\text{California}}$ is a coarsening of $\underline{\text{Oakland}}$, $\underline{\text{03/01/2015}}$ is a coarsening of $\underline{\text{03/01/2015 18 h}}$ and so on. However, different arguments may need different generalization processes, particularly the time and space. For example, a time interval can eventually be generalized to a time instant while a region might be simplified.

In short, the granularities-based model limitations are: (*i*) one cannot assign complex descriptions to statements' arguments; (*ii*) and, independently of the statement's argument, the generalization process is the same. The limitations found are addressed in this work by extending the granularities-based model with granular terms.

3 Extending Granularities-Based Model with Granular Terms

Under the granularities-based model (Silva et al. 2015b), a phenomenon is modeled through a collection of statements. These are built based on a definition of a predicate. A predicate P contains a set of arguments $Args(P)$, and its signature specifies the set of granularities that can be used in each argument $\mathcal{G}_{(\mathrm{P,\,arg})}$.

The phenomenon about tornadoes in the USA[1] is an example of a phenomenon where the limitations of the granularities-based model become visible. To illustrate our proposals, let's introduce the *tornado* predicate in order to model the phenomenon about tornadoes in the USA. The *tornado* predicate contains three arguments (space, time, victims), and its signature defines that the granularities $Raster(0.5\ \mathrm{km}^2)$,

[1] Tornadoes logged as spatiotemporal events: http://www.ncdc.noaa.gov/stormevents/details.jsp.

Raster$\left(2\ \mathrm{km}^2\right)$, *Raster*$\left(8\ \mathrm{km}^2\right)$ *Raster*$\left(32\ \mathrm{km}^2\right)$ and *Raster*$\left(128\ \mathrm{km}^2\right)$ can be used in the argument *space* (denoted as $\mathcal{G}_{(\mathrm{tornado,\,space})}$); the granularities *Minutes*, *Hours*, *Weeks*, *Months* and *Years* can be used in the argument *time* (denoted as $\mathcal{G}_{(\mathrm{tornado,\,time})}$) and, the granularity *Natural Numbers*, defined over $\mathbb{N}$ where each granule corresponds to an element of the corresponding domain, can be used in the argument *victims* (denoted as $\mathcal{G}_{(\mathrm{tornado,\,victims})}$).

To allow the assignment of complex descriptions to statements' arguments rather than a single granule, we propose granular terms. Granular terms are built based on function symbols and granules from a single granularity. Let f be an n-ary or a variadic function symbol and G a granularity. A n-ary function symbol has a fixed arity while a variadic function symbol has an indefinite arity. A granular term is $f(g_1, \ldots, g_n)$ such that $g_i \in \mathrm{G}$ for all $1 \le i \le n$ or a granular term is $f(t_1, \ldots, t_n)$ such that t_i is a term defined using the granularity G for all $1 \le i \le n$. Granular terms in the form of $f(g_1, \ldots, g_n)$ are simple, and the ones in the form of $f(t_1, \ldots, t_m)$ are compound. Finally, granular terms can also be built using the identity function symbol $I(g \in G)$.

Interval(03/01/2015 18 h, 03/01/2015 20 h) is an example of a simple granular term using granules from *Hour*s; *MultiInterval*(*Interval*(03/01/2015 18 h, 03/01/2015 19 h), *Interval*(03/01/2015 21 h, 03/01/2015 22 h)) is an example of a compound granular term using granules from *Hours*; and, *I*(Oakland) is an example of a granular term built based on the identity function symbol and the granularity *Counties*.

This work formalizes the following function symbols in the subsequent sections: *Instant*, *Interval*, *Cell* and *RasterRegion*. These allow modeling time instants, time intervals, cells or raster regions, respectively, using granules from granularities. Notice that, the *Interval* function symbol needs to establish additional constraints in order to disallow improper granular terms of *Interval* (see Sect. 3.1) like the ones in the form of *Interval (Interval(a, b), Instant (c))*. Similarly, the function symbol *RasterRegion* also needs additional constraints in order to have well-formed raster regions (see Sect. 3.2). Therefore, a function symbol allows granular terms be defined by using a collection of granules that represent a particular concept. As such, each function symbol contains its own signature establishing the needed restrictions to build granular terms of type f. Other examples of function symbols can be pointed out but we left for future work their formalization. For example, it's common the need to represent spatial features in vector space like points, lines, polygons, and a set of polygons, among others examples. Also, in several applications scenarios, the concept of trajectory is crucial to model the trajectories made by people, cars, animals, among others. The needed function symbols depend on the phenomenon under study.

By introducing granular terms, the predicate concept (Silva et al. 2015b) is extended as follows. The predicate signature declares a set of granularities $\mathcal{G}_{(\mathrm{P,\,arg})}$ and the function symbols $\mathcal{F}_{(\mathrm{P,\,arg})}$ that can be used in each argument. This way, a well-formed atom (i.e., a statement) is now in the form of $\mathrm{P}(\tau)$ with $\tau = \{(\mathrm{arg},\ \mathrm{granular\ term}) \mid \mathrm{arg} \in \mathrm{Args(P)} \wedge$ granular term is defined by using

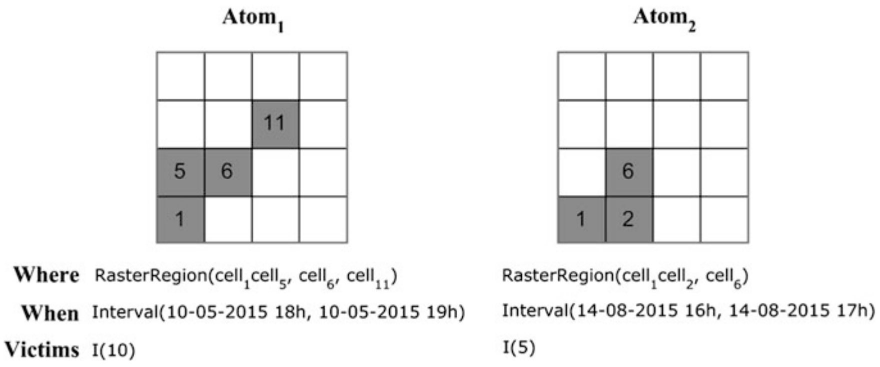

Atom$_1$
Atom$_2$
11
5
6
1
6
1
2
Where
RasterRegion(cell$_1$cell$_5$, cell$_6$, cell$_{11}$)
RasterRegion(cell$_1$cell$_2$, cell$_6$)
When
Interval(10-05-2015 18h, 10-05-2015 19h)
Interval(14-08-2015 16h, 14-08-2015 17h)
Victims
I(10)
I(5)

Fig. 2 Schematic representation of two atoms of the *tornado* predicate

granules of a single a valid granularity $G \in \mathcal{G}_{(\mathrm{P},\,\mathrm{arg})}$ and using a function symbol $f \in \mathcal{F}_{(\mathrm{P},\,\mathrm{arg})}\}$.

The signature of the *tornado* predicate defined previously is extended with: $\mathcal{F}_{(\mathrm{tornado,\ space})} = \{Cell,\ RasterRegion\}$, $\mathcal{F}_{(\mathrm{tornado,\ time})} = \{Instance,\ Interval\}$ and, $\mathcal{F}_{(\mathrm{tornado,\ victims})} = \{I\}$ (I is the identity function symbol). Figure 2 displays two atoms of the *tornado* predicate where each one is describing the occurrence of a tornado. This example shows the granular term concept being used to describe the regions where tornadoes occurred at granularity $Raster(0.5Km^2)$, the time interval during which they happened at granularity *Hours*, and the resulting victims specified at granularity *Natural Numbers*.

The granularities-based model follows an automated approach to generalize a phenomenon from one LoD to a coarser one (see Sect. 2). With the introduction of granular terms, the generalization process defined in (Silva et al. 2015b) is no longer applicable. There is a need to introduce new instruments in order to generalize atoms to coarser LoDs.

This work proposes that each function symbol must have associated generalization rules $\mathbb{G}_{\mathrm{f}}$. These are applied when an atom is generalized. The generalization can turn a time interval into a time instant, simplify a raster region, or even turn a raster region into a cell (i.e., generalization-reduction process). The formalization of the function symbols *Instant*, *Interval*, *Cell* and *RasterRegion* and their generalization rules are presented in the next sections.

3.1 Temporal Granular Terms

In order to represent time, we introduce temporal granular terms, which are built using temporal granularities.

Concerning the time domain, time instants have no duration. In contrast, a time interval is the set of all time instants between a starting point and an ending point.

Let T be a temporal granularity. To represent a time instant of T, we introduce the *Instant* function symbol defined as follows: $\mathit{Instant}(t)$ where $t \in T$.

Two granules of T can be related through the induced complete relationship $<^{\mathrm{C}}$ (see Sect. 2) in order to tell whether a granule of T occurs before another one. In order to represent time intervals of T, we introduce the *Interval* function symbol.

Definition 1 (*Time Interval*) Let *Interval* be a function symbol and its arity is equal to two; let t^- and t^+ be granules of T such that $t^- <^C t^+$ (also mentioned as the endpoints of the interval); a time interval of $T\, \{\mathrm{t_i} \in T \mid t^- <^C t_i <^C t^+\}$ is denoted by $\mathit{Interval}(t^-, t^+)$.

Granular terms of *Instant* or *Interval* should be interpreted in the context of the temporal granularity used to build them. For instance, a granule from a granularity *Hours* represents an hour of time and it should not be considered a time interval in the context of this work but rather an indivisible moment of time. Recall that, a granule is a non-decomposable entity. Therefore, granules from a temporal granularity T are interpreted as time instants.

Based on the temporal granular terms presented, we can build atoms describing that something occurred in some time instant or time interval of T. A well-formed atom describing a tornado event is for example:$\mathrm{o_1} = \mathrm{tornado}\left(\mathit{Cell}\left(\mathrm{cell}_1^{0.5\,\mathrm{km}^2}\right),\right.$ $\left.\mathit{Interval}(10/5/2014\; 16{:}40, 10/5/2014\; 6{:}45), I(2)\right)$. In this example, the interval of time is a granular term defined at granularity *Minutes*.

Allen (1983) and Vilain (1982) and point algebras model topological relations between time intervals, time intervals and time instants (or vice versa), and time instants, respectively, which are defined over the time domain. As we can bring the relations of the domain into the granularities (see Sect. 2), we transpose the topological relations for temporal granular terms.

Let $a = \mathit{Instant}(\alpha), b = \mathit{Instant}(\beta)$ be simple granular terms of T. a can occur before b ($\alpha <^C \beta$), both time instants can be equal ($\alpha = \beta$), or a can occur after b ($\alpha >^C \beta$). On the other hand, let $c = \mathit{Interval}(\alpha^-, \alpha^+)$ and $d = \mathit{Interval}(\beta^-, \beta^+)$ be simple granular terms of T. c and d can be related as follows (the symmetric relations are not displayed):

- c before d iff $\alpha^+ <^{\mathrm{C}} \beta^-$
- c equals d iff $(\alpha^- = \beta^-) \wedge (\alpha^+ = \beta^+)$
- c overlaps d iff $(\alpha^+ <^{\mathrm{C}} \beta^-) \wedge (\alpha^+ >^{\mathrm{C}} \beta^-) \wedge (\alpha^+ <^{\mathrm{C}} \beta^+)$
- c meets d iff $\alpha^+ = \beta^-$
- c starts d iff $\alpha^- = \beta^- \wedge \alpha^+ <^{\mathrm{C}} \beta^+$
- c during d iff $\alpha^- >^{\mathrm{C}} \beta^- \wedge \alpha^+ <^{\mathrm{C}} \beta^+$
- c finishes d iff $\alpha^+ = \beta^+ \wedge \alpha^- >^{\mathrm{C}} \beta^-$

Last but not least, let $e = \mathit{Instant}(\alpha)$ and $f = \mathit{Interval}(\beta^-, \beta^+)$ be a granular terms of T. e and ff can be related as follows (the symmetric relations are not displayed):

- e before f iff $\alpha <^{C} \beta^{-}$
- e starts f iff $\alpha = \beta^{-}$
- e during f iff $\beta^{-} <^{C} \alpha <^{C} \beta^{+}$
- e finishes f iff $\alpha = \beta^{+}$
- e after f iff $\beta^{+} <^{C} \alpha$

Generalization rules are defined for each function symbol so that the generalization of atoms can be performed automatically. We define generalization rules applicable to temporal granular terms. When a temporal granular term is generalized, an instant or an interval of time can remain an instant or an interval, correspondingly, but with less precision; or, a time interval can become a time instant. The generalization of temporal terms is formalized as follows.

Let T_1 and T_2 be temporal granularities such that $T_1 \preccurlyeq T_2$. An instant of time $a_1 = \mathit{Instant}(\alpha)$ of T_1 can be generalized into an instant of time $a_2 = \mathit{Instant}(\alpha')$ of T_2 through $\mathbb{G}_{\text{Instant}}\colon (a_1, T_1) \rightarrow (a_2, T_2)$ if and only if $\exists!\ \alpha' \in T_2\colon \mathrm{E}(\alpha) \subseteq \mathrm{E}(\alpha')$, that is, if there is exactly one granule α' belonging to T_2 such that the extent of α is contained by the extent of α'. For example, the $\mathit{Instant}(10-5-2014\ 16{:}40)$ at granularity *Minutes* is generalized into the $\mathit{Instant}(10-5-2014\ 16\,\mathrm{h})$ at granularity *Hours*.

An interval of time $a_1 = \mathit{Interval}(\alpha^{-}, \alpha^{+})$ of T_1 can be generalized into an interval of time $a_2 = \mathit{Interval}(\alpha'^{-}, \alpha'^{+})$ of T_2 through $\mathbb{G}_{\text{Interval}}\colon (a_1, T_1) \rightarrow (a_2, T_2)$ if and only if $\exists!\ \alpha'^{-} \in T_2\colon \mathrm{E}(\alpha^{-}) \subseteq \mathrm{E}(\alpha'^{-})$, and $\exists!\ \alpha'^{+} \in T_2\colon \mathrm{E}(\alpha^{+}) \subseteq \mathrm{E}(\alpha'^{+})$. That is, if there is exactly one granule α'^{-} belonging to T_2 such that the extent of α^{-} is contained by the extent of α'^{-} and, if there is exactly one granule α'^{+} belonging to T_2 such that the extent of α^{+} is contained by the extent of α'^{+}. Moreover, an interval of time $a_1 = \mathit{Interval}(\alpha^{-}, \alpha^{+})$ of T_1 can be generalized into an instant of time $a_2 = \mathit{Interval}(\alpha')$ of T_2 through $\mathbb{G}_{\text{Interval}}\colon (a_1, T_1) \rightarrow (a_2, T_2)$ if and only if $\exists!\ \alpha' \in T_2\colon \mathrm{E}(\alpha^{-}) \subseteq \mathrm{E}(\alpha') \wedge \mathrm{E}(\alpha^{+}) \subseteq \mathrm{E}(\alpha')$. That is, if there is exactly one granule α' belonging to T_2 such that the extent of α^{-} and α^{+} is contained by the extent of α'. For example, the $\mathit{Interval}(10-5-2014\ 16{:}\ 40,\ 10-5-2014\ 17{:}\ 45)$ at granularity *Minutes* is generalized into $\mathit{Interval}(10-5-2014\ 16\,\mathrm{h},\ 10-5-2014\ 17\,\mathrm{h})$ $10-5-2014\ 17\,\mathrm{h})$ at granularity *Hours*, or into $\mathit{Instant}(10-5-2014))$ at granularity *Days*.

The generalization of temporal granular terms may affect the temporal topological relationships held between pairs of atoms. On one hand, the type of relationship may change. For instance, we might have a relation between two time intervals that may turn into a relation between a time interval and a time instant. On the other hand, there are scenarios where the type of topological is kept but the actual relation (e.g., before) is changed (e.g., to equal).

A detailed study about the possible changes in all scenarios, as well as in what conditions they occur is provided.[2]

[2] http://staresearch.net/resources/agile2017/temporal_terms_appendix.pdf.

3.2 Spatial Granular Terms

In order to represent spatial features in raster space, we introduce simple spatial granular terms. These are built based on granularities defined over two-dimensional space where granules have equal square sized extents, i.e., raster granularities.

For the contexts of raster data, points are represented as cells, and regions are groups of contiguous cells that portray the shape of an area. Using granules from raster granularities, one may want to use granular terms to describe cells or raster regions. There are different definitions of raster regions (Kong and Rosenfeld 1989; Egenhofer and Sharma 1993). These definitions rely on the neighborhood concept. The 4-neighbors of a cell consist in the cells that share the vertical and horizontal sides, and the 8-neighbors are the ones sharing diagonal sides in addition to the 4-neighbors.

A region (without holes) is, in general, defined by a Jordan curve which divides a raster space in two parts. However, if we consider 4-adjacency or 8-adjacency, a paradox emerges in some curves (Kong and Rosenfeld 1989). One approach to overcome this problem is to consider different adjacency rules regarding a region and its complement (Kong and Rosenfeld 1989). In this work, we do not aim to propose a new raster region definition and we will adopt the mixed adjacency model (8, 4) to define a raster region, i.e., a raster region is 8-connected and its complement is 4-connected (Kong and Rosenfeld 1989).

Let S be a raster granularity. To represent a cell of S, we introduce the *Cell* function symbol defined as follows: $Cell(\mathrm{c})$ where $\mathrm{c} \in S$. In order to represent a raster region of S, we introduce the *RasterRegion* function symbol as follows.

Definition 2 (*Raster Region*) Let *RasterRegion* be a variadic function symbol; let $\mathrm{c}_1, \ldots, c_n$ be granules of a granularity S, i.e., $c_i \in S$ for all $1 \leq i \leq n$; a raster region of S is denoted by $RasterRegion(\mathrm{c}_1, \ldots, c_n)$ where $\{\mathrm{c}_1, \ldots, c_n\}$ is a set and their elements are 8-connected, the $\mathrm{S}\backslash\{\mathrm{c}_1, \ldots, c_n\}$ is 4-connected, and $n > 1$.

Based on the spatial granular terms presented, we can define atoms describing that something occurred in a particular cell or region of S. For example: $o_2 = tornado\left(RasterRegion\left(\mathrm{cell}_1^{0.5\,\mathrm{km}^2}, \mathrm{cell}_2^{0.5\,\mathrm{km}^2}, \mathrm{cell}_3^{0.5\,\mathrm{km}^2}\right), Interval(7-10-2014\ 15{:}25, 7-10-2014\ 15{:}33), I(2)\right)$.

Let S_1 and S_2 be raster granularities such that $S_1 \preccurlyeq S_2$. When a spatial granular term is generalized, a cell or a raster region can remain cell or a raster region, correspondingly, but with less precision; or, a raster region can become a cell. The generalization of spatial granular terms is formalized as follows.

A granular term $a_1 = Cell(\mathrm{c})$ of S_1 can be generalized to a granular term $a_2 = Cell(\mathrm{c}')$ of S_2 through $\mathbb{G}_{\mathrm{Cell}}: (a_1, S_1) \rightarrow (a_2, S_2)$ if and only if $\exists!\ \mathrm{c}' \in S_2: \mathrm{E}(\mathrm{c}) \subseteq \mathrm{E}(\mathrm{c}')$, that is, if there is exactly one granule c' belonging to S_2 such that the extent of c is contained by the extent of c'.

Let $a_1 = RasterRegion(\mathrm{c}_1, \ldots, c_n)$ be a granular term of S_1. It can be generalized to a granular term $a_2 = RasterRegion(c_1', \ldots, c_m')$ of S_2 $(\mathrm{m} \leq \mathrm{n})$ through $\mathbb{G}_{\mathrm{RasterRegion}}: (a_1, S_1) \rightarrow (a_2, S_2)$ if and only if $\forall \mathrm{i} \in \{1 \ldots, \mathrm{n}\}\ \mathrm{c}_i \in S_1\ \exists! \mathrm{j} \in$

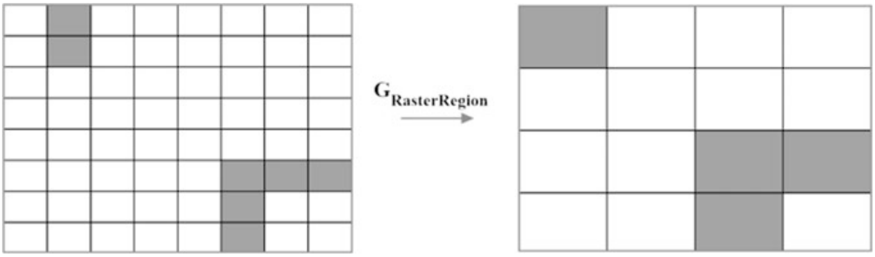


Fig. 3 Illustration of the generalization rules associated to $\mathbb{G}_{\mathrm{RasterRegion}}$

$\{1\ldots,\mathrm{m}\}c_j^{'} \in S_2$: $\mathrm{E}(\mathrm{c}_i) \subseteq \mathrm{E}(c_j^{'})$, that is, if for any granule c_i defining the raster region a_1 there is exactly one granule $c_j^{'}$ belonging to S_2 such that the extent of c_i is contained by the extent of $c_j^{'}$. Moreover, the granular term $a_1 = RasterRegion$ $(\mathrm{c}_1, \ldots, c_n)$ of S_1 can be generalized to a granular term $a_2 = Cell(c^{'})$ of S_2 through $\mathbb{G}_{\mathrm{RasterRegion}}$: $(a_1, S_1) \rightarrow (a_2, S_2)$ if and only if $\exists!\ c^{'} \in S_2 \forall \mathrm{i} \in \{1\ldots,\mathrm{n}\}\ \mathrm{c}_i \in S_1$: $\mathrm{E}(\mathrm{c}_i) \subseteq \mathrm{E}(\mathrm{c}^{'})$, that is, there is exactly one granule $c^{'}$ belonging to S_2 such that any granule c_i defining the raster region a_1 has its extent contained by the extent of $c^{'}$.

To illustrate the generalization rules associated to the function symbol *RasterRegion*, $\mathbb{G}_{\mathrm{RasterRegion}}$, Fig. 3 shows two raster regions being generalized to a coarser granularity. The left one changes from a *RasterRegion* to *Cell* while the right one remains a *RasterRegion* with less precision. A study about the generalization of spatial granular terms and its impact on the topological relations (e.g., disjoint, meets, contains) between them was left for future work.

3.3 Granularities-Based Model

Throughout this paper, we have extended the predicate concept as well as how atoms can be produced in order to handle granular terms. For each function symbol introduced, generalization rules were also defined.

In order to have spatiotemporal phenomena at multiple LoDs, the generalization concept takes an atom in one LoD and expresses it at a coarser one. In our previous work (Silva et al. 2015b) the granularities-based model did not handle granular terms. As a result, the generalization of atoms was based on the relationship *coarsening* that occurs between granules: a granule g_1 is a coarsening of another granule g_2 if the extent of g_1 contains the extent of g_2, i.e., $\mathrm{E}(\mathrm{g}_2) \subseteq \mathrm{E}(\mathrm{g}_1)$. This generalization process was the same for any argument of a predicate *P*. As opposed, we have introduced generalization rules for each function symbol, allowing different generalization processes for each predicate's argument.

The generalization of atoms occurs from one LoD to a coarser one. One LoD of the predicate *P* consists of a set of argument pairs and a valid granularity (Silva et al. 2015b). Two examples of LoDs of the *tornado* predicate are the $\mathrm{LoD_x} = \{(\text{space}, \; \mathit{Raster}(0.5\,\mathrm{km}^2)), \; (\text{time}, \; \mathit{Minutes}), \; (\text{victims}, \; \mathit{Natural\,Numbers})\}$ and $\mathrm{LoD_y} \; = \{(\text{space}, \mathit{Raster}(2\,\mathrm{km}^2)), (\text{time}, \mathit{Hours}), (\text{victims}, \mathit{Natural\,Numbers})\}$.

The set of all valid LoDs of a predicate *P* is denoted by $\mathcal{L}^{\mathrm{P}}$. The generalization of atoms occurs from a LoD_i *more detailed than* LoD_j. For example, the $\mathrm{LoD_x}$ *is more detailed than* $\mathrm{LoD_y}$ because $\mathit{Raster}(0.5\,\mathrm{km}^2) \preccurlyeq \mathit{Raster}(2\,\mathrm{km}^2)$, $\mathit{Minutes} \preccurlyeq \mathit{Hours}$ and $\mathit{Natural\,Numbers} \preccurlyeq \mathit{Natural\,Numbers}$. An atom is always described in some LoD belonging to $\mathcal{L}^{\mathrm{P}}$. This remains true after the introduction of granular terms, once a granular term is defined based on a single granularity. This way, the generalization of atoms corresponds to applying the generalization rules to each granular term specified in each argument of an atom.

4 Related Work and Discussion

There are several works in the literature for modeling spatiotemporal phenomena at multiple LoDs under different terminologies like multirepresentation, multiresolution and granular computing.

Similar to multirepresentation approaches (Parent et al. 2009), each predicate provides a representation of the phenomenon, but unlike them there is no need to define everything at the instances level. Furthermore, multirepresentation approaches do not have pre-defined operations that take data from one spatial and/or temporal LoD to another (Parent et al. 2009).

Multiresolution approaches (Weibel and Dutton 1999) focus essentially in the generalization of spatial features (Stell and Worboys 1998). The generalization of spatial data may involve object simplification (e.g., at less precise resolutions, a building may be defined using less vertices than originally); a change in the object geometry (e.g., a building can be represented by a polygon at a precise resolution, and by a point at a less precise resolution); and existence (e.g., eventually displaying that building is no longer relevant) (Weibel and Dutton 1999; Brahim et al. 2015; Zhou et al. 2004). This sequence of operations were coined as Generalization-Reduction-Disappearance process (Laurini 2014). Unlike the multiresolution approaches, the granularities-based model can express a phenomenon in several LoDs (which is a concept formally defined), and not just in several spatial LoDs. Using the granularities-based model, atoms can be expressed into other coarser LoDs in an automatic way, as the user just needs to define the generalization rules applicable to each function symbol. In this process, less detailed phenomenon representations can be achieved as the generalization-reduction process can be applicable to any argument of a predicate *P*.

Granular computing approaches make use of granules in complex problem solving (Yao et al. 2013). Keet (2008) developed a formal, domain-independent theory of granularity that can be used for computational reasoning. This theory was developed to model phenomena at different LoDs and it was applied to biological sciences. Based on the criterion of granulation, Keet (2008) proposed a domain-independent taxonomy of types of granularity. This taxonomy makes explicit both the ways of granulation, and how entities are organized within a granular level. However, Keet's (2008) theory of granularities does not fully support dealing with the complexity of temporal granularities (Keet 2008) which is crucial for modeling spatiotemporal phenomena.

Camossi et al. (2006) propose a granular computing approach to represent spatiotemporal information (vector approach) in object-oriented database management systems (DBMSs) extending the ODMG standard. Bravo and Rodriguéz (2014) propose a multi-granular database model and a query language in order to query data using different granularities. Camossi et al. (2006) are indexing information at different spatial or/and temporal granularities, and Bravo and Rodriguéz (2014) are aggregating and querying data at different granularities.

As opposed to the current granular computing approaches, which are mainly concerned about indexing and aggregating data at different granularities, the granularities-based model provides different phenomena's representations for each LoD. Once the atoms at the lowest LoD of the predicates are produced, the phenomenon can be expressed into other coarser LoDs in an automatic way based on the generalization rules for each function symbol.

To the best of our knowledge, there is no other model devised to model spatiotemporal phenomena, containing together such characteristics.

5 Demonstration Case

The demonstration case is about tornadoes occurred in the USA between 1990 and 2015. This phenomenon is described by a collection of 32 570 geo-referenced spatiotemporal events. The F1 tornadoes were excluded since their impact in terms of victims is not significant and their spatial coordinates were not accurate in general. So we kept 27 182 geo-referenced spatiotemporal events representing tornadoes with categories ranging from F2 to F5 (in Fujita scale). These events were modeled through a tornado predicate, with three arguments (space, time, victims). The most detailed spatial granularity is based on a grid of 32768 × 32768 cells that cover the analyzed spatial extent of the phenomenon, and each cell has an area of 0.13 km^2. The other coarser spatial granularities were obtained by dividing by a factor of 2 the number of cells in the grid. So the used granularities for space were rasters with cell sizes of 0.13 km^2, 0.5 km^2, 2 km^2, 8 km^2, 32 km^2. The used time granularities were *Minute*, *Hour*, *Day*, *Week*, *Month*.

The considered granular terms required to model these events were: *Instant* and *Interval* for the time argument; *Cell* and *RasterRegion* for space argument.

Table 1 Percentage of atoms using the proposed granular terms

	Instant (%)	Interval (%)	Total (%)
Cell	27	16	43
Raster region	4	54	57
Total	30	70	

The raw data (tornadoes) were encoded at the lowest LoD using the time granularity of *Minute* and the space granularity 0.13 km^2 The temporal granular terms *Instant* and *Interval*, and the spatial granular terms *Cell* and *RasterRegion* were used with the tornado predicate according to the data.

As shown in Table 1, at this LoD (0.13 km^2 and Minute), some tornadoes were described using:

- the *Cell* and *Instant* granular terms (27%), which are those with a very short duration and very little spatial expression;
- the *Cell* and *Interval* granular terms (16%), which are those with a very little spatial expression but with a time duration larger than a single minute; the average size of the intervals is 8 min and 22 s;
- the *Raster Region* and *Instant* granular terms (4%), which are those few that have a duration not larger than a minute but with a spatial expression that requires more than one *Cell*; the average number of cells for the raster regions is 70.6;
- the *Raster Region* and *Interval* granular terms (54%), which are those few that have a duration larger than a minute and with a spatial expression that requires more than one *Cell*; the average number of cells for the *Raster Region* is 39.6 and the average number of minutes for the *Interval* is 7 min and 12 s.

Notice that, the most (70%) of tornadoes require a granular term *Interval*. Also, most (57%) tornadoes require a granular term *Raster Regions*. The description of those tornadoes would be impossible, or at least very hardly to be encoded without the concept of granular terms and in particular the *Intervals* and *Raster Regions*.

The generalization rules presented in Sect. 3.1 and 3.2 were adopted, enabling the automatic computation of the model at coarser LoDs. Given all tornadoes encoded at the finest LoD (0.13 km^2 and *Minute*) using the *tornado* predicate with the appropriate granular terms, the model has been computed at coarser LoDs, at all combinations of space and time granularities.

To illustrate how the granularity affects our perception about temporal topological relationships between atoms, let's consider the three F4 tornadoes that occurred in western Iowa at May 27 of 1995. The first one (identified as A) started at 18:22 and ended at 19:47. The second one (identified as B) started at 18:55 until 20:24. The third one (identified as C) started at 18:56 until 20:08. At granularity *Minutes*, A overlaps B and A overlaps C. However, when the tornadoes are observed at granularity *Hours*, our perception is changed and therefore A starts B and A starts C.

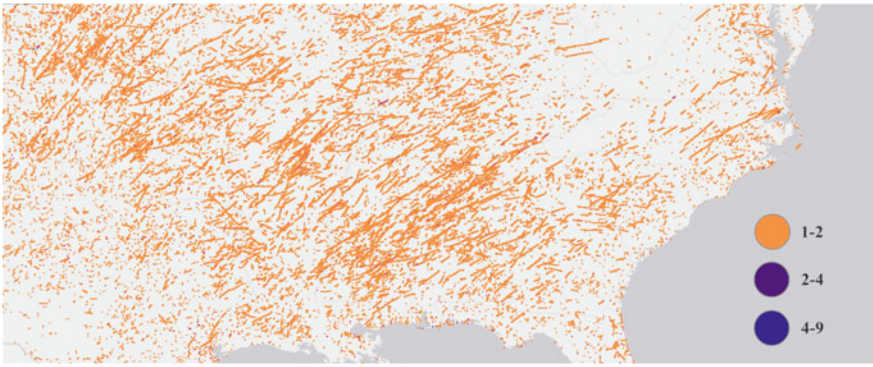


Fig. 4 An overview of all atoms in a LoD containing the granularity Raster(0.5 km^2)

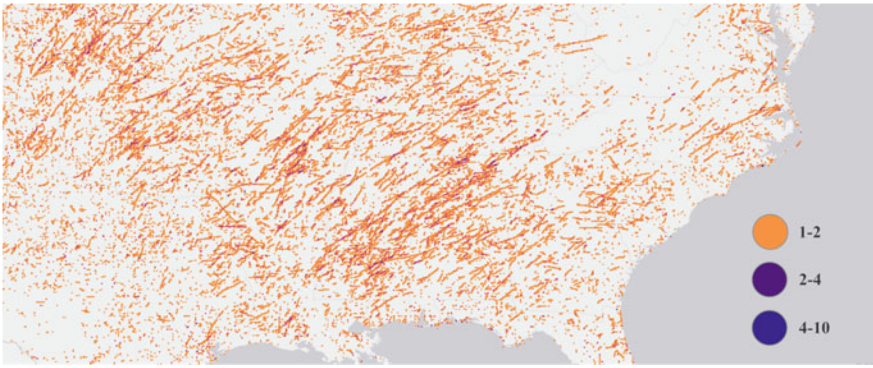


Fig. 5 An overview of all atoms in a LoD containing the granularity Raster(8 km^2)

To study the co-occurrence of tornadoes in space, we compute the total atoms that exist in each spatial granule considering LoDs where atoms (the space argument) are defined using the granularities $Raster(0.5\,\text{km}^2)$, $Raster(8\,\text{km}^2)$ and $Raster(32\,\text{km}^2)$.

For each scenario, we display on a map (Figs. 4, 5 and 6) the spatial granules colored based on the number of atoms in it. Oranges show low values while pinks and dark blues show high values. Looking at Fig. 6, the spatial co-occurrence of tornadoes becomes clear in LoDs where the spatial granular terms are built using granules from Raster(32 km^2) (see Fig. 6). This supports the claim of some authors (Andrienko et al. 2010; Goodwin et al. 2016; Silva et al. 2016) that the LoD plays a crucial role during the analytical process.

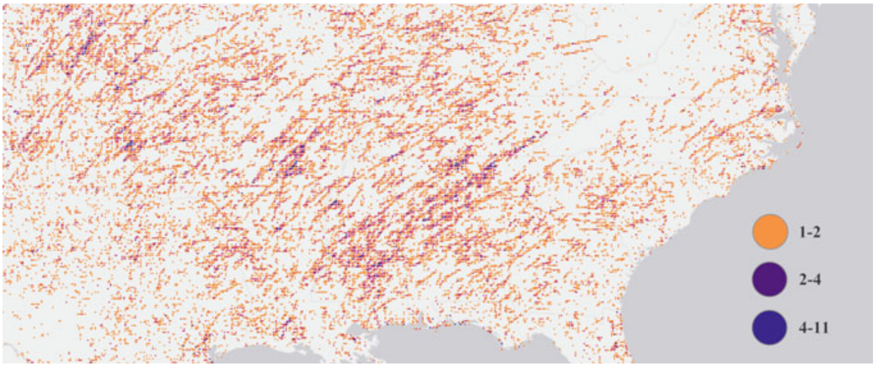


Fig. 6 An overview of all atoms in a LoD containing the granularity Raster(32 km^2)

6 Conclusions and Future Work

A theory of granularities has been proposed for modeling spatiotemporal phenomena at multiple LoDs (Pires et al. 2014; Silva et al. 2015a, b). This theory relies on a granularities-based model that models phenomena through atoms. Although there are advantages, we found two major drawbacks regarding the granularities-based model: (*i*) one cannot assign complex descriptions to the atoms' arguments; and (*ii*) independently of the atom's argument, the generalization process is the same. To address both issues, this paper introduces the concept of granular term. This way, one can assign granular terms to the atoms' arguments, and the generalization process depends on the function symbol used to build a granular term.

Based on the general concept of granular term, spatial granular terms (*Cell* and *RasterRegion*) and temporal granular terms (*Instant* and *Interval*) were formalized. Regarding the latters, we transpose the temporal topological relations to the temporal granular terms. A theoretical analysis was made for reasoning about what happens to the topological relations (in the context of temporal granular terms) when these are generalized. Our proposals were implemented and a demonstration case was conducted with a real dataset about tornadoes in the USA. The results obtained show an enhancement of application scenarios, from the extended granularities-based model, as well as its ability of providing different phenomenon's representations in each LoD while keeping its original strengths.

Future work can be directed to the definition of topological relations between the spatial granular terms introduced in Sect. 3.2 as well as a detailed study about what happens to topological relations between spatial granular terms when these are generalized. Also, new function symbols and their generalization rules can be defined. For example, function symbols to model spatial features in vector space like points, lines, polygons. Furthermore, function symbols to model trajectories can be crucial for several applications scenarios.

Acknowledgements This work has been supported by FCT—Fundação para a Ciência e Tecnologia MCTES, UID/CEC/04516/2013 (NOVA LINCS) and UID/CEC/00319/2013 (ALGORITMI), and COMPETE: POCI-01-0145-FEDER-007043 (ALGORITMI).

References

Allen JF (1983) Maintaining knowledge about temporal intervals. Commun ACM 26(11):832–843

Andrienko G et al (2010) Space, time and visual analytics. Int J Geogr Inf Sci 24(10):1577–1600

Bettini C, Jajodia S, Wang S (2000) Time granularities in databases, data mining, and temporal reasoning. Springer

Brahim L, Okba K, Robert L (2015) Mathematical framework for topological relationships between ribbons and regions. J Vis Lang Comput 26:66–81

Bravo L, Rodríguez MA (2014) A multi-granular database model. In Foundations of information and knowledge systems. Springer, pp 344–360

Camossi E, Bertolotto M, Bertino E (2006) A multigranular object-oriented framework supporting spatio-temporal granularity conversions. Int J Geogr Inf Sci 20(5):511–534

Egenhofer MJ, Sharma J, (1993) Topological relations between regions in ρ2 and ℤ2. In: Advances in spatial databases, pp 316–336

Goodwin S et al (2016) Visualizing multiple variables across scale and geography. IEEE Trans Vis Comput Graph 22(1):599–608

Keet CM (2008) A formal theory of granularity. Free University of Bozen-Bolzano

Keim D et al (2008) Visual analytics: definition, process, and challenges. In Kerren A et al (ed) Information visualization. Lecture notes in computer science. Springer, Berlin, pp 154–175

Kong TY, Rosenfeld A (1989) Digital topology: introduction and survey. Comput Vis Graph Image Process 48(3):357–393

Laurini R (2014) A conceptual framework for geographic knowledge engineering. J Vis Lang Comput 25(1):2–19

Parent C et al (2009) Multiple representation modeling. In: Liu L, Özsu MT (eds) Encyclopedia of database systems. Springer, US, pp 1844–1849

Pires JM, Silva RA, Santos MY (2014) Reasoning about space and time: moving towards a theory of granularities. In: Computational science and its applications–ICCSA 2014. Springer, pp 328–343

Silva RA, Pires JM, Santos MY (2015a) A granularity theory for modelling spatio-temporal phenomena at multiple levels of detail. Int J Bus Intell Data Min 10(1):33

Silva RA, Pires JM, Santos MY (2015b) Aggregating spatio-temporal phenomena at multiple levels of detail. In: AGILE 2015. Springer Science Business Media, pp 291–308

Silva RA et al (2016) Enhancing exploratory analysis by summarizing spatiotemporal events across multiple levels of detail. In: AGILE 2016

Sips M et al (2012) A visual analytics approach to multiscale exploration of environmental time series. IEEE Trans Vis Comput Graph 18(12):2899–2907

Stell J, Worboys M (1998) Stratified map spaces: a formal basis for multi-resolution spatial databases. In: Proceedings 8th international symposium on spatial data handling. Department of Computer Science, Keele University, Staffordshire, UK ST5 5BG, pp. 180–189

Vilain MB (1982) A system for reasoning about time. In: AAAI, pp 197–201

Weibel R, Dutton G (1999) Generalising spatial data and dealing with multiple representations. Geogr Inf Syst 1:125–155

Yao JT, Vasilakos AV, Pedrycz W (2013) Granular computing: perspectives and challenges. IEEE Trans Cybern 43(6):1977–1989

Zhou X et al (2004) Multiresolution spatial databases: making web-based spatial applications faster. In: Yu J et al (ed) Advanced web technologies and applications SE—5. Lecture notes in computer science. Springer, Berlin, pp 36–47

Assessing Accuracy and Geographical Transferability of Machine Learning Algorithms for Wind Speed Modelling

Fabio Veronesi, Athina Korfiati, René Buffat and Martin Raubal

Abstract Machine learning is very popular in the environmental modelling community and has recently been demonstrated to be a useful tool for wind resource assessment as well. Despite the popularity of wind resource assessment, research in the field of machine learning for this purpose is in its infancy. Only few algorithms have been tested and only for specific areas, making it difficult to draw any conclusions in regards to the best wind estimation method at the global scale. In this study, we compared several machine learning algorithms with validation techniques specifically employed to not only assess their accuracy but also their transferability. In particular, we tested cross-validation techniques designed to test the accuracy of the estimation in the context of autocorrelation. This way we performed a benchmarking experiment that should provide end users with practical rules of application for each algorithm. We tested three families of popular algorithms, namely linear models, decision trees and support vector machines; each was tested using wind mean speed data and several environmental covariates as predictors. The results demonstrated that no single algorithm could consistently be used to estimate wind globally, even though decision-tree based methods seemed to be often the best estimators.

Keywords Machine learning · Geographical cross-validation · Geographic information systems · Wind resource assessment

1 Introduction

In recent years, machine learning algorithms have become very popular in the environmental modelling community, and have been employed for estimating numerous types of environmental variables, few examples are Recknagel (2001),

F. Veronesi (✉) · A. Korfiati · R. Buffat · M. Raubal
Institute of Cartography and Geoinformation, ETH Zurich,
Stefano-Franscini-Platz 5, 8093 Zurich, Switzerland
e-mail: 8093Zurich.fveronesi@ethz.ch

A. Bregt et al. (eds.), *Societal Geo-innovation*, Lecture Notes
in Geoinformation and Cartography, DOI 10.1007/978-3-319-56759-4_17

Grimm et al. (2008), Foresti et al. (2011), Marjanović et al. (2011). Of particular importance in the context of an increased push towards renewables that many countries are including in their energy policies, is the possibility to use machine learning as an alternative way to estimate the wind resource. The industry generally employs estimation methods that solve the physical equations that govern the motion of air in the atmosphere, i.e. numerical wind flow models. This class of algorithms tend to be very accurate, albeit extremely time-consuming and computationally expensive, but they are not perfect and in places they may record errors, intended as root mean square error (RMSE), of 1–2 m/s (Meteotest 2016). However, numerical wind flow models do not have a way to assess the variance of their estimates, in contrast to statistical methods such as kriging, meaning that planners cannot know the exact error of the model for each estimated location. As a consequence, maps created with numerical wind flow models are never used alone to plan new wind farms, but planners always set up additional measuring campaigns of several months or years before building (Argyle and Watson 2014), thus increasing the total cost of a project.

A potential way to solve all issues with numerical wind flow models, which was explored by the research community, is the use of statistical algorithms. For example, Luo et al. (2008) compared seven spatial interpolation methods to estimate wind mean speed across the UK, obtaining an average RMSE of 1.47 m/s. Another example is the work by Foresti et al. (2011), who presented multiple kernel learning regression methods as a modelling tool to map the wind field on complex terrain in Switzerland, obtaining RMSEs that ranged from a minimum of 0.98 up to 1.27 m/s. In Douak et al. (2013), the authors tested Kernel ridge regression to map wind speed in Algeria, obtaining a minimum RMSE of 1.4 m/s. In the most recent work, Veronesi et al. (2015, 2016), Veronesi and Grassi (2016) mapped wind speed and direction in the UK and obtained a RMSE of 0.7 m/s, which is the lowest ever reported for statistical wind resource assessment. Moreover, in this study the authors presented an uncertainty map that planners could use to assess the confidence interval of the energy output for each potential site. These results suggest that machine learning (contrary to geostatistical interpolation that performed poorly in literature tests) could be used to at least partly replace numerical wind flow models, and provide a better output that planners can readily adopt to locate ideal sites for wind farm projects, both in terms of long term estimates (e.g. production of wind atlases) and temporal distribution. However, a proper comparison between algorithms has never been carried out in the context of wind resource assessment. For this reason, researchers do not currently know which method is better suited for their particular type of analysis. To solve this issue we set up an experiment where several algorithms, previously employed in the literature for wind resource assessment, were thoroughly tested aiming not only at assessing their overall accuracy, but also at identifying their level of transferability, i.e. the possibility to use a model trained in one area to estimate another (Wenger and Olden 2012).

Generally speaking, the most widely used method of validation is cross-validation, which is a technique where part of the original dataset is excluded from training, and then used to test the estimation accuracy. The most common algorithm is the k-folds cross-validation, which divides the dataset into random sets or folds (James et al. 2013), then iteratively excludes one for testing while the others are used to train the algorithm. However, these machine learning algorithms generally "assume that data are identically distributed, and training and validation samples are independent" (Arlot and Celisse 2010), which cannot be guaranteed for environmental data that are affected by spatial autocorrelation. This implies that if the validation process selects the folds randomly, as it happens in k-folds, chances are that values almost identical to the test set would end up in the training set. This method is therefore not ideal for testing the transferability of the algorithm, since it may underestimate the error of the model (Ruß and Brenning 2010). However, transferability is a crucial factor to test since it allows for a correct estimation of the accuracy of the algorithm to estimate the wind resource in areas or even countries where no direct observations are available, which is of the uppermost importance for the wind industry.

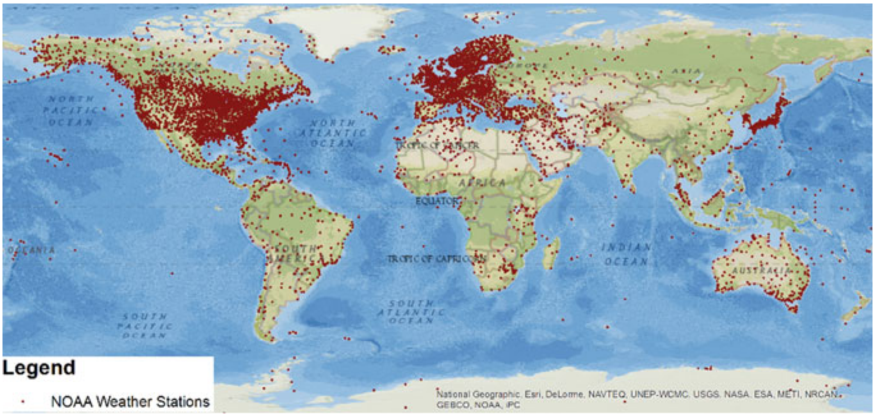
Based on these considerations, in this research we compared several machine learning algorithms, estimating the long-term wind speed average at the global scale. In order to not only test the overall accuracy of the methods, but also their transferability, this study employs various cross-validation techniques. The standard k-folds cross-validation was used as a benchmark, and in addition cross-validations based on geographical, elevation and land-use clustering were employed. These techniques allowed a detailed comparison of the accuracy and transferability of the machine learning algorithms, aiming at providing practitioners with basic rules to correctly and effectively apply machine learning to wind resource assessment.

Section 2 presents the weather data and the environmental predictors employed to perform this comparison. In this section, we also present a brief technical overview of the machine learning algorithms we compared, and the cross-validation techniques we used to assess the differences between algorithms. Section 3 presents the results of the comparison and discussions about the accuracy and degree of geographical transferability of each algorithm. Finally, Sects. 4 and 5 present the main conclusions of this work, its potential limitations and the future work that can be done to further increase our knowledge about machine learning applied to environmental modelling.

2 Materials and Methods

2.1 Dataset and Predictors

For this research we used the global wind speed dataset provided by the National Oceanic and Atmospheric Administration (NOAA 2016). This is a collection of daily weather averages from over 9000 stations worldwide. We collected a total of five years of daily data to calculate a robust wind mean speed for each weather station, which represents the variable we are estimating in this experiment. It is



National Geographic, Esri, DeLorme, NAVTEQ, UNEP-WCMC, USGS, NASA, ESA, METI, NRCAN, GEBCO, NOAA, iPC

Fig. 1 Global distribution of meteorological stations (NOAA dataset)

worth mentioning that in this experiment we are not dealing with time-series, but with the long term wind distribution. We calculated the wind mean speed over the five years we considered and that was our target variable. Figure 1 shows the distribution of meteorological stations globally.

As predictors for the analysis we used several environmental raster data, namely: environmental raster data from NASA (National Aeronautics and Space Administration), the land-use map by ESA (European Space Agency) and the global digital terrain model by USGS (United States Geological Survey). The Surface meteorology and Solar Energy dataset (Atmospheric Science Data Center) offered by NASA provides satellite raster data with global coverage. The dataset consists of monthly and annually averaged values of several parameters, e.g. solar radiation and climatic information, such as temperature, atmospheric pressure and relative humidity. For land-use information we used the GlobCover 2009 (Bontemps et al. 2011), provided by ESA. This dataset is the result of an automated classification of 22 land-use classes, purposively designed for global consistency. Finally, the Digital Elevation Model (DTM) utilized for this research was the Global Multi-resolution Terrain Data 2010 (GMTED2010) that is freely available from the USGS (Danielson and Gesch 2011). For each algorithm, all of the available predictors were used for training.

2.2 Machine Learning Algorithms

We assumed that more complex machine learning algorithms would provide higher accuracy compared with simpler ones, thus linear regression was selected to provide a lower baseline for the comparison. Then we included Lasso (Tibshirani 1996), which is a method based on linear regression but which also includes a penalty in solving the least squares equation. Thus, Lasso decreases the number of predictors in the estimation (this process is done automatically by the algorithm). Since it is well established that feature selection is an important point in a machine learning experiment (Buisson et al. 2010), Lasso was selected assuming that it would provide higher accuracy compared to simple linear regression. Another simple method we tested is k-nearest neighbor regression. This is probably the most simple estimation method available in the literature (James et al. 2013), but it allows more flexibility compared to linear regression. This method works by finding the closest subset of observations close to the estimation location (closeness defined in term of geographical distance) and averaging their values.

As mentioned, Veronesi et al. (2016) recently published a paper in which they used random forest (Breiman 2001), which is based on decision trees. To test whether decision tree based methods can be extensively employed to map the wind resource globally, we included some of them in this comparison. The simplest algorithm of the family is CART (Breiman et al. 1984), which basically fits a single decision tree to the entire dataset. This method is well established and easy to interpret, but it is not very accurate (James et al. 2013), and for this reason it was not included in our experiment. Even though fitting a single regression tree does not estimate the variable of interest with much accuracy, creating methods that fit multiple trees to the same dataset can solve that. These are referred to as ensemble methods and in this research random forest and extra tree regression were included. As mentioned, instead of fitting a single tree to the full dataset, these methods use an approach to fit multiple trees, based on random resampling of the original dataset through bootstrapping (Efron and Tibshirani 1994), which means that the resamples are drawn with replacements. For each resampled dataset, a tree is fitted with the same approach used in CART. The difference between random forest and extra tree regression comes after this initial resampling. In random forest for each tree a subset of the predictors is randomly chosen. Subsequently, for each node in the tree the best split among the subset of predictors is used to create it. This has the ability to decorrelate the trees, thus increasing the overall accuracy of the method (James et al. 2013). Extra tree regression (Pedregosa et al. 2011) goes one step further in terms of randomization. It starts from the same random subset from the pool of predictors, but then for each node instead of using the best possible split among the predictors, it first creates random thresholds values, i.e. potential splits, from all the predictors in the subset and then selects the best among them, meaning that it may not necessarily be the best overall split.

The final two methods were selected again because previously used in the literature for wind speed estimation. The first is support vector regression (Cortes and Vapnik 1995), which was used by Mohandes et al. (2004) for predicting wind speed. This is a mathematically complex model that uses nonlinear mapping to decrease the dimensionality of a high-dimensional feature space and then fits a linear regression model to it (Behrens and Scholten 2006). Another method closely related to support vector regression is kernel ridge regression (Murphy 2012), which was used by Douak et al. (2013) for wind resource assessment.

A mathematically comprehensive analysis of these two methods is provided by Smola and Schölkopf (2004).

2.3 Cross-Validation Methods

The most common method for performing cross-validation is the k-folds, where the dataset is divided randomly in either 5 or 10 subsets or folds (James et al. 2013). Since machine learning algorithms assume that training and testing folds are identically distributed and independent (Arlot and Celisse 2010), k-folds may not be ideal in the presence of spatial autocorrelation. In fact, Ruβ and Brenning (2010) argued that k-folds underestimates the prediction error in case of autocorrelation, and proposed a form of geographical cross-validation in which the folds are selected based on geographical clustering with a k-means algorithm. Similarly, Wenger and Olden (2012) studied the problem from an ecological perspective, and proposed a form of cross-validation again based on geographical selection of the folds, but in this case achieved by dividing the study area subjectively based on data density. In this research we decided to use the automatic approach based on clustering (Ruß and Brenning 2010), since it guarantees a statistically robust way of dividing the dataset. Since we wanted to maintain consistency among the validation techniques we decided to use 10 folds for all the validation methods tested.

Since wind speed is highly affected by changes in land use and elevation, additional cross-validation techniques were developed. We started with an elevation cross-validation, where the weather stations were clustered using only elevation as variable; this method was employed to divide the dataset into 10 folds. Additionally, we included a land-use cross-validation where the 22 classes provided in the GlobCover 2009 map were divided into 10 relatively homogeneous, in terms of similar land-uses, folds.

3 Results and Discussion

An important issue with the current network of weather stations is that it does not have sufficient coverage to be used alone for wind farm planning, as can be easily identified in regions such as South America, Africa and Asia from Fig. 1. Weather observations may have acceptable coverage in certain areas of the US, Europe or Japan, but in other areas they are very sparsely located. This is the main reason that justifies the development of new techniques for extensive wind resource assessment at high resolution.

In this study we tested several machine learning algorithms with varying degrees of complexity, ranging from the simple linear regression to the complex kernel ridge regression. To test the accuracy of each algorithm four methods of cross-validation were used to both calculate the overall accuracy of the models, and

Table 1 k-folds cross-validation results (k = 10)

Algorithm	RMSE (m/s)
Random forest	2.058
Extra trees regression	2.063
Lasso regression	2.181
Linear regression	2.190
Kernel ridge regression	2.193
k-neighbors regression	2.520
Decision tree regression	2.589
Support vector regression	3.105

also to test their transferability. We started our analysis by using a 10-folds cross-validation to benchmark and rank the machine learning algorithms. The results are presented in Table 1, where it is evident that there is not much difference between the accuracy of the various algorithms. When excluding support vector regression, the others have differences in terms of accuracy of maximum 0.5 m/s. The best overall model resulted to be random forest, but its results are very similar to simpler methods such as Lasso and linear regression.

To explain this observation, we need to take a step back and introduce the variance/bias trade-off. Machine learning algorithms try to model the pattern of the dependent variable of interest, in this case wind mean speed, as a function of the explanatory variables, in this case the environmental predictors, thus solving (Eq. 1):

$$Y = f(X) + \epsilon, \tag{1}$$

where Y is the target variable that has to be modelled, and $f(X)$ is a mathematical function of the predictors that the estimator creates to model Y. The component ϵ is a generic random error that is independent of X and has mean zero. It can be mathematically demonstrated that ϵ is the sum of two quantities, which are referred to as bias and variance (these are terms used in machine learning and may not have the same significate as the same terms used in other fields). Bias refers to the approximation error created by using strict functions. A linear model serves as an example: no matter how many observations are in the dataset and their general pattern, linear regression will always model them using a line. This creates an error that is intrinsic to the fact that the general shape of the function does not change. Thus linear regression is a biased method. On the contrary, variance measures the amount of change that the function experiences with changes in the training set. An example of a method with high variance can be a cubic spline. If applied to a dataset, since it fits a local polynomial of third order, it will probably fit most of the observations very closely. However, substantially changing the training data will also drastically modify the shape of the curve, since it will again try to fit all observations. Thus, this method has high variance and low bias. For a more comprehensive explanation please refer to James et al. (2013).

This trade-off is of crucial importance in explaining the results of the study. As we can observe from Table 1, biased methods are not the best estimators but are

Table 2 Results of the geographical cross-validation

Algorithm	RMSE (m/s)
Extra trees regression	2.211
Random forest	2.235
Linear regression	2.459
Kernel ridge regression	2.473
Lasso regression	2.476
k-neighbors regression	2.623
Decision tree regression	2.629
Support vector regression	3.036

certainly very close to the best result, achieved by random forest. This indicates that the pattern of wind speed and predictors is mostly linear, with some divergence that is picked up by decision trees. This may also be the reason why support vector regression, which assumes nonlinearity, resulted to be the worst overall. However, these results concerned the k-folds cross-validation, where training and test sets may be similar and thus it may underestimate the error of the models. This is the reason why we tested other cross-validation methods.

The first alternative method was the geographical cross-validation, in which we created 10 folds by clustering the geographical locations of the weather stations. This method of validation should be better suited for spatially autocorrelated data, since it is not affected by neighboring effects, and should also allow for testing the transferability of the algorithms. The results of the geographical cross-validation are presented in Table 2 and Fig. 2. The first thing to notice is that RMSE values have increased slightly compared to the k-folds, but not substantially, for almost all algorithms. The exception is support vector regression, for which the RMSE decreased by 0.1 m/s. Furthermore, algorithms based on ensembles of decision trees are once again the best performers, even though random forest is now in second place, followed by linear regression.

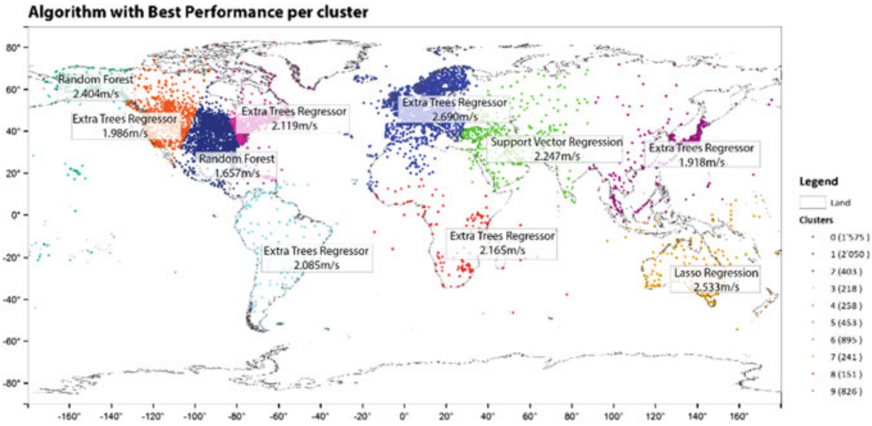


Fig. 2 Best performing algorithm for each geographical cluster

Going back to the variance/bias trade-off, these are very interesting results. We expected that this strict subdivision of the observations in areas where the predictors have values that may differ considerably from neighboring regions, would badly affect models with high variance, since this quantity is affected by changes in the training set. This did not happen, and in fact the ranking of the algorithms remained similar to the results from k-folds, meaning that decision tree based methods were the best performers. However, there is a slight but important difference between the results presented in Tables 1 and 2: extra tree regression overtook random forest as the best algorithm. As mentioned in Sect. 2.2, extra tree regression differs from random forest in the way in which it splits the predictors into nodes. While random forest takes the best split among the subset of predictors to create the nodes, extra tree regression creates random splits with all the predictors and takes the best split among this set, without necessarily picking the best split overall. This peculiarity allows a decrease of the variance and consequently an increase in bias (Pedregosa et al. 2011). In other words, this means extra tree regression is less affected by changes in the training set, compared to random forest. This may explain the results of the geographical cross-validation. In fact, by looking at the individual results from each geographic fold, we can better observe the increased accuracy of extra tree regression. Figure 2 clearly demonstrates that extra tree regression is the algorithm with the higher transferability among the ones tested. However, this is not the only important result. For example, it is interesting to notice that in Australia, linear regression is the best estimator with a RMSE of 2.5 m/s, which is 0.5 m/s higher compared to the overall best accuracy for Japan. Since Australia is environmentally so different from the rest of the world we may speculate that estimating it may be difficult. Therefore, it may be that linear regression becomes the best method simply because it is one of the few that can estimate outside the range of the predictors' values: we can just draw a longer line to cover uncharted territory. It is in fact well known that random forest is generally unable to estimate outside the range of predictors used for training (Veronesi et al. 2016). This is something practitioners should carefully consider when designing machine learning experiments.

Another important result is the fact that support vector regression recorded the lowest RMSE in the cluster around West Asia with a value of 2.2 m/s, much lower than the 2.8 m/s recorded by random forest. Since support vector regression assumes non-linearity in the data, this result is at odds with our previous discussions. However, it may well be that the wind pattern in this area is actually non-linearly correlated with the environmental predictors. Therefore, the only method able to identify this non-linearity is support vector regression. This may suggest to always consider including diverse machine learning models when planning to create ensembles of models, because even though the dataset has a general linear pattern, local changes may occur.

To further increase the level of details in the comparison we decided to build upon the concept of geographical cross-validation with two new methods based on elevation and land-use, which are two parameters that highly affect the wind

Table 3 Results of the elevation cross-validation

Algorithm	RMSE (m/s)
Kernel ridge regression	2.323
Lasso regression	2.332
Extra trees regression	2.335
Linear regression	2.379
Random forest	2.386
k-neighbors regression	2.747
Decision tree regression	2.957
Support vector regression	3.053

resource. In the first we clustered elevation, again creating 10 folds that were used in a similar fashion to the standard k-folds. The results are presented in Table 3 and Fig. 3.

This validation provided again very interesting results. In fact, the overall best model resulted to be kernel ridge regression, which in the k-folds cross-validation resulted to be the fifth overall in terms of accuracy, even though the first five methods had only slight differences in RMSE. However, it is interesting to notice that when elevation is specifically taken into account for validation, a method that assumes non-linearity resulted to be the best performer, while the other method with the same assumption, i.e. support vector regression, was last. This result supports findings from the literature. In particular, Foresti et al. (2011) used a method based on kernel regression to estimate the wind resource in Switzerland, obtaining relatively good results for an Alpine country with extremely complex terrain. Our results coupled with the findings from the literature suggests that kernel ridge regression is the most appropriate method to map complex terrain and high

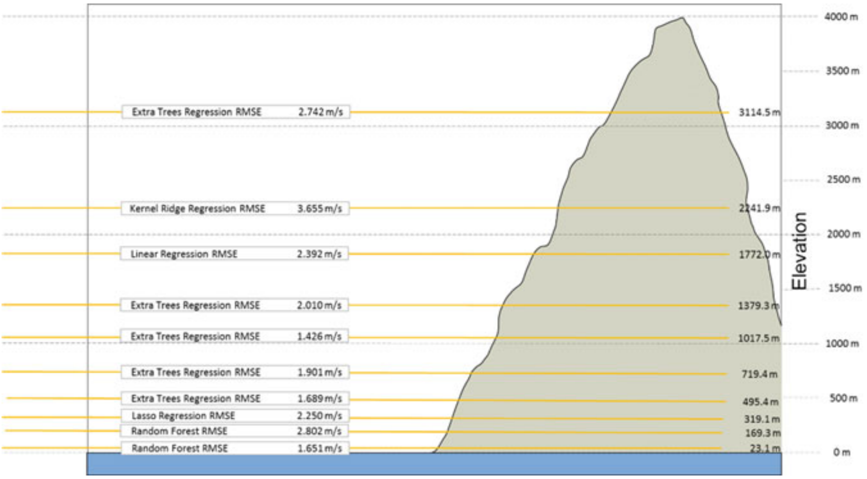


Fig. 3 Best performing algorithm for each elevation cluster

Table 4 Results of the land-use cross-validation

Algorithm	RMSE (m/s)
Random forest	1.893
Extra trees regression	1.911
Kernel ridge regression	1.949
Linear regression	1.952
Lasso regression	1.980
k-neighbors regression	2.318
Decision tree regression	2.462
Support vector regression	2.736

altitudes. Figure 3, which represents the best models divided by clusters of elevation, shows that decision tree based methods are generally the best performers. However, they all have relatively higher error rates at higher elevations, and this is the reason why on average kernel ridge has the lower error rate.

The final type of cross-validation we tested was based on land-use. The 22 classes provided by GlobCover 2009 were divided into 10 folds with the following general characteristics: Cropland, Mixed uses, Forest, Shrubland, Sparse vegetation, Urban, Bare ground area, Water body, Permanent snow and ice, Grassland.

The results of this validation are presented in Table 4, where all algorithms present lower RMSE values compared to all other cross-validations tested. It seems that land-use cross-validation overestimates the accuracy of the algorithms, even though from our analysis we cannot say for certain why that is. However, in terms of ranking these results follow what was found for k-folds and geographical cross-validation, i.e. decision tree based methods seem to be better suited for estimating global wind speed. This can also be concluded by looking at the results divided by land-use type, presented in Table 5.

Extra trees regression achieved the best performance for four out of the ten clusters (i.e. water body, forest, sparse vegetation, and mixed land-use). Random forest provided the best estimation results for three clusters (i.e. urban, grassland,

Table 5 Best performing algorithm for each land-use cluster

Land-Use	Algorithm	RMSE (m/s)
Cropland	Kernel ridge regression	3.465
Mixed	Extra tree regression	2.079
Forest	Extra tree regression	1.517
Shrubland	Random forest	2.167
Sparse vegetation	Extra tree regression	2.022
Urban	Random forest	1.250
Bare ground area	Linear regression	1.542
Water body	Extra tree regression	1.513
Permanent snow and ice	Kernel ridge regression	1.512
Grassland	Random forest	1.541

and shrubland). Kernel ridge and linear regression resulted in the best algorithm for another three land-uses: cropland, bare ground and permanent ice. Again, these results seem to suggest that when the conditions in the test set are very dissimilar to the training set, decision tree based algorithms do not perform well.

4 Conclusions

The main purpose of this research was the comparison and ranking of several machine learning algorithms, aiming at estimating wind speed globally. Such a comparison was never done in the literature, and may provide useful information to the environmental modelling community. Even though decision-tree based methods tend to consistently show a good level of accuracy, we cannot conclude that they can be used extensively for wind resource assessment. In fact, results consistently show that these algorithms tend to have a low level of transferability and when the test area is different from the training area, for example in terms of land-use or terrain complexity, these methods tend to have relatively low accuracy. Therefore, a reasonable conclusion from this study is that it is always a very good idea to know exactly the type of data available in terms of coverage, and the potential differences between training and test areas. If these differences are substantial, decision-tree based models are not the optimal algorithms, since their transferability is low. Additionally, it can also be concluded that even when the correlation between variable and predictors is mostly linear, one should try to test also methods that assume non-linearity, because it may be that in some subsets this assumption holds, as it was the case in our study for West Asia.

This research also demonstrated that k-folds cross-validation does not provide the full picture in terms of accuracy, and should be used carefully with data affected by spatial autocorrelation. Since the assumption of k-folds does not hold in case of autocorrelation, at least considering more than one cross-validation may be a wise way to accurately assess the accuracy of the model.

5 Limitations and Future Work

A thorough comparison of machine learning algorithms for global wind resource assessment has never been done in the literature and we believe our results, albeit limited in scope, may provide practitioners useful insights into this topic. However, having prioritized the comparison we did not focused on other aspects that will be covered in future research work. For example, we have only worked with environmental predictors such as DTM and other data from NASA, which are commonly used in geostatistical analyses. However, other satellite data are available and their use may increase the overall accuracy of the machine learning algorithms. Their relationships should not change, meaning that if regression trees resulted in the best estimators with our predictors, they would probably result in the best ones also with satellite data. Their accuracy may increase though, and this may help us draw conclusions about how machine learning compares with numerical wind flow models.

This brings us to the second point that we are planning to explore in future work, namely a benchmarking of machine learning algorithms in comparison with numerical wind flow models. Even though wind maps at high resolution at the global scale were never produced with numerical wind flow models, because it would require years of supercomputer work, some small areas were estimated and these results are available in the literature. For this reason, our next step is comparing our results with numerical models in areas where we have data to see their differences.

Acknowledgements The authors acknowledge the CTI—Commission for Technology and Innovation (CH), and the SCCER-FURIES—Swiss Competence Center for Energy Research—Future Swiss Electrical Infrastructure, for their financial support to the research activity presented in this paper. The authors would also like to thank NOAA, NASA, ESA and the USGS for providing open data that allowed us to perform this research.

References

Argyle P, Watson SJ (2014) Assessing the dependence of surface layer atmospheric stability on measurement height at offshore locations. J Wind Eng Ind Aerodyn 131:88–99

Arlot S, Celisse A (2010) A survey of cross-validation procedures for model selection. Stat Surv 4:40–79

Behrens T, Scholten T (2006) A comparison of data-mining techniques in predictive soil mapping. Dev Soil Sci 31:353–617

Bontemps S, Defourny P, Bogaert EV, Arino O, Kalogirou V, Perez JR (2011) GLOBCOVER 2009-Products description and validation report

Breiman L (2001) Random forests. Mach Learn 45:5–32

Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. CRC press

Buisson L, Thuiller W, Casajus N, Lek S, Grenouillet G (2010) Uncertainty in ensemble forecasting of species distribution. Glob Change Biol 16:1145–1157

Cortes C, Vapnik V (1995) Support-vector networks. Machine Learn 20:273–297

Danielson JJ, Gesch DB (2011) Global multi-resolution terrain elevation data 2010 (GMTED2010). US Geological Survey

Douak F, Melgani F, Benoudjit N (2013) Kernel ridge regression with active learning for wind speed prediction. Appl Energy 103:328–340

Efron B, Tibshirani RJ (1994) An introduction to the bootstrap. CRC press

Foresti L, Tuia D, Kanevski M, Pozdnoukhov A (2011) Learning wind fields with multiple kernels. Stoch Env Res Risk Assess 25:51–66

Grimm R, Behrens T, Märker M, Elsenbeer H (2008) Soil organic carbon concentrations and stocks on Barro Colorado Island—digital soil mapping using random forests analysis. Geoderma 146:102–113

James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning. Springer

Luo W, Taylor MC, Parker SR (2008) A comparison of spatial interpolation methods to estimate continuous wind speed surfaces using irregularly distributed data from England and Wales. Int J Climatol 28:947–959

Marjanović M, Kovačević M, Bajat B, Voženílek V (2011) Landslide susceptibility assessment using SVM machine learning algorithm. Eng Geol 123:225–234

Meteotest (2016) Windpotentialanalyse für Windatlas.ch Jahresmittelwerte der modellierten Windgeschwindigkeit und Windrichtung. BFE

Mohandes MA, Halawani TO, Rehman S, Hussain AA (2004) Support vector machines for wind speed prediction. Renew Energy 29:939–947

Murphy KP (2012) Machine learning: a probabilistic perspective. MIT press

National Oceanic and Atmospheric Administration (NOAA) (2016) Global surface summary of the day—GSOD

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830

Recknagel F (2001) Applications of machine learning to ecological modelling. Ecol Model 146:303–310

Ruß G, Brenning A (2010) Data mining in precision agriculture: management of spatial information. In: Computational intelligence for knowledge-based systems design. Springer, pp 350–359

Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. Stat Comput 14:199–222

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B (Methodol) 267–288

Veronesi F, Grassi S (2016) Generation and validation of spatial distribution of hourly wind speed time-series using machine learning. In: Journal of Physics: Conference Series. IOP Publishing, p 12001

Veronesi F, Grassi S, Raubal M (2016) Statistical learning approach for wind resource assessment. Renew Sustain Energy Rev 56:836–850

Veronesi F, Grassi S, Raubal M, Hurni L (2015) Statistical learning approach for wind speed distribution mapping: the UK as a case study. In: AGILE 2015. Springer, pp 165–180

Wenger SJ, Olden JD (2012) Assessing transferability of ecological models: an underappreciated aspect of statistical validation. Methods Ecol Evol 3:260–267

On Measures for Groups of Trajectories

Lionov Wiratma, Marc van Kreveld and Maarten Löffler

Abstract We present a list of measures for a single trajectory, including measures that require the presence of other trajectories, such as the centrality of a trajectory amidst other trajectories. Then, we introduce three different views in order to extend measures of a single trajectory to a group, namely the representative view, the complete view and the area view. Furthermore, we give measures that exist only for a group of trajectories, like density and formation stability. We also show that it may be possible to define new measures by combining trajectory data with data from other sources, such as the environment where the entities move. Finally, we discuss several tasks: settlement selection, visualization and segmentation, where measures on groups of trajectories are necessary.

Keywords Trajectories · Groups of trajectories · Movement attributes · Measures

1 Introduction

With the increased use and quality of GPS and other positioning devices, the analysis of *trajectory* data has become a mainstream topic in GIScience. A trajectory is the model of a moving entity. Trajectory data comes, for example, from vehicle, animal, hurricane, pedestrian, and sports player tracking.

We can use the *abstract model* or the *data model* for trajectories. In the abstract model, we consider the trajectory to be the representation of a moving object,

L. Wiratma (✉) · M. van Kreveld · M. Löffler
Department of Information and Computing Sciences, Utrecht University,
Utrecht, The Netherlands
e-mail: l.wiratma@uu.nl; lionov@unpar.ac.id

M. van Kreveld
e-mail: m.j.vankreveld@uu.nl

M. Löffler
e-mail: m.loffler@uu.nl

L. Wiratma
Department of Informatics, Parahyangan Catholic University, Bandung, Indonesia

A. Bregt et al. (eds.), *Societal Geo-innovation*, Lecture Notes
in Geoinformation and Cartography, DOI 10.1007/978-3-319-56759-4_18

assumed to be a point, that moves without discontinuities. In the data model, we realize that data on a moving object is generally collected by a device that gives a location at certain times where the location is sampled.

In the abstract model, a *trajectory* is a function T that maps a time interval $I = [t_\alpha : t_\beta]$ to the plane or 3-space. The time interval is the domain of this function. The image of the function on I is referred to as the *path* of the trajectory. The path has a shape and a direction but no time component. The location at the start time t_α is the *origin* of T and the location at the end time t_β is the *destination* of T.

In the (GPS) data model, a trajectory is a time-ordered sequence of triples $(x_1, y_1, t_1), ..., (x_m, y_m, t_m)$ where at time t_i, the moving object was recorded to be at location (x_i, y_i). We essentially do not know where the object was at times in between the samples. We could assume that locations are (sufficiently) precise, or incorporate imprecision in the analysis. The sampling rate and geometric precision are the two most important aspects of data quality for trajectories. Note that in the abstract model such quality aspects do not exist. If the sampling rate is sufficiently high and the precision as well, we can interpolate location and time between consecutive triples to make a continuous mapping from time to space. The triples can still be the most suitable representation.

We can distinguish the following main types of trajectory analysis:

- Segmentation (Anagnostopoulos et al. 2006; Buchin et al. 2011c): Partitioning a trajectory into segments so that within each segment, certain attributes of the trajectory are uniform. It may be used to split a trajectory into different movement behaviors.
- Similarity analysis (Buchin et al. 2011b; Liu and Schneider 2012): Analysing how much two trajectories appear alike. It can be used to determine how similar movement of a single entity is on different days, or to determine how similar the movement of two different entities is. The reasons for appearing alike can be rather varied: we could define similarity based on visiting the same locations, or based on having the same speed development (e.g. slow in the beginning, then linearly increasing speed).
- Clustering (Nanni and Pedreschi 2006; Lee et al. 2007): Grouping the trajectories in a collection based on similarity.
- Outlier detection (Lee et al. 2008a): Finding trajectories that do not belong to any cluster.
- Classification (Lee et al. 2008b): Assigning a trajectory to a cluster of trajectories, based on similarity.
- Hotspot detection (Gudmundsson et al. 2013): Finding places that are visited frequently and/or by many trajectories.
- Pattern detection: Finding interesting characteristics in one or more trajectories on any sort. Examples are leadership patterns (Andersson et al. 2007) and commuting patterns (Buchin et al. 2011a).
- Flocks, group, herd detection (Benkert et al. 2008; Buchin et al. 2015; Huang et al. 2008): Finding a special type of pattern determined by spatial proximity of a subset of the entities over a period of time. It is related to clustering, but

in clustering we generally consider the whole trajectory when making clusters, whereas in grouping a single entity can be in different groups at different times, or even at the same time.

All analysis types depend on *measures* defined on trajectories. In the listing above, we mentioned location, speed, and similarity as attributes that are defined in the abstract model, and can be computed in the data model. In fact, an attribute like speed gives rise to multiple measures: average speed, variation of speed, total stationary time, etcetera.

The purpose of this paper is to discuss measures for *groups of trajectories*. Recently, a number of different definitions of groups have been given, accompanied by algorithms that compute them from a collection (Hwang et al. 2005; Jeung et al. 2008; Buchin et al. 2015). Many analysis tasks that apply to trajectories exist for groups of trajectories as well. For example, we can imagine segmentation of the whole group at once, or doing a similarity analysis on two groups. To perform such analyses, we also need measures for the whole group. Groups are a natural unit of aggregation which can be analysed at once and which can be visualized using a single stroke. With the ongoing trend of dealing with larger and larger collections of data, aggregation is one of the approaches to cope with the growth.

In this paper we first provide an overview of the measures that exist for trajectories. We classify them into three types: measures for a single trajectory in isolation, measures for a single trajectory amidst other trajectories, and measures for a single trajectory in other environments. Then we proceed with extensions of these measures to groups, while at the same time including new measures that do not exist for single trajectories. We use the same classification into three types. When developing group measures, similar approaches are sometimes taken. We will highlight such more general approaches because they can potentially be used for other measures needed in specific applications.

2 Measures for a Single Trajectory

When defining measures for trajectories we will use the abstract model, since it is mathematically more clean and also representation-independent. All measures can be converted in various ways to measures in a data model.

Since a trajectory is a function from a time interval to the plane (or to space), the image of the function gives the locations where the entity is at the relevant times. The derivative of the function is the velocity, which is a vector, and the second derivative is acceleration, which is also a vector. Velocity has two components, namely speed and heading. These at any time measurable features (location, speed, heading, …) of a trajectory are called *attributes*, and attributes are often the basis of measures defined over a whole trajectory. For example, from speed we can derive the measures *average speed*, *maximum speed*, *standard deviation of speed*, *percentage of standstill*, and more. These measures give a single value for the whole trajectory.

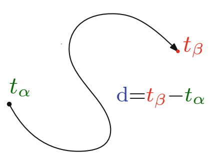
There are also attributes for trajectories amidst other trajectories. For example, for a given trajectory we have the attribute closest distance to another trajectory, which exists at any time. We can use this as the basis for a measure that intuitively corresponds to *degree of isolation* by averaging.

In this section we give an overview of various general-purpose measures that can be defined for a single trajectory. We first discuss measures for a single trajectory in isolation, then measures for a single trajectory amidst other trajectories, and then measures for a single trajectory in various contexts.

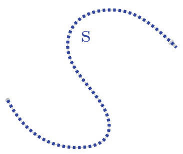
2.1 Measures for a Trajectory in Isolation

In this section we describe measures that relate to a single trajectory independent of other trajectories and the environment. Our list is not complete; we describe a number of existing useful, general-purpose measures. We begin with a number of basic, well-known measures capturing properties of the whole trajectory at once. Note that most of these common measures also appeared in similar lists of measures given in previous research (Laube et al. 2007; Dodge et al. 2008; Andrienko et al. 2008, 2013).

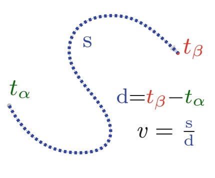
Duration

Description	*The length of the time interval of the trajectory (d)*			
Unit	second	Range	$[0,\infty)$	
Derived from	-			
Related works	(Calenge et al. 2009)			

Traversed distance

Description	*The total length of the path of the trajectory (s)*			
Unit	meter	Range	$[0,\infty)$	
Derived from	location			
Related works	(Yuan et al. 2010)			

Average speed

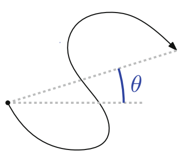
Description	*The ratio of the traversed distance and the duration (v)*			 s t_β t_α $\mathrm{d}=t_\beta-t_\alpha$ $v=\frac{\mathrm{s}}{\mathrm{d}}$
Unit	meter per second	Range	$[0,\infty)$	
Derived from	speed			
Related works	(Buard et al. 2011)			

While the average speed is a useful measure, it tells little about the variation in speed during the time interval. The entity could have been moving with constant speed equal to the average speed, or it could have stood still for half the time and then moved to the destination with speed double the average. Since this distinction is often important, we would like to have measures for it. The most obvious candidate is the standard deviation of speed over *I*.

In principle speed is an attribute that has a value at any time in *I*. To capture speed development over *I* in a more extensive manner, there are many options, both quantitative and qualitative. For example, we could split up *I* into *k* equal-duration subintervals and use the average speed in each of them. This may not be a good description of speed development because it might be more fine-grained than a chosen value of *k*. Ideally, one wants a partition of *I* into subintervals where the speed is similarly behaved. This can be obtained by segmentation.

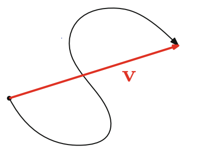
A derivative of speed is *acceleration* (and deceleration). Since acceleration exists at any time, we have a similar measure choices as for speed. For example, we can use the standard deviation of the acceleration to describe its variation.

Global direction

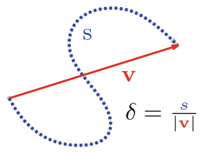
Description	*The direction of the vector from the origin to the destination of T (θ)*			
Unit	radian	Range	$[0,2\pi]$	
Derived from	location			
Related works	(LaPoint et al. 2013; Safi et al. 2013; Ranacher and Tzavella 2014)			

Also direction can be described more finely than with just a global direction, and as with speed, we can use subintervals of *I* to describe it. These subintervals may be obtained by equal durations or by segmentation.

Global velocity

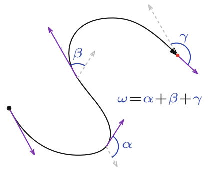
Description	*The vector from the origin to the destination of T divided by total duration (***v***)*			
Unit	vector	Range	$\mathbb{R}^2$	
Derived from	velocity, or location and time			
Related works	(Hanks et al. 2011)			

Detour

Description	*The ratio of the traversed distance and the length of the global velocity vector (δ)*			
Unit	-	Range	$[1,\infty)$	
Derived from	-			
Related works	(Buchin et al. 2011c; Boinski and Garber 2000)			

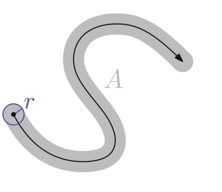
In other works, the notion of detour is described under different names, for example, sinuosity (Andrienko et al. 2013) and straightness index (Benhamou 2004).

Total angular change

Description	*The total change of the heading angle (ω)*			
Unit	radian	Range	$[0,\infty)$	
Derived from	heading			
Related works	(Calenge et al. 2009)			

The *total angular change* is a common shape descriptor that does not depend on time. It is the global version of the attribute representing the change in heading. It describes the shape in just one number on an angular scale.

Area covered

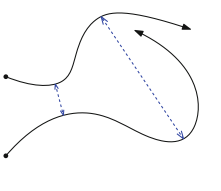
Description	*The area covered by a disc with radius r that moves along the trajectory (A)*			
Unit	square meter (m^2)	Range	$[0, \infty)$	
Derived from	location			
Related works	(Giuggioli et al. 2011)			

Finally, it may be useful to define the area covered by the trajectory. Andrienko et al. (2013) describe this measure as the *spatial extent* of a trajectory, which can be defined in several ways. A simple definition is to use a distance parameter r and say that the moving entity covers the whole area within distance r from its location. The covered area is then the total swept area of a disk centered at the entity when it follows its path; multiply swept areas count only once.

2.2 Measures for a Trajectory Amidst Other Trajectories

Some trajectory measures require the presence of other trajectories, for example measures for notions like similarity and centrality. For now we only consider other trajectories that exist at the same time as the trajectory that we observe. However, it is also possible to take other trajectories into account that occur later or earlier, by adapting the measures.

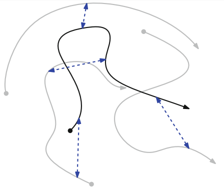
Similarity

Description	*The average distance to another trajectory*			
Unit	meter	Range	$[0, \infty)$	
Derived from	-			
Related works	(Nanni and Pedreschi 2006)			

Similarity has been studied extensively. It is a distance measure between two trajectories that can refer to average distance between the entities, maximum distance between the entities, or maybe just similarity in shape of the path of the trajectory. Well-known similarity measures for paths of trajectories are the Hausdorff distance (Alt et al. 1995) and the Fréchet distance (Alt and Godau 1995). Well-known similarity measures for trajectories are the time-focused distance, the dynamic time-warping distance and the edit distance. All of these similarity measures take the location into account. However, one can easily define speed similarity or acceleration similarity if location is not relevant in the application. Another similarity measure, introduced for shapes, is the turn function similarity. It is rotation-invariant, and therefore particularly useful for comparing shape only.

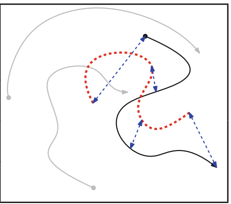
In this work, we define the similarity as the average distance between two trajectories during their time interval. To get the average distance, we can take the integral of the distance over the time interval and divide by the duration of the trajectories. Alternatively, one can take other possible options such as taking the minimum (or maximum) value, or integrating over space (reparameterize the trajectory).

Closeness

Description	*Average distance to the nearest other trajectory at each time*			
Unit	meter	Range	$[0, \infty)$	
Derived from	distance to the closest trajectory			
Related works	(Giardina 2008)			

The closest entity to a single moving entity at a single time can change over its time interval. Therefore, the average distance can show how close a trajectory is to other trajectories.

Centrality

Description	*Average distance to the central position of several trajectories*			
Unit	meter	Range	$[0, \infty)$	
Derived from	centrality			
Related works	(Agarwal et al. 2005; Laube et al. 2007)			

Centrality is also an intuitive concept that allows various different definitions. The centrality of an entity amidst other entities can change over time, so we first consider centrality measures at a fixed time. In other words, we have a set of points in the plane and one specific point p, and we want to describe how central p is. One option is to define the center of mass as the most central location, and use the distance of p to the center of mass to define p's centrality. We can do something similar with the median location, whose coordinates are the median x-coordinate and the median y-coordinate of all points. Another option is to use the distance to the center of the smallest enclosing circle of the points. For one trajectory, we take the average of centrality over its time interval, which is defined by an integral. Similar with the two previous measures, other options than taking the average are also possible.

A more combinatorial notion of centrality is the minimum number of points of the set that must be removed to bring p to the convex hull of the points.

Discussion. Other notions that can be turned into measures are isolation and sociality. To define these, we could use nearest-neighbor distance or k-nearest neighbor distances.

2.3 Measures for a Trajectory in an Environment

Since the environment influences how an entity moves, it is natural to consider trajectories in the context of other data. Other data can be internal or external, that is, it can refer to the moving entity or to its environment. Depending on the source of the data, internal attribute data can be heart rate, brain activity, CO_2 level in exhaust, and produced noise. External attribute data can be current landcover, elevation, weather, quality of road, water currents, and visibility information.

Environmental factors can influence how the measures defined before may be redefined. For example, two trajectories at distance 80 m in the open field can be deemed more similar than two trajectories at 60 m, one of which is in the field and one in the forest. A speed of 10 m/s in the open field may be considered similar to a speed of 8 m/s in bushland.

Moreover, the environment may give rise to measures that do not exist without it, like a measure for the diversity of landcover types crossed by the moving entity. Here, the locations of the moving entity is used to select other data (the environment) to which the measure relates.

The definitions of measures for trajectories in environments are very much application-dependent, and therefore we omit further discussion in this paper that aims to take a general view on measures for trajectories and groups.

3 Measures for a Group of Trajectories

When sufficiently many moving entities are close enough for a relatively long period of time, they form a *group*. Closely related concepts are flocks, clusters, herds, and convoys. Formal definitions are given in (Gudmundsson and van Kreveld 2006; Jeung et al. 2008; Buchin et al. 2015; van Kreveld et al. 2016; Li et al. 2010; Kalnis et al. 2005). Here, we assume that a group is a fixed set of entities during a fixed period of time. Intuitively, the existence of a group relies on some conditions, such as a minimum number of entities in it, a minimum entity inter-distance, a minimum duration, etc. It is possible to add further requirements to a group definition, like similarity of heading of the group entities.

Just like for a single trajectory, many measures exist for a group of trajectories. Some of these are direct extensions from the single trajectory case, while others only occur for a group. In general, there are three views of treating a group when defining these measures:

Representative view: A single trajectory is used (not necessarily from the group) to define the measure. For example, we can take the mean location at every time and use this to define a mean trajectory. This mean trajectory can then be used to define group speed, for instance.

Complete view: All entities in the group are used to define the measure. For example, to define group speed in a different way than in the representative view, we can take the average speed of each entity over the group duration, and average this over the entities.

Area view: The area the group occupies is used to define the measure. For example, to define group density, we could define the area that the group occupies and let the density be the ratio of the number of entities and the area.

Not all views make sense for all measures we will define. Often, these different views give rise to different possible measures for the same concept, like for our example of group speed. The representative view gives us a way to generalize all single trajectory measures to groups.

In the following subsections we define group measures that do not require other input, group measures that exist for a group amidst other groups of moving entities, and group measures that require other types of input. The measures exist for any definition of a group, assuming that the number of entities in the group does not change in the interval during which this group exists. We also assume that the group exists during exactly one time interval, so we do not allow a group to form, break up, and later form again and call it the same group. Both assumptions can be lifted if required by the application at hand; they are made for the sake of clarity of the definitions of the measures.

3.1 Measures for a Single Group in Isolation

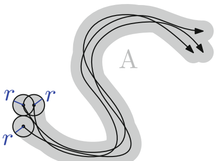
3.1.1 Measures Extended from a Single Trajectory

Almost all measures for a single trajectory in isolation can be extended directly to a group. Both the representative view and the complete view work well with these measures (except for the covered area). For example, the traversed distance of a group can be defined in two ways: as the average of the traversed distance of all entities in the group or as the traversed distance of a representative.

Furthermore, each of these two views can have more variations. With the complete view, instead of taking the average of the measure from each trajectory, we can also choose to take, for example, either the minimum or the maximum value. There are also several possibilities to get a representative trajectory. For example, to define the representative trajectory, we can take the mean or median point of all entities in a group at a single time, or just pick one trajectory to represent the group.

Only one measure for a single trajectory cannot be extended directly to a group:

Area covered

<table>
<tr><td>Description</td><td colspan="3"><i>The union of the total area covered by each trajectory (using a disc of radius r) in a group (A)</i></td><td rowspan="4">
</td></tr>
<tr><td>Unit</td><td>square meter (m^2)</td><td>Range</td><td>$[1, \infty)$</td></tr>
<tr><td>Derived from</td><td colspan="3">location</td></tr>
<tr><td>Related works</td><td colspan="3">(Boinski and Garber 2000)</td></tr>
</table>

Clearly, we cannot sum up the covered area from all entities in a group because they might overlap. For the same reason, it is also not possible to aggregate the covered area by a group at each single time. Therefore, we define the area covered of a group as a union of the total area covered by its entities.

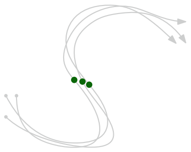
Note that this measure relies on a parameter *r* to define the disc which is used to sweep the covered area.

The movement behavior of a group influences this measure. For instance, when entities in a group move by following each other (*single file behavior*), then the group covers roughly the same area as one of its entities. A more compact group will also yield a smaller value for this measure than a more spread out group.

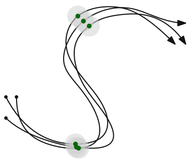
3.1.2 Measures that only Exist for a Group

In addition to attributes of a group that exist in a single trajectory, there are also attributes that only naturally exist when we observe multiple entities at a single time. For example, the number of entities in a group, (which is fixed for the whole duration of a group), or the *density* of a group.

Size

<table>
<tr><td>Description</td><td colspan="3"><i>The number of entities in a group (n)</i></td><td rowspan="4"></td></tr>
<tr><td>Unit</td><td>-</td><td>Range</td><td>$[2, \infty)$</td></tr>
<tr><td>Derived from</td><td colspan="3">the number of entities in a group</td></tr>
<tr><td>Related works</td><td colspan="3">(Beauchamp 2012)</td></tr>
</table>

Density

Description	*The average of the density of a group over its duration (ρ)*			
Unit	per square meter (m^{-2})	Range	$[\frac{1}{\pi r^2}, \infty)$	
Derived from	density			
Related works	(Peters and Krisp 2010; Beauchamp 2012)			

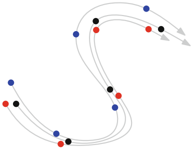
Intuitively, density relates a count to an area. Therefore, it is natural to define the density using an area view: a ratio of the size of a group and the covered area at a single time. Then, we integrate it over time to get a single value to represent the density of the group.

Other options to define the density of a group are based on the distance between entities in a group. We describe several possible definitions:

- The average of all distances between pairs of entities in a group.
- The weighted average of all distances, where small distances get a higher weight. Closer entities have more influence on the density than the farther ones.
- The ratio between the farthest distance occurring between two entities in the group and the maximum distance possible between two entities in any one group of the same size.

These distance-based measures have the advantage that they do not depend on a parameter to be determined.

Formation stability

Description	*The average of the magnitudes of the velocity difference of each entity and the group (F)*			
Unit	meter per second	Range	$[0, \infty)$	
Derived from	velocity			
Related works	(Heppner 1997; Viscido et al. 2004)			

A group of moving entities may move in a mostly stable formation, for instance, the V-formation of migrating birds. Often, entities keep the same formation for several reasons, like increasing the efficiency of movement by using the least energy.

At a fixed time, we define this measure using the magnitude of the difference in velocity of each entity and the group. We can average these magnitudes over the group entities and over time. Lower values for this measure indicate that the group has a low deformation rate. The formula to compute F for a group $G = \{e_1, e_2, \ldots, e_n\}$ over the duration $[t_\alpha, t_\beta]$ is:

$$F = \frac{1}{t_\beta - t_\alpha} \int_{t_\alpha}^{t_\beta} \frac{1}{n} \sum_{i=1}^{n} \|\mathbf{e}_i(t) - \mathbf{G}(t)\| dt$$

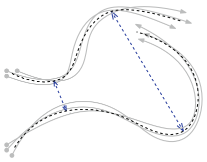
where $\mathbf{e}_i(t)$ is the velocity vector of entity e_i at time t and $\mathbf{G}(t)$ is the velocity vector of G, which is the average over all of its entities at time t.

Discussion. Many more application-specific group measures are conceivable. For instance, the occurrence of a *leader*. A leader can be defined as a single entity that is usually in front of the rest of the entities in the group, or as an entity whose movement drives the movement of the group (Boinski and Garber 2000; Andersson et al. 2007). Some groups are more clearly led than others, which can be captured in a measure. Because such measures depend more on behaviour interpretation than a clear objective aspect of groups, we will not suggest any formal definition in the present article.

3.2 Measures for a Group Amidst Other Groups

Most of the measures for a group among other groups can be directly defined using the representative trajectory of each group and apply to measures for a trajectory amidst other trajectories.

Similarity

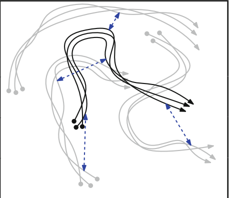
Description	*The average distance to another group*			
Unit	meter	Range	$[0, \infty)$	
Derived from	location			
Related works	(Bento 2016)			

To define similarity between two groups of entities we need to be aware of a few issues: (i) The two groups may have different size, and we still want a similarity measure that makes sense. (ii) The trajectories in one group may be dissimilar among each other, and if this occurs in the same way in the other group, then the groups are similar. These considerations suggest a many-to-many matching-based definition. For each trajectory in each group, consider the similarity to the most similar trajectory in the other group, and average over these similarity values. If the one group has n trajectories and the other m, then we use $n + m$ similarity values to average over. Alternatively, we could take all $n \cdot m$ pairs of trajectories, one from each group, and take their average, but this does not allow groups to be similar if their trajectories are not similar internally.

Alternatively, we can compare the similarity between the two groups at a single time and then integrate the result over the duration of the groups.

If we ignore the temporal component, then the area view can be used to compare the shape of the groups by comparing the shape of their covered area for a group shape similarity measure.

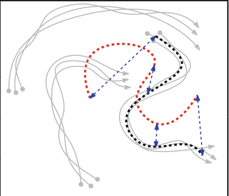
Closeness

Description	*The average distance to the nearest other group at each time*			
Unit	meter	Range	$[0,\infty)$	
Derived from	distance to the closest group			
Related works	-			

Although this measure is a straightforward extension from the measure for a single trajectory, applying it directly to the representative trajectories might not give the proper estimation about how close the two groups are. Intuitively, a group is near to the other groups if one or more entities in the group is close to entities in other groups. However, because the representative trajectory is located roughly in the middle of other entities in the group, there must be other pair of entities (each from different groups) that are closer.

Since the nearest other group might change over the duration of the group, it is better to first define the distance between two groups at a fixed time. Then, we can pick the minimum distance to one other group and get the average of this value over the whole duration of the group. At a single time, the distance between two sets of points can be defined in several ways. The simplest is to take the closest distance from any entity in one group to any entity in the other. Other options include the Hausdorff Distance (Alt et al. 1995) and the Earth Mover's Distance (Rubner et al. 2000). The question is whether we consider two groups to be very close when they are completely interspersed or when they move closely side by side as a whole group.

Centrality

Description	*The average distance to the central position of other groups*			
Unit	meter	Range	$[0,\infty)$	
Derived from	centrality			
Related works	-			

Centrality can be defined using the representative trajectories and then apply the centrality measure for a single trajectory. We can also use the complete view of a group by computing the centrality of each entity w.r.t the central position of all entities and then average them over time to get the centrality of the group.

Discussion. So far in this section, we only considered measures that assume all groups start and end at the same time. It is also possible to define measures if we drop this assumption. For example, we can measure how many other groups exist (on average) during the lifetime of a group, and we can measure how similar a group moves with respect to any previously existing group.

3.3 Measures for a Group of Trajectories in an Environment

Similar to the case for a single trajectory, we can consider environmental factors when defining the measure of a group. For example, entities in a group may change their formation because they are moving in single file formation while crossing a river, and may change back to the previous formation on the opposite bank. In this case, we probably want to say that the formation is consistent if the formation is the same in each terrain type. We cannot use formation stability to capture this.

Different types of moving entities show different behavior when moving as a group, and this is influenced by different types of external factors. Consequently, integrating those factors into measures for a group depends on the type of moving entities that we want to analyze. Therefore, measures that take external factors into account are application-dependent, and general purpose measures are less easily defined.

4 Applications of Group Measures

With the vast increase of trajectory data, large collections of trajectories must be organized well and good tools for selection, visualization, and analysis must be available. We discuss how group measures may be used for these tasks. We assume that a group structure exists and has been computed.

Selection. To explore a collection of trajectories organized in groups, we may want to select the ten or twenty most important or characteristic groups. What is important about a group is application-dependent and can be discussed, but generally one can say that large groups with a long duration are more important than small groups with a short duration. Any of the other attributes can contribute to the importance of a group too: perhaps fast-moving groups (high average speed) or groups with a large area covered are important. The more measures play a role, the more dimensions the Pareto-optimal front has, and the more Pareto-optimal choices there are.

While the single most important group may be defined by weighing the different measures that contribute to importance, selecting a set of ten most characteristic and important groups is more complex. In many applications the selection should be *varied*, implying that choosing one group in the set influences which other groups should also be chosen. The corresponding issues have been studied in the classical

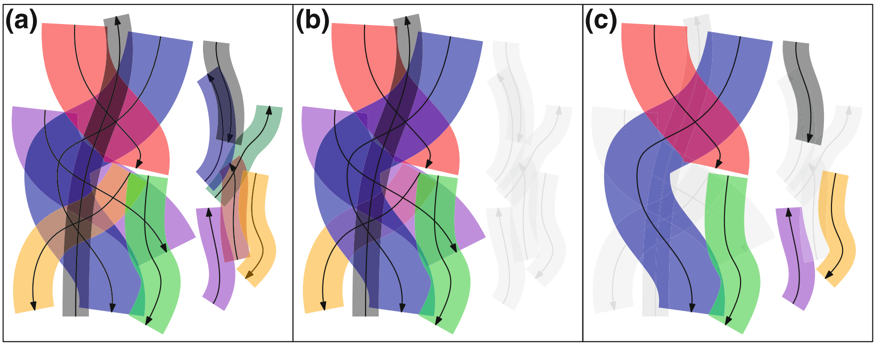


Fig. 1 **a** Each group is represented using a single *black* trajectory and an area with different colors. The width of the area shows the size of the group. **b** Only the six longest and biggest groups are selected, shown in color. **c** A better selection of six groups for diversity of the groups with respect to location.

problem of *settlement selection* (van Kreveld et al. 1997; Samsonov and Krivosheina 2012): not necessarily the ten biggest cities should be chosen, they should also be spatially distributed over the region of interest. Various models for such selections have been suggested in the literature. Transferring these issues to groups in collections of trajectories, it may be important to get variation in where in the plane (in space) the group occurred. If the six biggest and longest duration groups all occurred in the west, the next six ones occurred in the east, and we wish to select six groups, then it would probably be undesirable to select all groups in the west; instead, three or four in the west and three or two in the east gives a better idea of the data (see Fig. 1).

While a full treatment on when groups are important, salient or characteristic, and when a selection is an appropriate selection is beyond the scope of this paper, it is clear that group measures do play an important role in the process.

Visualization. Visualization of large collections of trajectories is difficult for several reasons. First of all, it is not easy to show time in an intuitive manner, besides animations. Second, simply showing all paths of trajectories usually creates a chaos in which little can be seen. If the trajectories come in groups, it is attractive to identify these groups and use visual encodings of relevant attributes or measures. A representative can be used to show each group (determined by trajectory measures), the group size can be visualized by line thickness of the representative, and the group speed or density can be visualized by color or other visual variable.

Segmentation. The segmentation of a single trajectory has been discussed in several papers (Mann et al. 2002; Buchin et al. 2011c; Aronov et al. 2013), but it may be more appropriate to *segment a group* in case the group as a whole shows certain behavior types. A herd of bison may be roaming, migrating, or sleeping. Segmentation is typically determined by within-segment similarity and across-segment dissimilarity and hence it relies on suitable measures. How to extend trajectory segmentation to trajectory group segmentation is an interesting research question in which it is clear that group measures will play an important role.

5 Conclusions and Future Work

In this paper, we discussed measures for a single trajectory and a group of trajectories arising from moving entities. These measures provide extra information than just the spatial and temporal component needed to be able to analyze the trajectory data better. First, we gave basic measures for a single trajectory and differentiated them into three types: measures for a single trajectory, measures for a trajectory amidst other trajectories, and measures for a trajectory in the context of other data.

For a group of trajectories, we introduced three different views to define measures for a group: the representative view, the complete view and the area view. Most measures for a single trajectory can be extended directly to groups using at least one of the three views. We also introduced exclusive measures for a group like density and formation stability. For measures for group amidst other groups, we discussed their differences with the same measure for a single trajectory and gave alternative approaches to define them.

Finally, we considered several tasks where group measures are essential to analyze collections of trajectories data.

Our discussion of measures gives rise to various directions of future research. Firstly, we can consider real-world trajectory data and investigate the need for measures beyond the ones we have given. It is also interesting to analyze how external factors like geographic context should influence measures in various applications. Secondly, the three cases selection, visualization, and segmentation of groups have only been addressed briefly, and a complete study of each of these tasks would be valuable. Thirdly, we may examine the measures further on robustness to "outliers" within a group. Finally, measures need to be computed by algorithms, which in several cases still need to be developed.

Acknowledgements Lionov Wiratma is supported by the Ministry of Research, Technology and Higher Education of Indonesia (138.41/E4.4/2015). Marc van Kreveld is partially supported by the Netherlands Organisation for Scientific Research (NWO) under project "More Content with Geometric Content". Maarten Löffler is partially supported by the Netherlands Organisation for Scientific Research (NWO) under project no. 614.001.504.

References

Agarwal PK, de Berg M, Gao J, Guibas LJ, Har-Peled S (2005) Staying in the middle: exact and approximate medians in R1 and R2 for moving points. In: Proceedings of the 17th Canadian conference on computational geometry, CCCG'05, University of Windsor, Ontario, Canada, 10–12 Aug 2005, pp 43–46

Alt H, Godau M (1995) Computing the fréchet distance between two polygonal curves. Int J Comput Geom Appl 05(01n02):75–91

Alt H, Behrends B, Blömer J (1995) Approximate matching of polygonal shapes. Ann Math Artif Intell 13(3):251–265

Anagnostopoulos A, Vlachos M, Hadjieleftheriou M, Keogh EJ, Yu PS (2006) Global distance-based segmentation of trajectories. In: Proceedings of the twelfth ACM SIGKDD international conference on knowledge discovery and data mining, Philadelphia, PA, USA, 20–23 Aug 2006, pp 34–43

Andersson M, Gudmundsson J, Laube P, Wolle T (2007) Reporting leadership patterns among trajectories. In: Proceedings of the 2007 ACM symposium on applied computing (SAC), Seoul, Korea, 11–15 Mar 2007, pp 3–7

Andrienko GL, Andrienko NV, Bak P, Keim DA, Wrobel S (2013) Visual analytics of movement. Springer, Berlin

Andrienko N, Andrienko G, Pelekis N, Spaccapietra S (2008) Basic concepts of movement data. In: Giannotti F, Pedreschi D (eds) Mobility, data mining and privacy, 1st edn. Springer

Aronov B, Driemel A, van Kreveld MJ, Löffler M, Staals F (2013) Segmentation of trajectories for non-monotone criteria. In: Proceedings of the twenty-fourth annual ACM-SIAM symposium on discrete algorithms, SODA 2013, New Orleans, Louisiana, USA, 6–8 Jan 2013, pp 1897–1911

Beauchamp G (2012) Flock size and density influence speed of escape waves in semipalmated sandpipers. Anim Behav 83(4):1125–1129

Benhamou S (2004) How to reliably estimate the tortuosity of an animal's path: straightness, sinuosity, or fractal dimension? J Theor Biol 229(2):209–220

Benkert M, Gudmundsson J, Hübner F, Wolle T (2008) Reporting flock patterns. Comput Geom 41(3):111–125

Bento J (2016) A metric for sets of trajectories that is practical and mathematically consistent. CoRR abs/1601.03094

Boinski S, Garber P (2000) On the move: how and why animals travel in groups. Nature/Science. University of Chicago Press

Buard E, Brasebin M, IGN S (2011) Visual exploration of large animal trajectories. In: 25th international cartographic conference (ICC11). ICC 2011, 3–8 July 2011 Paris, France, p 7

Buchin K, Buchin M, Gudmundsson J, Löffler M, Luo J (2011a) Detecting commuting patterns by clustering subtrajectories. Int J Comput Geom Appl 21(3):253–282

Buchin K, Buchin M, van Kreveld MJ, Luo J (2011b) Finding long and similar parts of trajectories. Comput Geom 44(9):465–476

Buchin M, Driemel A, van Kreveld MJ, Sacristán V (2011c) Segmenting trajectories: a framework and algorithms using spatiotemporal criteria. J Spat Inf Sci 3(1):33–63

Buchin K, Buchin M, van Kreveld MJ, Speckmann B, Staals F (2015) Trajectory grouping structure. JoCG 6(1):75–98

Calenge C, Dray S, Royer-Carenzi M (2009) The concept of animals' trajectories from a data analysis perspective. Ecol Inform 4(1):34–41

Dodge S, Weibel R, Lautenschütz AK (2008) Towards a taxonomy of movement patterns. Inf Vis 7(3):240–252

Giardina I (2008) Collective behavior in animal groups: theoretical models and empirical studies. HFSP J 2(4):205–219

Giuggioli L, Potts JR, Harris S (2011) Animal interactions and the emergence of territoriality. PLoS Comput Biol 7(3):1–9

Gudmundsson J, van Kreveld MJ (2006) Computing longest duration flocks in trajectory data. In: Proceedings of 14th ACM international symposium on geographic information systems, ACM-GIS 2006, 10–11 Nov 2006, Arlington. Virginia, USA, pp 35–42

Gudmundsson J, van Kreveld MJ, Staals F (2013) Algorithms for hotspot computation on trajectory data. In: 21st SIGSPATIAL international conference on advances in geographic information systems, SIGSPATIAL 2013, Orlando, FL, USA, 5–8 Nov 2013, pp 134–143

Hanks EM, Hooten MB, Johnson DS, Sterling JT (2011) Velocity-based movement modeling for individual and population level inference. PLoS ONE 6(8):1–17

Heppner F (1997) Three-dimensional structure and dynamics of bird flocks. In: Parrish JK, Hamner WM (eds) Animal groups in three dimensions: how Species Aggregate, 1st edn. Cambridge University Press

Huang Y, Chen C, Dong P (2008) Modeling herds and their evolvements from trajectory data. In: Cova TJ, Miller HJ, Beard K, Frank AU, Goodchild MF (eds) Geographic information science, Proceedings of the 5th international conference, GIScience 2008, Park City, UT, USA, 23–26 Sept 2008. Lecture notes in computer science, vol 5266. Springer, pp 90–105

Hwang S, Liu Y, Chiu J, Lim E (2005) Mining mobile group patterns: A trajectory-based approach. In: Ho TB, Cheung DW, Liu H (eds) Advances in knowledge discovery and data mining, Proceedings of the 9th Pacific-Asia conference, PAKDD 2005, Hanoi, Vietnam, 18–20 May 2005. Lecture notes in computer science, vol 3518. Springer, pp 713–718

Jeung H, Yiu ML, Zhou X, Jensen CS, Shen HT (2008) Discovery of convoys in trajectory databases. PVLDB 1(1):1068–1080

Kalnis P, Mamoulis N, Bakiras S (2005) On discovering moving clusters in spatio-temporal data. In: Medeiros CB, Egenhofer MJ, Bertino E (eds) Advances in spatial and temporal databases, Proceedings of the 9th international symposium, SSTD 2005, Angra dos Reis, Brazil, 22–24 Aug 2005. Lecture notes in computer science, vol 3633. Springer, pp 364–381

van Kreveld M, van Oostrum R, Snoeyink J (1997) Efficient settlement selection for interactive display. In: Auto-Carto XIII Proceedings of the International Symposium on Computer-Assisted Cartography, pp 287–296

van Kreveld M, Löffler M, Staals F, Wiratma L (2016) A refined definition for groups of moving entities and its computation. In: Proceedings of the 27th international symposium on algorithms and computation, ISAAC 2016, 12–14 Dec 2016, Sydney, Australia, pp 48:1–48:12

LaPoint S, Gallery P, Wikelski M, Kays R (2013) Animal behavior, cost-based corridor models, and real corridors. Landsc Ecol 28(8):1615–1630

Laube P, Dennis T, Forer P, Walker M (2007) Movement beyond the snapshot dynamic analysis of geospatial lifelines. Comput Environ Urban Syst 31(5):481–501

Lee J, Han J, Whang K (2007) Trajectory clustering: a partition-and-group framework. In: Proceedings of the ACM SIGMOD international conference on management of data, Beijing, China, 12–14 June 2007, pp 593–604

Lee J, Han J, Li X (2008a) Trajectory outlier detection: a partition-and-detect framework. In: Proceedings of the 24th international conference on data engineering, ICDE 2008, 7–12 Apr 2008, Cancún, México, pp 140–149

Lee J, Han J, Li X, Gonzalez H (2008b) TraClass: trajectory classification using hierarchical region-based and trajectory-based clustering. PVLDB 1(1):1081–1094

Li Z, Ding B, Han J, Kays R (2010) Swarm: mining relaxed temporal moving object clusters. PVLDB 3(1):723–734

Liu H, Schneider M (2012) Similarity measurement of moving object trajectories. In: Proceedings of the third ACM SIGSPATIAL international workshop on GeoStreaming, ACM, New York, NY, USA, IWGS '12, pp 19–22

Mann R, Jepson AD, El-Maraghi TF (2002) Trajectory segmentation using dynamic programming. In: 16th international conference on pattern recognition, ICPR 2002, Quebec, Canada, 11–15 Aug 2002, pp 331–334

Nanni M, Pedreschi D (2006) Time-focused clustering of trajectories of moving objects. J Intell Inf Syst 27(3):267–289

Peters S, Krisp JM (2010) Density calculation for moving points. In: Proceeding of the 13th AGILE international conference on geographic information science, Guimaraes, Portugal, 11–14 May 2010, pp 43–46

Ranacher P, Tzavella K (2014) How to compare movement? a review of physical movement similarity measures in geographic information science and beyond. Cartogr Geogr Inf Sci 41(3)

Rubner Y, Tomasi C, Guibas LJ (2000) The earth mover's distance as a metric for image retrieval. Int J Comput Vis 40(2):99–121

Safi K, Kranstauber B, Weinzierl R, Griffin L, Rees EC, Cabot D, Cruz S, Proaño C, Takekawa JY, Newman SH, Waldenström J, Bengtsson D, Kays R, Wikelski M, Bohrer G (2013) Flying with the wind: scale dependency of speed and direction measurements in modelling wind support in avian flight. Mov Ecol 1(1):4

Samsonov T, Krivosheina A (2012) Joint generalization of city points and road network for small-scale mapping. In: Proceedings of seventh international conference on geographic information science GIScience, Columbus, Ohio, USA, 18–21 Sept 2012, pp 18–21

Viscido SV, Parrish JK, Grünbaum D (2004) Individual behavior and emergent properties of fish schools: a comparison of observation and theory. Mar Ecol Prog Ser 273:239–249

Yuan J, Zheng Y, Zhang C, Xie W, Xie X, Sun G, Huang Y (2010) T-drive: driving directions based on taxi trajectories. In: Proceedings of the 18th ACM SIGSPATIAL international symposium on advances in geographic information systems, ACM-GIS 2010, 3–5 Nov 2010, San Jose. CA, USA, pp 99–108

Beyond Pairs: Generalizing the Geo-dipole for Quantifying Spatial Patterns in Geographic Fields

Rui Zhu, Phaedon C. Kyriakidis and Krzysztof Janowicz

Abstract With their increasing availability and quantity, remote sensing images have become an invaluable data source for geographic research and beyond. The detection and analysis of spatial patterns from such images and other kinds of geographic fields, constitute a core aspect of Geographic Information Science. Per-cell analysis, where one cell's characteristics are considered (geo-atom), and interaction-based analysis, where pairwise spatial relationships are considered (geo-dipole), have been widely applied to discover patterns. However, both can only characterize simple spatial patterns, such as global (overall) statistics, e.g., attribute average, variance, or pairwise auto-correlation. Such statistics alone cannot capture the full complexity of urban or natural structures embedded in geographic fields. For example, empirical (sample) correlation functions established from visually different patterns may have similar shapes, sills, and ranges. Higher-order analyses are therefore required to address this shortcoming. This work investigates the necessity and feasibility of extending the geo-dipole to a new construct, the geo-multipole, in which attribute values at multiple (more than two) locations are simultaneously considered for uncovering spatial patterns that cannot be extracted otherwise. We present experiments to illustrate the advantage of the geo-multipole over the geo-dipole in terms of quantifying spatial patterns in geographic fields. In addition, we highlight cases where two-point measures of spatial association alone are not sufficient to describe complex spatial patterns; for such cases, the geo-multipole and multiple-point (geo)statistics provide a richer analytical framework.

Keywords Spatial interaction · Multiple-point (geo)statistics · Geographic field analysis · Spatial pattern · Geo-multipole

R. Zhu (✉) · K. Janowicz
STKO Lab, Department of Geography, University of California, Santa Barbara,
Santa Barbara, USA
e-mail: ruizhu@geog.ucsb.edu

K. Janowicz
e-mail: jano@geog.ucsb.edu

P.C. Kyriakidis
Department of Civil Engineering and Geomatics, Cyprus University of Technology,
Limassol, Cyprus
e-mail: phaedon.kyriakidis@cut.ac.cy

A. Bregt et al. (eds.), *Societal Geo-innovation*, Lecture Notes
in Geoinformation and Cartography, DOI 10.1007/978-3-319-56759-4_19

1 Introduction and Motivation

The geo-atom, defined as $\langle x, Z, z(x)\rangle$, plays an important role in Geographic Information Science as the core representation of spatial information (Goodchild et al. 1999, 2007). The geo-atom associates a spatial location x with an attribute feature Z via the functional mapping $z(x)$. In terms of analysis, the geo-atom is applied in the computation of classical statistics used to describe aspects of spatial pattern in geographic fields. Examples of such statistics include the mean, variance, proportion of specific attribute values, and so on. Cell-based analysis of remotely sensed images with multiple attributes being available at each cell, form another common example of the usage of the geo-atom. Per-cell classification of low spatial resolution multispectral images is an example of such analytical operations. Finally, the geo-atom representation also applies to the object-driven perspective of spatial analysis, whereby each atom refers to an object rather than a cell.

The geo-atom considers each location x independently from other locations. This independence, however, ignores any interaction between locations, a critical aspect of geographic pattern (Goodchild et al. 2007). To address this shortcoming, Goodchild et al. (2007) introduced the concept of a geo-dipole, $\langle x, x', Z, z(x, x')\rangle$, whereby the interaction of variables between two locations x and x' is described via the two-point function $Z(x, x')$, with $z(x, x')$ as its one realization or estimation from a probabilistic perspective. Such interaction function often involves measures of similarity of attribute values at location pairs, along with their geographical (or other) distance. Statistics relying on the geo-dipole for exploring spatial patterns include the distance to nearest neighbor, Ripley's K, Moran's I, the correlogram, the semivariogram, and so forth. The same can be argued for interpolation, e.g., the interpolation of a temperature surface based on data obtained at monitoring stations, as well as for geographic contextual classification (Atkinson and Lewis 2000; Lu and Weng 2007; Congalton 1991), e.g., the improved classification of land use categories accounting for image texture information. Two-point statistics, such as the variogram that quantifies spatial auto-correlation, have been used several decades before the term geo-dipole was introduced; it was Goodchild et al. (2007), however, that first brought this notion into a more conceptual and theoretical level, which is the main focus of our work as well.

The geo-dipole considers a particular type of spatial interaction; namely, pairwise interactions. Those pairwise or two-point interactions are often (linearly) combined to arrive at interactions characterizing multiple locations, as, is done, for example, in spatial interpolation. In Kriging interpolation, in particular, the semivariogram model is first used to link each sample data location with a single interpolation location, and then such elementary two-point relations are combined through the Kriging system to arrive at interpolation weights. The entire procedure is based on explicit prior probabilistic models, such as the classic multivariate Gaussian model which is fully determined by its first-order statistics—the *mean* component—and its second-order statistics—the pairwise *covariance* function (Remy et al. 2009). However, these two-point models can only capture relatively simple spatial interactions, such as regularity, randomness, or clustering in attribute values. The identification and analysis of more complex spatial interactions, like those associated with curvilinear or other types of geometric structures, call for higher-order or multiple-point statistics.

In this work, we propose the *geo-multipole* as a new conceptual model in which the interactions among multiple locations are simultaneously quantified. To model this kind of multiple-point interactions, we employ higher-order statistics, namely multiple-point (geo)statistics together with their estimating approaches. In order to illustrate the necessity and feasibility of the geo-multipole, we compare it against the geo-dipole and classical two-point statistics for the recognition of urban spatial patterns. Although the geo-multipole concept could be employed to both object- and field-based representations of geographic information, this work focuses on methods and applications to geographic fields only.

The remainder of this paper is structured as follows: Sect. 2 briefly summarizes related work on analyzing and predicting geographic field patterns. In Sect. 3, the geo-atom and geo-dipole are approached from a probabilistic perspective in the context of geographic field analysis and then generalized to arrive at the notion of a geo-multipole. To motivate the need for the geo-multipole, Sect. 4 presents five contrasting spatial patterns extracted from remotely sensed images and compares them using two-point statistics and multiple-point statistics under the geo-dipole and the geo-multiple frameworks, respectively. Finally, Sect. 5 summarizes our results and highlights future research directions.

2 Related Work

In this section we review related work on geographic information representation and analysis required for the understanding of the proposed geo-multipole, as well as background material on multiple-point (geo)statistics.

2.1 Geographic Conceptualization

The conceptualization of geographic information has been discussed in GIScience since its emergence (Goodchild 1992a, b; Couclelis 2010). The core challenge is how to model (and distinguish) field-based and object-based views on geographic occurrences. Corresponding work includes the geographic field (G-Field) and object (G-Object), field object, object field, general field, and so on (Goodchild et al. 2007; Liu et al. 2008; Cova and Goodchild 2002; Voudouris 2010). To unify the multitude of concepts, Goodchild et al. (1999) introduced the *geo-atom* by which the former concepts can be generalized. In a later work, Goodchild et al. (2007) argued that these concepts are designed for describing the static distribution of features and attributes on the Earth surface, whereas dynamic processes of geographic phenomena require different conceptual models, i.e. interaction models. Goodchild et al. (2007) went further to propose the *geo-dipole*, in which the interaction between two locations is modeled. The authors demonstrated that the geo-dipole is capable of representing many analytical interaction models, such as object fields, metamaps, object pairs, and association classes. One common characteristic of these analytical models, however, is the property of pairwise interactions, which is also the conceptual foundation of the geo-dipole. For more complex, but nonetheless very frequent spatial patterns, e.g., those emerging in urban environments, the geo-dipole might not suffice to adequately model the complexity of spatial interactions, as more than two locations may be involved simultaneously in defining a pattern. For example, it makes sense to study a central market place located in a dense residential area with many individual private units, but it would be very limiting to observe only one pair, i.e., a private unit and the market center. Considering the interaction between many private units and the market center simultaneously is different from considering pairwise interactions between each private unit and the market center, e.g., when the task is to uncover a star-shaped pattern formed by the market and the incoming streets with their residential units. To the best of our knowledge, multiple-point interaction has been seldom formalized in conceptual models in GIScience, an exception being the concept of Markov (random) fields where spatial interaction is defined using higher-order cliques encompassing groups (triplets, quadruplets, and so forth) of pixels. In terms of applications, however, such higher-order interactions are rarely quantified, and inference in such fields accounts to considering pair-wise (two-point clique) interactions only.

2.2 Geographic Field Analysis

As discussed in Sect. 2.1, the *field* is one core concept of geographic information science (Kuhn and Frank 1991; Kuhn 2012). The detection and analysis of patterns from geographic fields, constitutes a critical task not only in geography, but also in related sciences, such as geology, environmental sciences, ecology, oceanography, and so on. Whether spatial context is explicitly considered or not distinguishes analytical approaches into non-contextual analysis and contextual analysis. Non-contextual analysis only focuses on individual cells and no interactions with neighbors are taken into account (Settle and Briggs 1987; Rollet et al. 1998; Fisher 1997). This type of approach is commonly used in the classification of hyper-spectral remote sensing images or the spatial prediction of many other multivariate geographic fields (Lu and Weng 2007). In contrast, contextual analysis introduces spatial patterns into the process of prediction and is frequently applied to high spatial resolution geographic fields (Li et al. 2014), including remotely sensed images. Depending on the way of incorporating spatial information, the analysis of fields can be categorized into distance-based and object-based approaches (Li et al. 2014).

Distance-based Analysis

In this approach, spatial patterns are described by pairwise dissimilarities between attribute values measured at locations separated by specific *distance* lags (Cressie 1993); examples include the variogram, the correlogram or transition probability diagrams. Such distance-based or two-point statistics are widely employed for incorporating spatial auto-correlation into interpolation and classification of field information (Atkinson and Lewis 2000; Remy et al. 2009). For classification purposes, in particular, distance-based spatial interaction pertaining to multiple attributes (reflectance values recorded in different spectral bands at each cell or pixel) has been used as a model of field (image) texture, and incorporated into the classification procedure via: (1) local (within a neighborhood template) sample or modeled variograms used as additional entries of the feature vector at each cell (Carr 1996; Carr and De Miranda 1998; Ramstein and Raffy 1989); and (2) multivariate variograms altering the weights originally attributed to entries of the feature vector, had classification been performed without accounting for spatial information (Oliver and Webster 1989; Bourgault et al. 1992). Variogram-based analysis of geographic fields, however, constitutes a two-point representation of spatial interactions, and typically invokes the rather limiting assumption of second-order stationarity (Remy et al. 2009).

Object-based Analysis

In object-based image analysis (OBIA), the field is first segmented into homogeneous areas, regarded as objects, and then the predictions about the cells contained within these objects are assumed to be the same (Blaschke 2010; Blaschke et al. 2014; Li et al. 2014). In OBIA, spatial information is considered in the process of segmentation, e.g., for Markovian methods (Jackson and Landgrebe 2002) and watershed methods (Salembier et al. 1998). Object-based analysis is commonly used in the classification and simulation of remotely sensed images and related work demonstrated the improvement over cell-based analysis (Blaschke 2010; Ceccarelli et al. 2013). However, OBIA is limited in terms of the assumption of homogeneous objects, the sensibility to segmentation algorithms, as well as the difficulty of using a large amount of conditioning data when it comes to generating patterns in a simulation setting (Remy et al. 2009).

2.3 Multiple-Point (Geo)statistics

Multiple-point (geo)statistics (MPS) were initially proposed to overcome the limitations inherent in variogram-based and object-based analysis for the identification of complex spatial patterns in the subsurface (Guardiano and Srivastava 1993). The core idea behind MPS is that, since variogram models are commonly estimated from data pertaining to analog deposits or outcrops or even expert-drawn images due to data limitations regarding the subsurface, why not directly borrow entire (conceptual) images as depositories of spatial patterns (Remy et al. 2009). It is these images that domain experts use to visually detect spatial patterns from and, thereby, estimate variogram model parameters. In addition, variograms being two-point statistics cannot capture spatial patterns resulting from complex earth processes. Implementing this idea, MPS abandons any explicit statistical model, but regards the training image as one realization of non-analytically defined random field pertaining to the actual (target) region being studied. The key assumption under MPS is that the training image contains adequate (in terms of complexity and number) replicates over the patterns deemed to occur at the target region (Strbelle 2002; Journel and Zhang 2006). Multiple-point statistics, e.g., the probability of three or more grid cells having simultaneously a particular lithological class, are then directly learned from the training image.

So far, most applications of multiple-point (geo)statistics (MPS) are limited to the domain of geology, in which subsurface heterogeneities, such as those found in porous media and reservoirs, are modeled and simulated (Strebelle et al. 2001). Several MPS algorithms have been implemented for applications in geology. Examples include simple normal equation sampling (Strébelle and Journel 2000), filter-based simulation (Zhang et al. 2006), and direct sampling (Mariethoz et al. 2010).

In recent years, two threads of applications of multiple-point (geo)statistics (MPS) can be distinguished for classifying geographic features, such as roads, buildings, vegetation, and open water-bodies, using remotely sensed images. Tang et al. (2016) incorporated MPS as new weights into K-nearest neighbor (KNN) classification and illustrated the improved performance compared to other supervised learning models such as Bayesian classifiers and Support Vector Machines. Others (Ge et al. 2008; Ge and Bai 2010, 2011; Ge 2013) introduced the Classification by Combining Spectral Information with Spatial Information in Multiple-point Simulation (CCSSM), in which MPS-based spatial classification is combined with pixel-based spectral classification using fusion techniques, such as consensus-based and probability-based fusion. The performance of the CCSSM approach compared favorably to traditional classification approaches, such as Maximum Likelihood Classification.

While these studies aim at improving the classification performance for remotely sensed images by applying multiple-point (geo)statistics (MPS), our work focuses on investigating the necessity and value of applying multiple-point interactions in analyzing geographic information, particularly geographic fields. We do so by generalizing the geo-dipole to stay within the conceptual framework proposed by Goodchild and others. Using our approach, MPS are not limited to classifications problems, but can also be used for interpolation, simulation, and so forth. Going beyond its recent practice in remote sensing, MPS could also be extended to other types of fields such as model outputs and irregular tessellations. Lastly, by introducing the geo-multipole, we hope to foster the development of GIScience-specific MPS algorithms that suit the needs and application areas of our community, e.g., for studying urban environments.

3 Introducing the Geo-multipole

In this section we introduce the geo-multipole as a conceptual generalization of the geo-dipole and also provide a probabilistic perspective on the geo-atom.

3.1 Conceptual Models

Capitalizing on the previously established conceptual models of the geo-atom $\langle x, Z, z(x)\rangle$ and the geo-dipole $\langle x, x', Z, z(x, x')\rangle$, we define the geo-multipole as follows:

$$\textbf{Geo-multipole} : \langle x, t_N, Z, z(x, t_N)\rangle$$
$$\text{where } t_N = \{x_1, \ldots, x_N\} \text{ are the } N \text{ neighbors of } x.$$

Here, we categorize conceptual models into three groups: (1) single-point data models, namely the geo-atom where no interactions between locations are considered; (2) two-point data models, namely the geo-dipole where pairwise interactions are considered; and (3) multiple-point data models, namely the proposed geo-multipole, which can be regarded as a generalized conceptualization of spatial interactions as defined by the geo-dipole. With respect to the geo-multipole, a neighborhood t_N, with N locations, is defined for each target location x. Then the interaction between x and its neighborhood t_N in terms of variable Z is defined as $Z(x, t_N)$, with the $z(x, t_N)$ as its one realization or estimation in a probabilistic perspective. The key difference between the geo-dipole and the geo-multipole is the fact that locations $x_1, \ldots, x_N$ in t_N are *simultaneously* considered (along with the corresponding attribute values) when modeling their interactions with x. In contrast, interactions are considered in *pairs* under the conceptualization of the geo-dipole despite that multiple pairwise interactions could be combined in *sequence*. It is important to note that simultaneously modeling interactions between the target and all its neighbors is mathematically different from simply combining pairwise interactions between each neighbor and that target. Namely, $Z(x, t_n = \{x_1, \ldots, x_N\})$ does not imply $f(Z(x, x_1), \ldots, Z(x, x_N))$.

3.2 Probabilistic Perspective

Geographic fields are frequently assumed to be generated from stochastic processes, and are thus regarded as realizations of a random field. Along the same lines, this work approaches the three conceptual models from a probabilistic perspective. Therefore, we discuss their descriptive statistics, as well as relevant estimation approaches in what follows.

3.2.1 Geo-atom

To summarize geographic fields in terms of the geo-atom, single-point statistics could be employed. The *mean* and *standard deviation* are the most commonly used examples. They are capable of describing the average magnitude, as well as the spread, of values of the attribute of interest across the domain. Other common statistics include *quantiles*, the *number* of cells whose attribute values satisfy a particular query, and the *probability density function* (PDF) of the attribute values:

$$f(z, x) = prob(Z(x) = z \pm \varepsilon)$$

where ε denotes an infinitesimally small value. It should be noted that, in this work, we use the term PDF also for the case of a categorical attribute Z, instead of the more correct notion of a probability mass function (PMF) for the sake of simplicity. For the same reason, we drop the $\pm\varepsilon$ notation from the PDF in what follows.

Optimal prediction at each location x, requires knowledge of the PDF $f(z, x)$. In the univariate case, the estimation of $f(z, x)$ only depends on the location x itself and no other variables at this location are provided. In addition, there is no interaction between the attribute value at this location and other locations. Therefore, unless the probability density function $f(z, x)$ is estimated by domain experts using physical models or experience considering a limited number of sample data $z(x_s; s = 1, \dots, S)$, it is challenging, if not impossible, to estimate the function from a probabilistic perspective.

In the multivariate case where the target variable Z is co-located with other variables Z', the relation between $Z(x)$ and $Z'(x)$ could be modeled through sample data; hence, the multivariate version of $f(z, x)$, i.e. $f(z, z', x) = prob(Z(x) = z|z'(x))$, could be estimated. This second case is common in GIScience. For example, if Z^{temp} is an unknown temperature field and we observe elevation z^{elev} and solar radiation z^{solar} as known fields, by using the sample data $\{z^{temp}(x_s), z^{elev}(x_s), z^{solar}(x_s); s = 1, \dots .S\}$, the relation between Z^{temp} and $\{Z^{elev}, Z^{solar}\}$ can be modeled through either linear or non-linear models. Then, the conditional PDF of the random variable Z^{temp} can be estimated by substituting z^{elev} and z^{solar} in the trained model. In remote sensing applications, per-cell classification is another example of this case, whereby reflectance values at different spectral bands form multiple feature variables and the class code at each cell forms the target categorical field.

3.2.2 Geo-dipole

Since the interaction between two points is now considered, concepts such as distance and neighborhood are key components of the geo-dipole. Statistics that could be used for describing spatial patterns via geo-dipoles are spatial autocorrelation measures, such as *Moran's I* and *Geary's C*, or their multiple lag-distance analogs, the correlogram and the semivariogram, for continuous data, and *transition probabilities* for categorical data. In addition, the *conditional PDF* in this case is modeled as:

$$f(z, x, z(x')) = prob(Z(x) = z|z(x'))$$

To predict $f(z, x, z(x'))$ within the geo-dipole framework, the key is to model the interaction $Z(x, x')$. Geostatistics provides approaches to model such interaction or association based on distance. For instance, under first- and second-ordered stationarity, $Z(x, x')$ could be characterized through semivariogram models whose parameters are estimated by sample data. Note here, that interaction among data of two different attributes can be also defined via cross-semivariogram models (Goovaerts 1997). Given the interaction model $Z(x, x')$, the conditional PDF of the random variable $Z(x)$ could be estimated through an observed variable as in the univariate case, or multiple observed variables as in the multivariate case. One example for the univariate case is interpolation, whereby a, say, temperature field can be interpolated using limited sample data. Interpolation methods, such as inverse distance weighting and Kriging, account for pairwise interactions between $Z^{temp}(x)$ and $Z^{temp}(x_s; s = 1, \dots, S)$. Land use classification using multi-spectral remote sensing images is an example of the multivariate case. In contrast to incorporating data pertaining to only one spectral band, multiple bands, together with their modeled cross-interactions, are used to arrive at land use classifications.

3.2.3 Geo-multipole

In contrast to the geo-dipole, the geo-multipole takes the N neighbors of x into account *simultaneously*. Rather than two-point statistics, higher-order statistics are thus required to model such a multiple-point interaction $Z(x, t_N)$. Similar to the geo-dipole, such multiple-point statistics could be obtained through sample data. However, the size of sample data sets is typically relatively small for such a multiple-point inference endeavor; this might result into biased estimates. A more promising approach is to use training images, which are assumed to contain spatial patterns deemed representative of the actual field under study. Multiple-point interactions $Z(x, t_N)$ are then directly learned from the training image without building any parametric model. Specific algorithms to accomplish this are discussed in Sect. 3.3. The *conditional PDF* of the random variable $Z(x)$ at a target location x can then be built, from which an optimal prediction can be derived for the attribute Z at that location:

$$f(z, x, z(t_N)) = prob(Z(x) = z|z(t_N))$$

The geo-multipole is appropriate for analyzing geographic fields that have rather complex spatial patterns. Examples include categorical fields that pertain to urban structures, such as roads that exhibit curvilinearity patterns or rooftops that have polygonal shapes. The geo-multipole could also be used for spatial interpolation, e.g., for air pollution patterns, and spatial simulation, e.g., of urban growth. Concrete examples of utilizing the geo-multipole, together with a comparison to the geo-dipole are given in Sect. 4.2.

3.3 *Higher-Order Statistics*

The geo-multipole concept employs higher-order statistics with respect to the geo-dipole, whereby two-point statistics are considered. Therefore, the question of how to efficiently compute or model such higher-order or multiple-point statistics becomes a key challenge. Although different algorithms exist for implementing multiple-point (geo)statistics (MPS), the core idea is to use the training image as an analog for learning higher-order spatial patterns. Basic elements of MPS algorithms are (Mariethoz and Caers 2014):

- **Training images (*TI*)** that contain spatial patterns; see Fig. 2
- **Template (*T*)** for scanning training images; see column 1 of Fig. 4
- **Data events (*dev*(*x*))**, which are simultaneous (joint) combinations of attribute values at template cells; see column 2 of Fig. 4

Since templates are used to detect spatial patterns, attribute values at more than two points are simultaneously considered in MPS. After obtaining data events from training images, multiple point statistics, i.e. $prob(Z(x) = z|z(t_N))$, can be calculated (Honarkhah and Caers 2010). Together with actual or directly sampled data, e.g., land cover classes verified at particular cells from ground surveys, these learned conditional probability values can subsequently be applied to estimate, or simulate, attribute values at non-sampled locations.

In our work, we implement one of the many MPS algorithms available, namely simple normal equation simulation (SNESIM), to estimate the required higher-order statistics. We then employ simulation to generate synthetic images of fields, in order to visually explicate the patterns learned by MPS. Several steps are involved in SNESIM (Remy et al. 2009): (1) a search template T_j is first defined; (2) a search tree specific to template T_j is then constructed; (3) the conditioning data are located on the field (this step can be skipped for unconditional simulation); (4) a random path that visits all locations to be simulated is established; then for each location x along the path: (5) the conditioning data event $dev_j(x)$ defined by template T_j is selected; (6) the corresponding conditional probability from the search tree is retrieved; (7) and finally the simulated value from the conditional probability is generated and added to the conditioning data set. In this work, we make use of the Matlab library mGstat[1] to run SNESIM; illustrative examples are given in Sect. 4.2.

Note that higher-order statistics are different from classic map algebra or image processing operations, e.g., focal or zonal operations, or kernel filters. Higher-order statistics consider neighboring interactions simultaneously rather than splitting them into (weighted) linear combinations of pairwise interactions. In classical map algebra operations, neighbors are considered in a first-order (linear combination of neighboring attribute values) or at most second-order (neighboring attribute values weighted as function of pairwise distances) manner.

[1] http://mgstat.sourceforge.net/.

4 Case Study

In this section we demonstrate the utility of the geo-multipole concept in describing spatial patterns. We do so by means of employing multiple-point statistics to highlighted use cases, and comparing the results against those obtained by solely relying on the geo-dipole and therefore two-point statistics such as variograms.

4.1 Sample Patterns

To illustrate the benefits of introducing the geo-multipole, as well as the feasibility of applying multiple-point (geo)statistics, we extracted several spatial patterns (shown in Fig. 2) in the form of binary maps from remotely sensed images (shown in Fig. 1). The binary maps are derived from remotely sensed images by threshold-based brightness segmentation to distill target patterns. The proportions of black cells in those maps are quite similar (pattern 1: 0.2697, pattern 2: 0.258, pattern 3: 0.257, pattern 4: 0.267, pattern 5: 0.269). The spatial patterns in the five binary maps, however, are rather different. Pattern 1 is extracted from streams, thus showing curvilinear patterns; patterns 2 and 3 are extracted from vegetation of a park and a golf court, respectively, and thus show circular patterns; pattern 4 is extracted from a residential area with rectangular patterns; and finally pattern 5 is extracted from the public garden of a mission, showing bounded patterns of different simple shapes.

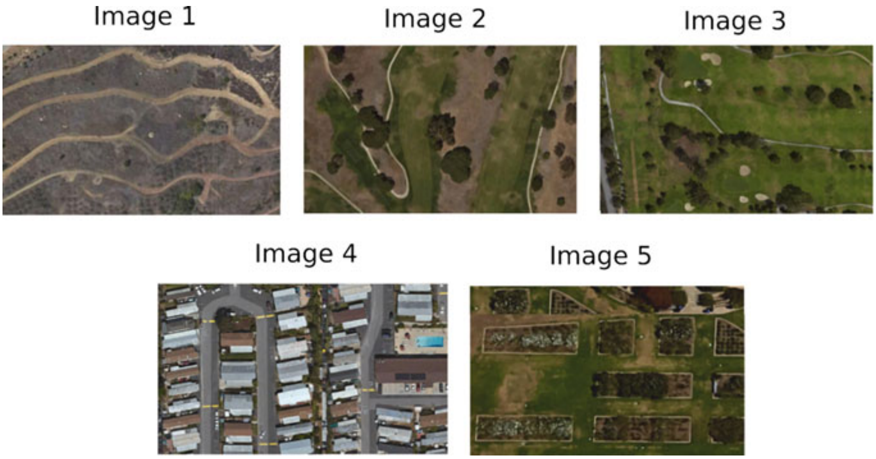


Fig. 1 Five remotely sensed images at 1 m spatial resolution

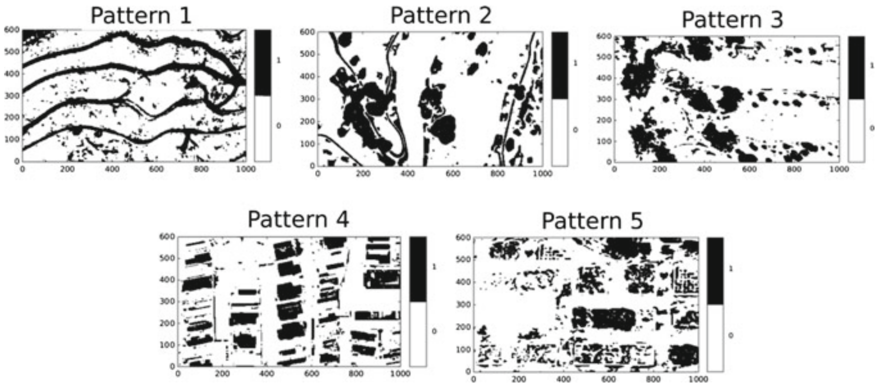


Fig. 2 Binary maps of five spatial patterns (600 × 1000)

4.2 *Experimental Results and Discussion*

The geo-dipole and the geo-multipole are compared in this section using the five patterns described in Sect. 4.1. Specifically, variogram-based and MPS-based approaches are applied for quantifying the selected patterns. Two sets of experiments are conducted in both approaches to highlight their differences: (1) a statistical description of the pattern, and (2) a simulation expression of the pattern, visualizing the information contents conveyed by this description.

4.2.1 Description of the Pattern

Directional semivariograms and conditional multiple-point probabilities are calculated to show their ability to characterize the selected spatial pattern. As the five examples show distinctive spatial patterns visually, the more different the results of the employed statistics are, the more successful the methods are in detecting distinct complex patterns.

Variogram-based Analysis

The two-directional semivariograms (i.e., West-East and North-South) for the five examples are illustrated in Fig. 3. As can be seen, despite the visually different patterns, their semivariograms for the two directions are generally similar, with a dramatic increase from distance lag 0 to about 50. The semivariograms also remain flat after the distance lag 60. The only salient characteristic is the bump at the distance lag 50 of the North-South semivariogram for pattern 1; this is due to the repetition of multiple elongated (along West-East) features of relatively regular width (pseudo-periodicity). This observation indicates that two-point (geo)statistics are

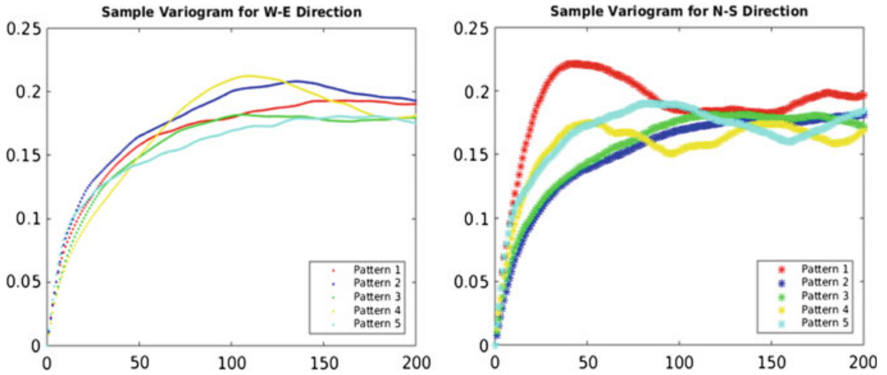


Fig. 3 Directional semivariograms for the five examples (*Left* West-East; *Right* North-South)

barely enough to capture the complex spatial patterns embedded in these (urban) structures.

MPS-based Analysis

In multiple-point (geo)statistics (MPS), one computes the conditional probability of class occurrence given nearby classes in the template directly from the training image. The order of the statistics employed is determined by the size and geometry of the template. The lager the template, the more neighboring locations will be simultaneously considered. To show the capability of MPS in detecting different spatial patterns, a simplified template (see the first column of Fig. 4) was used for pattern 1 and pattern 4. To determine the class, i.e., black (1) or white (0), at the central cell, its 8 neighbors are simultaneously considered as data events shown in column 2. The class of the central cell will be assigned to the one that has the highest conditional probability. A data event's conditional probability is calculated as the frequency of occurrence, for example:

$$prob(Z(x) = 0|z(t_N)) = \frac{\#(Z(x) = 0|z(t_N))}{\#(Z(x) = 0|z(t_N)) + \#(Z(x) = 1|z(t_N))}$$

There are 2^8 possibilities for such a neighborhood configuration; in this work we sampled 6 of them for illustration purposes. From Fig. 4, we can see that the conditional probabilities using the 3×3 template are different between pattern 1 and pattern 4. Note that only a relatively simple template is tested here; had a more complicated template, such as a 80×80 square template, been used, the conditional probabilities would be even more different. Such an observation indicates the capability of MPS for learning complex patterns compared to the simple semivariogram-based analysis.

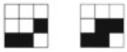
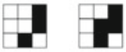
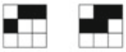
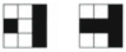
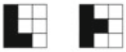
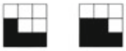
Template	Data Events	P(x| n₁, … , n₈)			
		Pattern 1		Pattern 4	
		P(x=0| n₁, … , n₈)	P(x=1| n₁, … , n₈)	P(x=0| n₁, … , n₈)	P(x=1| n₁, … , n₈)
$n_1 n_2 n_3$ $n_4 \times n_5$ $n_6 n_7 n_8$		0.000	**1.000**	**0.600**	0.400
		**0.600**	0.400	0.300	**0.700**
		0.500	0.500	**0.625**	0.375
		0.500	0.500	0.000	**1.000**
		**0.530**	0.470	0.458	**0.542**
		**0.610**	0.390	0.487	**0.513**

Fig. 4 Conditional multiple-point probabilities for patterns 1 and 4 (only 6 out of $2^8 = 256$ possibilities are shown)

4.2.2 Simulation of Pattern

To visualize the information content of variograms and multiple-point statistics, unconditional simulations are conducted using modeled variograms and multiple-point (geo)statistics (MPS), respectively. Our rationale here is that the more similar the simulated patterns are to the original examples, the more feasible the approach is in terms of learning spatial patterns.

Variogram-based Simulation

Unconditional moving average simulation via the Fast Fourier Transform (FFT) was used in this work to simulate realizations of 2-D multivariate Gaussian fields given the semivariogram model (i.e., exponential model); the resulting continuous images were then thresholded using suitable cutoff values so as to reproduce the same proportion of black cells as the corresponding original binary images. From the Fig. 5, we have mainly two observations: (1) although the original patterns 1 and 4 (in Fig. 2) show two different spatial patterns, their variogram-based simulations demonstrate similar spatial patterns; (2) the spatial patterns of both simulations in Fig. 5 are not consistent with the original patterns (i.e., the curvilinear pattern of pattern 1 and polygonal pattern of pattern 4). These two observations showcase the limitations of using variograms for simulating (and thus analyzing) spatial patterns; one should not

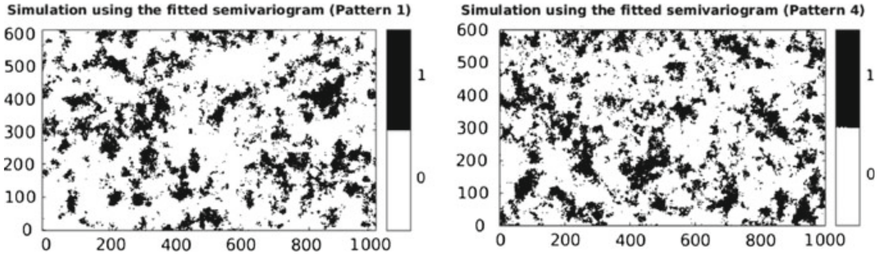


Fig. 5 Variogram-based simulations (pattern 1 and pattern 4)

expect a two-point variogram function to capture complex higher-order spatial patterns as those corresponding to elongated features or other curvilinear or geometric shapes.

MPS-based Simulation

The simple normal equation simulation (SNESIM) was applied in this work to generate the MPS-based simulation using the original patterns 1 and 4 as training images. The templates for both patterns were set to 80 × 80 squares. From the results in Fig. 6, we can observe that the two simulations show significantly different patterns, with pattern 1 showing more curvilinearity along the west-east direction and pattern 4 showing more polygonal geometries. Furthermore, comparing the simulated images with the original training images (in Fig. 2), we observe that the illustrated patterns in the simulations are relatively similar to the ones from the original images, although there are still inconsistencies between the two. A viable explanation for such inconsistencies is that the original training images (Fig. 2) are small in size and their patterns are rather complex with many elementary patterns being combined. For example, there are only four curved lines, which is the main pattern visually in the pattern 1, but there are also many small clusters across the domain. Summing up, despite some inconsistencies, the advantage of using MPS for learning spatial patterns is clearly highlighted by these simulations; particularly when compared to variogram-based approaches. Evidently, more work is required (possibly involving testing different MPS-based simulation methods) in order to improve the similarity between

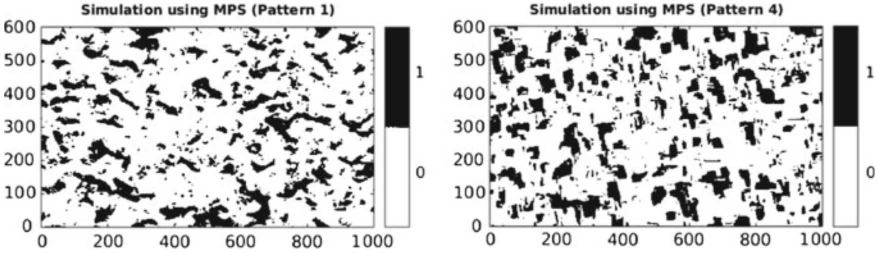


Fig. 6 MPS-based simulations (pattern 1 and pattern 4)

simulated and training images. It should be stressed, however, that the generation of spatial patterns with geometrical characteristics is a very improbable outcome when using variogram-based simulation algorithms.

5 Conclusions and Future Work

In this work, we generalized the traditional two-point interaction model, the geo-dipole, by introducing the geo-multipole concept whereby multiple-point interactions are simultaneously modeled. Furthermore, a general framework for geographic field analysis was discussed from both geographic and probabilistic perspectives. All three conceptual models, the geo-atom, the geo-dipole and the geo-multipole, are included in the framework, and they represent statistics of different order, i.e. first-order statistics for the geo-atom, second-order for the geo-dipole and higher-order for the geo-multipole. Different descriptive statistics, prediction techniques, and concrete examples were given to demonstrate such a framework.

This work also discussed the application of multiple-point (geo)statistics (MPS) as one potential approach for estimating higher-order statistics for geographic fields. In MPS, the training image is regarded as an explicit (non-parametric or better multi-parametric) model that replaces the role of implicit statistical models. The only assumption in using MPS is that the training image contains a representative collection of the spatial patterns expected at the target site; thus, the target field characteristics can be learned using approximate replicates contained in the training image. It should be noted, however, that since MPS places extreme “faith” in the training image, there is a risk of over-parameterization; thus, more attention should be placed on selecting appropriate training images, possibly considering more than one such images (Mariethoz and Caers 2014).

A series of experiments were conducted to illustrate the necessity and value of using the geo-multipole in quantifying patterns in field data. In short, we showed that spatial patterns extracted from multiple-point (i.e., MPS-based) interaction models are more realistic (better reproduce the complexity of patterns) compared to the ones extracted from two-point (i.e., variogram-based) interaction models.

There are several potential research directions for future work. First, the application details of multiple-point (geo)statistics (MPS) for quantifying spatial patterns in geographic phenomena should be further explored. For example, the sensitivity of template geometry and size, the impact of the training image size and pattern richness, as well as the feasibility of using other algorithms, should be studied in more depth. Second, in addition to using MPS for contextual classification, MPS could also be applied to spatial simulations. For example, the performance of cellular automata could be improved by incorporating information from training images using MPS. Last but not least, techniques for estimating multiple-point interactions could be extended from applications pertaining to field information to applications involving other types of geographic information as well. For example, higher-order interactions among different places (objects) in gazetteers could be considered to supplement spatial signatures for place types when learning alignments between geo-ontologies, as proposed in Zhu et al. (2016).

References

Atkinson PM, Lewis P (2000) Geostatistical classification for remote sensing: an introduction. Comput Geosci 26(4):361–371

Blaschke T (2010) Object based image analysis for Remote sensing. ISPRS J Photogramm Remote Sens 65(1):2–16

Blaschke T, Hay GJ, Kelly M, Lang S, Hofmann P, Addink E, Feitosa RQ, van der Meer F, van der Werff H, van Coillie F et al (2014) Geographic object-based image analysis-towards a new paradigm. ISPRS J Photogramm Remote Sens 87:180–191

Bourgault G, Marcotte D, Legendre P (1992) The multivariate (co) variogram as a spatial weighting function in classification methods. Math Geol 24(5):463–478

Carr JR (1996) Spectral and textural classification of single and multiple band digital images. Comput Geosci 22(8):849–865

Carr JR, De Miranda FP (1998) The semivariogram in comparison to the co-occurrence matrix for classification of image texture. IEEE Trans Geosci Remote Sens 36(6):1945–1952

Ceccarelli T, Smiraglia D, Bajocco S, Rinaldo S, De Angelis A, Salvati L, Perini L (2013) Land cover data from landsat single-date imagery: an approach integrating pixel-based and object-based classifiers. Eur J Remote Sens 46:699–717

Congalton RG (1991) A review of assessing the accuracy of classifications of remotely sensed data. Remote Sens Environ 37(1):35–46

Couclelis H (2010) Ontologies of geographic information. Int J Geogr Inf Sci 24(12):1785–1809

Cova TJ, Goodchild MF (2002) Extending geographical representation to include fields of spatial objects. Int J Geogr Inf Sci 16(6):509–532

Cressie N (1993) Statistics for spatial data: wiley series in probability and statistics, vol 15. Wiley-Interscience, New York, pp 105–209

Fisher P (1997) The pixel: a snare and a delusion. Int J Remote Sens 18(3):679–685

Ge Y (2013) Sub-pixel land-cover mapping with improved fraction images upon multiple-point simulation. Int J Appl Earth Obs Geoinf 22:115–126

Ge Y, Bai H (2010) Mps-based information extraction method for remotely sensed imagery: a comparison of fusion methods. Can J Remote Sens 36(6):763–779

Ge Y, Bai H (2011) Multiple-point simulation-based method for extraction of objects with spatial structure from remotely sensed imagery. Int J Remote Sens 32(8):2311–2335

Ge Y, Bai HX, Cheng Q (2008) New classification method for remotely sensed imagery via multiple-point simulation: experiment and assessment. J Appl Remote Sens 2(1):023537–023537

Goodchild MF (1992a) Geographical data modeling. Comput Geosci 18(4):401–408

Goodchild MF (1992b) Geographical information science. Int J Geogr Inf Syst 6(1):31–45

Goodchild MF, Egenhofer MJ, Kemp KK, Mark DM, Sheppard E (1999) Introduction to the varenius project. Int J Geogr Inf Sci 13(8):731–745

Goodchild MF, Yuan M, Cova TJ (2007) Towards a general theory of geographic representation in gis. Int J Geogr Inf Sci 21(3):239–260

Goovaerts P (1997) Geostatistics for natural resources evaluation. Oxford University Press on Demand

Guardiano FB, RM Srivastava (1993) Multivariate geostatistics: beyond bivariate moments. In Geostatistics Troia92. Springer, pp 133–144

Honarkhah M, Caers J (2010) Stochastic simulation of patterns using distance-based pattern modeling. Math Geosci 42(5):487–517

Jackson Q, Landgrebe DA (2002) Adaptive bayesian contextual classification based on markov random fields. IEEE Trans Geosci Remote Sens 40(11):2454–2463

Journel A, Zhang T (2006) The necessity of a multiple-point prior model. Math Geol 38(5):591–610

Kuhn W (2012) Core concepts of spatial information for transdisciplinary research. Int J Geogr Inf Sci 26(12):2267–2276

Kuhn W, Frank AU (1991) A formalization of metaphors and image-schemas in user interfaces. In: Cognitive and linguistic aspects of geographic space. Springer, pp. 419–434

Li M, Zang S, Zhang B, Li S, Wu C et al (2014) A review of remote sensing image classification techniques: the role of spatio-contextual information. Eur J Remote Sens 47:389–411

Liu Y, Goodchild MF, Guo Q, Tian Y, Wu L (2008) Towards a general field model and its order in GIS. Int J Geogr Inf Sci 22(6):623–643

Lu D, Weng Q (2007) A survey of image classification methods and techniques for improving classification performance. Int. J Remote Sens 28(5):823–870

Mariethoz G, Caers J (2014) Multiple-point geostatistics: stochastic modeling with training images. Wiley

Mariethoz G, Renard P, Straubhaar J (2010) The direct sampling method to perform multiple-point geostatistical simulations. Water Resour Res 46(11)

Oliver M, Webster R (1989) A geostatistical basis for spatial weighting in multivariate classification. Math Geol 21(1):15–35

Ramstein G, Raffy M (1989) Analysis of the structure of radiometric remotely-sensed images. Int J Remote Sens 10(6):1049–1073

Remy N, Boucher A, Wu J (2009) Applied geostatistics with SGeMS: a user's guide. Cambridge University Press

Rollet R, Benie G, Li W, Wang S, Boucher J (1998) Image classification algorithm based on the RBF neural network and k-means. Int J Remote Sens 19(15):3003–3009

Salembier P, Oliveras A, Garrido L (1998) Antiextensive connected operators for image and sequence processing. IEEE Trans Image Process 7(4):555–570

Settle J, Briggs S (1987) Fast maximum likelihood classification of remotely-sensed imagery. Int J Remote Sens 8(5):723–734

Strbelle S (2002) Conditional simulation of complex geological structures using multiple point geostatistics. Math Geol 34(1): 122Strbelle

Strébelle S, Journel A (2000) Sequential simulation drawing structures from training images

Strebelle SB, Journel AG et al (2001) Reservoir modeling using multiple-point statistics. In: SPE annual technical conference and exhibition. Society of Petroleum Engineers

Tang Y, Jing L, Atkinson PM, Li H (2016) A multiple-point spatially weighted k-nn classifier for remote sensing. Int J Remote Sens 37(18):4441–4459

Voudouris V (2010) Towards a unifying formalisation of geographic representation: the object-field model with uncertainty and semantics. Int J Geogr Inf Sci 24(12):1811–1828

Zhang T, Switzer P, Journel A (2006) Filter-based classification of training image patterns for spatial simulation. Math Geol 38(1):63–80

Zhu R, Hu Y, Janowicz K, McKenzie G (2016) Spatial signatures for geographic feature types: examining gazetteer ontologies using spatial statistics. Transactions in GIS

Part III
20 Years of AGILE

20 Years of AGILE

Hardy Pundt and Fred Toppen

Abstract Preceded by a ‘personal word’ from the authors, the focus of this paper is on a number of facts and activities that together describe how AGILE developed from an idea in the garden of the late Peter Burrough in Wageningen in July 1997 to a mature and well known organization nowadays. After a short description of those very first years where the reason for existence was discussed that resulted in the AGILE mission, the structure of AGILE as an organization is presented. How the mission was translated into a range of activities is described in a number of paragraphs, where the AGILE working groups, initiatives, the participation in EU projects and conference themes are discussed. The paper will finish with some personal impressions by those who acted as chairpersons of AGILE.

Keywords AGILE history · GI science · European co-operation

1 A Personal Word Beforehand

We thank the organization committee of the AGILE conference 2017 in Wageningen, giving us space in the proceedings to review AGILE as an organization including some “historic” issues, covering 20 years in which a lot of ideas and developments grew up.

During the past 20 years, AGILE (https://agile-online.org/) has developed towards a lively, well established, and widely appreciated organization with around 85 member organizations throughout Europe. The Association holds diverse international links to other communities. Even after 20 years, AGILE lives due to

H. Pundt
Harz University of Applied Sciences, Wernigerode, Germany
e-mail: hpundt@hs-harz.de

F. Toppen (✉)
Faculty of Geosciences, Utrecht University, Utrecht, The Netherlands
e-mail: f.j.toppen@uu.nl

A. Bregt et al. (eds.), *Societal Geo-innovation*, Lecture Notes
in Geoinformation and Cartography, DOI 10.1007/978-3-319-56759-4_20

its passionate members, who have strong interest in supporting the advancement of research and education in Geographical Information Science (GIScience).

An organization such as AGILE lives through the people who support the association in various ways: giving talks and writing papers for the conferences, coordinating working groups or initiatives to promote European research and teaching in GIScience, contributing opinions and findings in discussions, and written documents, or carrying out challenging workshops, as well as EU projects that attract the attention of many other interested people. Without the AGILE member organizations—but foremost the individuals who act as representatives of their laboratories at AGILE events—the association would not have survived.

The foregoing AGILE conferences, the inaugural meeting of which took place in Enschede, the Netherlands, in 1998, were great events. Excellent keynotes, remarkable sessions, inspiring workshops, fruitful discussions and unforgettable gala dinners are treasured as memories by many participants. Conferences, workshops, and sessions are brilliant places to communicate with colleagues. Linking with colleagues from different European countries and beyond means to enhance one's horizon, to broaden one's own worldview, and to learn. Breaks between sessions, the receptions and gala dinners are places where ideas were born that led to new initiatives, scientific papers, projects or at least to contacts between scientists that can be used to build new networks. Another aspect should be mentioned: a very important point of all the AGILE conferences has been the fact that people from many different nations and continents came together and discussed in an atmosphere of acceptance, friendliness, and respect. They spoke to each other peacefully and paying tribute to different views and perspectives. In such a sense, AGILE has provided a framework in which science and education, research and development can grow freely.

Twenty years of AGILE also means that many of us who accompanied the Association during the last two decades have been getting older. This is why the young researchers are of great importance to AGILE. With several activities, AGILE supports young researchers in GIScience. Grants to participate in conferences to present ideas and research, as well as the Ph.D. School are two of the measures that were initiated by AGILE to support young Ph.D. students that guarantee progress in our science. The young researchers are invited to engage in AGILE and to guarantee that AGILE will stay as a lively and communicable organization, bringing scientists from different countries and continents together, aiming at the goals of progress and advancement in geographic information science.

2 A Twenty-Year Journey Through GIScience

In 1998 the inaugural AGILE conference took place at the ITC Enschede, the Netherlands. Now, after two decades, the conference takes place again in the Netherlands. Wageningen, the "town of life sciences" is a well-chosen host for the twentieth AGILE conference.

Historically, AGILE has been built on over 10 years of experience of collaboration in GI research at the European level starting from the EGIS conference series (1990–94), the JEC-GI conferences (1995–1997) and the GISDATA programme of the European Science Foundation (1993–97). With the end of GISDATA and the JEC-GI series, a certain vacuum occurred concerning GIScience on a European level. Consequently, some researchers who were active within this sector for many years developed the idea of a new organization which could fill the empty space (Masser 1999).

In this context the garden session at Peter Burrough's place is memorable, not only because of the food served and the great weather, but because here it was decided to start AGILE and to have a first conference in Enschede in 1998. These early main drivers of geoinformatics in Europe believed strongly that an annual European conference dealing with geographic information science was urgently needed to guarantee continuous progress in this relatively young field of R&D. Since its advent, AGILE has been seeking to ensure that the views of the geographic information teaching and research community are fully represented in the discussions that take place on future European research agendas. For that reason, AGILE wanted to develop towards a permanent scientific forum where geographic information researchers can meet and exchange ideas and experiences at the European level.

Today we can conclude that these goals have been achieved. During the past 20 years the AGILE community has developed continuously, thus exploring the geospatial domain and seeking for innovative solutions that have an impact on society, exchanging ideas and carrying out fundamental spatial research.

3 The AGILE Mission

Since the first conference in Enschede, twenty conferences in total have been carried out with a large number of participants from nearly all European nations, as well as the US, Brazil, Japan, Israel, Russia, Africa, Australia and other countries. The conference, organized as an annual event, has led to high publicity of the acronym "AGILE". In a step-by-step process many measures were carried out to enhance the scientific quality of the conference (The Springer Series). The Springer book series, the first issue published at the conference 2007 in Aalborg (Fig. 1) is one of the outcomes of this process, with now ten books on the shelf.

The annual conference functions as a focus and this 20th international AGILE conference in Wageningen is the best proof that AGILE is an active association that assembles a significant number of geographic information laboratories throughout Europe. They follow AGILE's mission (Fig. 2) in several different ways.

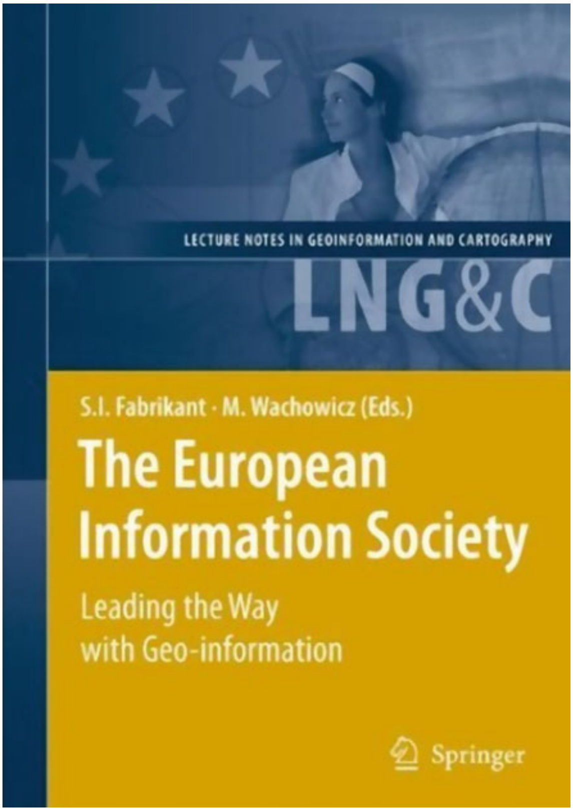
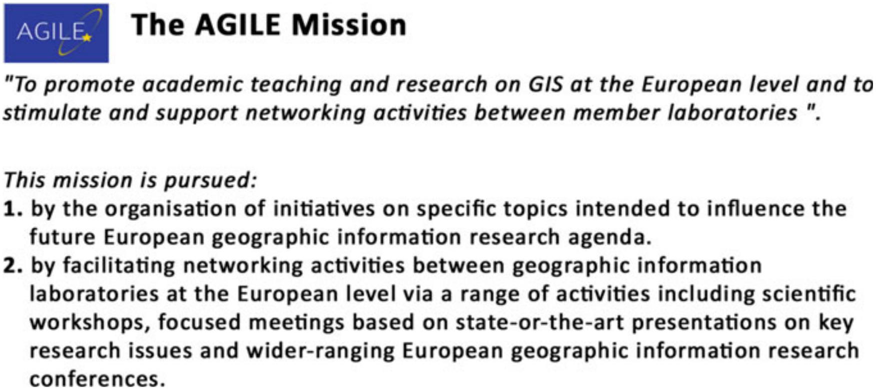

LECTURE NOTES IN GEOINFORMATION AND CARTOGRAPHY
LNG&C
S.I. Fabrikant · M. Wachowicz (Eds.)
The European Information Society
Leading the Way with Geo-information
Springer

Fig. 1 The first Springer book containing papers that underwent a blind peer review process with three reviewers for each paper



The AGILE Mission

"To promote academic teaching and research on GIS at the European level and to stimulate and support networking activities between member laboratories ".

This mission is pursued:

1. by the organisation of initiatives on specific topics intended to influence the future European geographic information research agenda.

2. by facilitating networking activities between geographic information laboratories at the European level via a range of activities including scientific workshops, focused meetings based on state-or-the-art presentations on key research issues and wider-ranging European geographic information research conferences.

Fig. 2 The AGILE mission statement

4 AGILE as an Organization

4.1 The Council

The association is steered by a Council of eight elected persons, each of whom works at an AGILE member institution. The Council meets twice a year and discusses strategic issues and makes decisions concerning all important aspects playing a role around AGILE, including the annual conference, the working groups and initiatives, the website, the research agenda, marketing activities, financial aspects, liaisons with other organizations, and others.

During the past 20 years, the membership of the Council has continually evolved. This is due to the rule that a Councilor is only allowed for a maximum of two periods of 4 years to be on the AGILE Council. The Council is composed of chair, secretary, treasurer, and five further members. It is supported by three external officers. One supports the secretary, one the treasurer and the third external officer manages the AGILE website. Barend Köbben from Twente University (ITC), the Netherlands, managed the website for a long time, followed by Christin Henzen from the Technical University of Dresden, Germany. Fred Toppen from the University of Utrecht (Netherlands) supports the AGILE treasurer. Since the beginning in 1998, Fred Toppen has been councilor (treasurer and secretary) and, afterwards, external officer of AGILE. In such a way he is the "pool of AGILE knowledge". Danny Vandenbroucke, supported by Ludo Engelen, both from the University of Leuven (Belgium), has served as AGILE secretary for many years as well. Maribel Yasmina Santos followed him, supported by Joao Galvao, both from the University of Minho, Portugal. Since 2015 Marek Baranowksi is secretary, supported by Aleksandra Furmankiewicz, both from the Institute of Geodesy and Cartography, Warsaw, Poland.

The Council is led by the Chair. Figure 3 provides an overview of the AGILE chairs from 1998–2017. Table 1 lists also the councilors who worked for AGILE during the years.

Councilors are responsible for a number of activities, next to regular duties such as administration and finances. These are described in Sect. 5. Other duties are external contacts (Sect. 4.3), but even more important are contacts with the members. For a long time AGILE chose to communicate with members by way of a newsletter, edited by Poulicos Prastacos (Greece), now the website is the main communication channel.

4.2 The Members

AGILE members are not individuals. Members are institutes, laboratories or departments in which the collection, administration, analysis and visualization of

Fig. 3 AGILE chairs 1998–2017

geospatial data is the main topic of interest. Around 85 of such institutions are members of AGILE.

One of the most challenging tasks of the work of the Council is to develop ideas on how to activate members to take part in the process of developing European GIScience. For instance, in 1999 it was decided to create AGILE working groups that would act as think tanks in different specific fields. The first working groups were on education and on environmental modeling. During several years, some of the working groups were quite active. They organized workshops at the annual conferences and in-between, they intensively discussed various topics, produced and presented ideas and new scientific approaches, and published special issues in well-known international journals.

Around 2010 the policy of the Council was changed and, instead of relatively “static” working groups, the idea of AGILE initiatives was born. This was thought as a more flexible and more productive scheme, giving AGILE members the chance to apply for some financial support for many possible ideas within the framework of geospatial information, carrying out workshops, exchanging ideas about new European projects, discussing specific themes that can arise at the research horizon, and many others. Still, every AGILE member can apply for support, only filling in and submitting a simple form (look at https://agile-online.org/index.php/initiatives). Working groups and initiatives often were the start of collaboration in EU projects. In some projects AGILE acted as partner (ETeMII, GI_N2K and GEOTHNK) but in most cases individual AGILE members joined EU projects often based on initial contacts at AGILE conferences.

The members are the most important “capital” of AGILE and the Council is still seeking for opportunities to convince all those laboratories that are concerned with spatial data processing to become members. One new and current measure,

Table 1 AGILE councilors 1998–2017

Councilor	Position	Period
Ian Masser	Chair	1998–2002
Werner Kuhn	Member	1998–2002
Anders Ostman	Treasurer	1998–2005
Fred Toppen	Treasurer/secretary	1998–2005
Tapani Sarjakoski	Member	1998–2000
Mauro Salvemini	Chair	1999–2006
Max Craglia	Member	1999–2003
Mike Gould	Chair	1999–2008
Marco Painho	Member	2000–2004
Juan Suarez	Member	2002–2009
Bela Markus	Member	2002–2006
Wolfgang Reinhardt	Treasurer	2003–2011
Monica Wachowicz	Chair	2004–2010
Martin Raubal	Chair/member	2005–2007, 2012–current
Danny Vandenbroucke	Secretary	2005–2013
Sara Fabrikant	Member	2006–2010
Irene Compte	Member	2006–2010
Lars Bernard	Chair/member	2007–2015
David Medyckyj-Scott	Member	2008–2010
Poulicos Prastacos	Member	2008–2012
Hardy Pundt	Treasurer	2009–2017
Mike Jackson	Chair	2010–2014
Bénédicte Bucher	Member	2010–2014
Maribel Yasmina Santos	Secretary	2011–2015
Joaquín Huerta	Member	2012–2015
Marinos Kavouras	Chair/member	2013–current
Tiina Sarjakoski	Member	2014–current
Alexis Comber	Member	2014–current
Marek Baranowski	Secretary	2015–current
Ali Mansourian	Member	2016–current
Mike Worboys	Member	2016–current

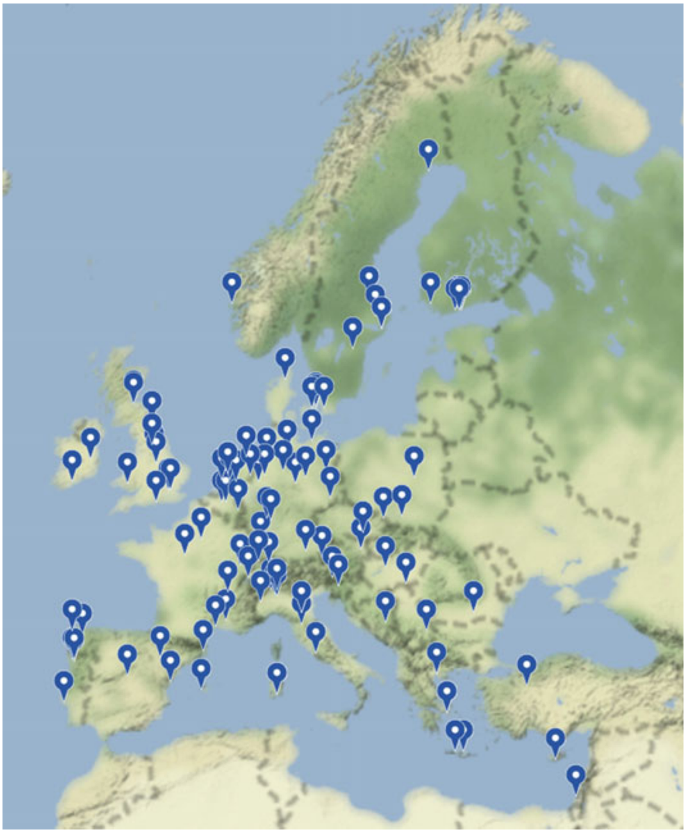


Fig. 4 Map of AGILE member laboratories

for instance, is to support financially the production of new ideas and work of early career researchers. They can apply for financial support within the framework of their specific project.

Figure 4 gives an overview of the AGILE member laboratories and the countries in which they are situated.

4.3 Links with Other Organizations

AGILE holds several linkages to other international organizations. These co-operations have been declared through common memoranda of understanding (MOU). Currently, relationships to a number of associations have been established, thus including MOUs with the main GI organizations, and a company (Esri), in the US and Australia (Fig. 5).

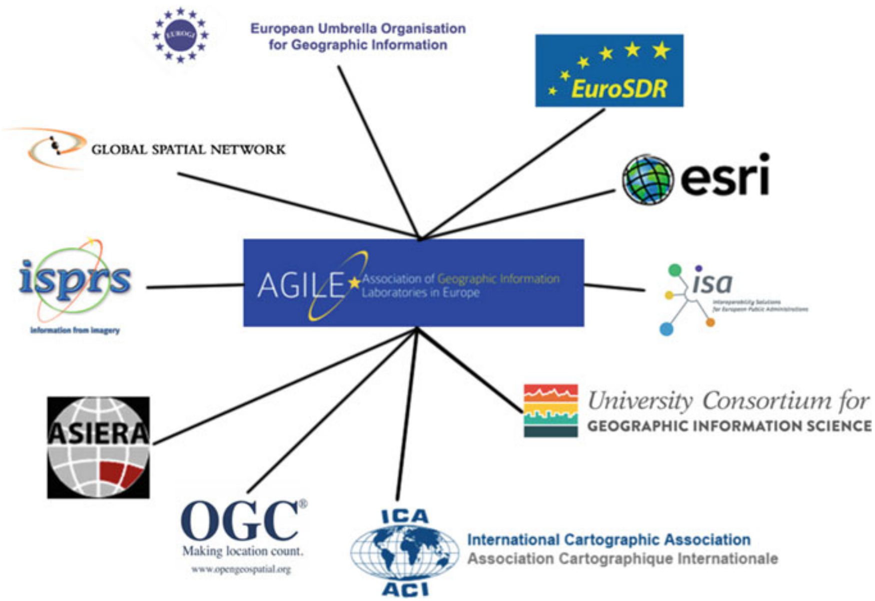

EUROGI
European Umbrella Organisation for Geographic Information
EuroSDR
GLOBAL SPATIAL NETWORK
esri
isprs
information from imagery
AGILE Association of Geographic Information Laboratories in Europe
isa
Interoperability Solutions for European Public Administrations
University Consortium for GEOGRAPHIC INFORMATION SCIENCE
ASIERA
OGC
Making location count.
www.opengeospatial.org
ICA
ACI
International Cartographic Association
Association Cartographique Internationale

Fig. 5 AGILE-links to partner organizations, manifested through MOUs

5 AGILE Activities

Without any doubt the AGILE conference is its primary activity. However, related activities such as working groups, initiatives, workshops, and EU projects support AGILEs goal setting of promoting GIScience and education. They serve also as a tool to meet, exchange and co-operate.

5.1 AGILE Conferences 1998–2017

Since 1998, 20 conferences have been carried out with a large number of workshops, presentations, keynotes and foremost uncountable personal talks between participants. Many initiatives, projects, common papers, special issues and ideas are outcomes of these events and form a significant contribution to the advancement of geographic information science and education in Europe. Figure 6 gives an overview of all conference venues.

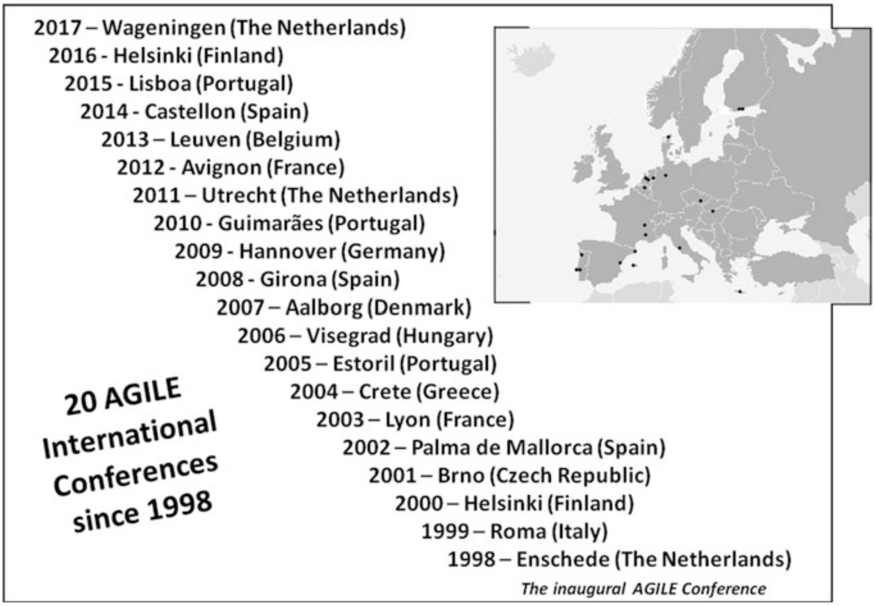

2017 – Wageningen (The Netherlands)
2016 - Helsinki (Finland)
2015 - Lisboa (Portugal)
2014 - Castellon (Spain)
2013 – Leuven (Belgium)
2012 - Avignon (France)
2011 – Utrecht (The Netherlands)
2010 - Guimarães (Portugal)
2009 - Hannover (Germany)
2008 - Girona (Spain)
2007 – Aalborg (Denmark)
2006 – Visegrad (Hungary)
2005 – Estoril (Portugal)
2004 – Crete (Greece)
2003 – Lyon (France)
2002 – Palma de Mallorca (Spain)
2001 – Brno (Czech Republic)
2000 – Helsinki (Finland)
1999 – Roma (Italy)
1998 – Enschede (The Netherlands)
The inaugural AGILE Conference
20 AGILE International Conferences since 1998

Fig. 6 AGILE conference venues 1998–2017

5.2 Working Groups, Initiatives and EU Projects

Working groups and initiatives were—and are—the two main streams thought to activate the members to contribute to the European research agenda, as well as developing European ideas and plans to promote and improve research and teaching in geographic information science, as is stated in the AGILE mission.

There were many significant outcomes of the working groups, as stated in Sect. 4.2. They were initiated, worked on specific subjects, and finished their work when members felt that goals were achieved. Table 2 summarizes working groups, initiatives and projects that were active or carried out during the last 20 years.

5.3 Workshops

Another indicator of the research carried out in AGILE laboratories is shown in the next table that lists workshops from 2009–2016. They were coordinated by AGILE members and attracted the attention of between 10 and 80 participants per workshop. The number of workshops responded to the increasing demand of AGILE members to discuss specific themes and to exchange ideas and results in a motivating environment. As a result, in recent years the first day of each conference is dedicated exclusively to the workshops.

Table 2 Examples of AGILE working groups, initiatives, and projects

Working groups Mostly active over several years	Long-term and short term initiative	Projects EU-funded
• Urban planning • Environmental modeling • Data Policy • Interoperability • Education	• GIS science publication rating • Crowdsourcing in international mapping • Persistent testbed • Body of knowledge I • Body of knowledge II • Semantics of geospatial information • Education mapping • Ph.D. schools (2012 in Wernigerode, Germany; 2013 at Lake Chiemsee, Germany; 2015 in Paris, France) • Access to spatial data for research and education • Ph.D. student survey • Research bursaries for Ph.D. project	• ETeMII • ESF–GI project • GI-N2 K • GEOTHNK

Some workshops acted as starting points for EU projects and other joint initiatives. The education-related workshops are a clear example of the need for members to discuss the progress in GI teaching. Already in the early years the UCGIS GI Body of Knowledge was an issue and several workshops were dedicated to the debate on to what extent a European version of this was needed. It is good to see that by way of the EU-projects GeoTHNK and GI-N2K, finally, a "European" approach has been realised (Table 3).

5.4 Research Themes

Since the early days of AGILE, members worked on the construction of a research agenda. A first initiative was to create a research map that gave access to information on the research activities of each individual AGILE member. Providing such information fostered co-operation, and was helpful in finding relevant partners for (EU) projects.

Another example was the "Green Paper for an Action-Oriented Research Agenda in the Geographic Information Science" (Craglia et al. 2001). One of the main reasons for this research agenda was the importance of the spatial dimension in national and EU policies. The green paper identified five priorities for such a research programme: GI Policy and Society; Theory of spatial-temporal information systems; Dynamic modeling of environmental and social processes; Semantic interoperability of spatial data and services; and Integration of social and physical sciences in their contribution to space. Browsing the titles of the sessions at AGILE conferences thus far, only the last topic seems to be somewhat neglected.

Table 3 Examples of workshops at AGILE conferences 2009–2016

Year	Workshop subject
2016	• Automated generalisation for on-demand mapping • CCM2 river and catchment database for Europe—applications • LinkVGI: Linking and analyzing volunteered geographic information (VGI) across different platforms • Code loves maps: cartographically oriented programming environments • 3rd AGILE workshop on geogames and geoplay • GIS with NoSQL • Visually-supported computational movement analysis • GI-N2K workshop: back-to-back @AGILE-2016
2015	• Geoprocessing on the web—science-driven and community-driven • RICH-VGI: enRICHment of volunteered geographic information (VGI): Techniques, practices and current state of knowledge • Geospatial Thinking: Research, educational and societal aspects • Assessing the fitness of citizens observatories for land cover/land use mapping and validation purposes • 2nd AGILE workshop an geogames and geoplay
2014	• Valarm-Esri ad hoc, real-time mobile sensor networks in the cloud • Digital earth: what the hack? • Geoprocessing on the web • Geogames and geoplay • Development augmented reality applications for google GLASS • COBWEB: citizen science, quality and standards
2013	• Analysing spatio-temporal data with R • Web cartography for national SDIs • Analysing eye-tracking data in real, virtual and mixed environments • Integrating 4D, GIS and cultural heritage • 3D urban modelling with Esri city engine • Action and Interaction in volunteered geographic information (ACTIVITY) • Understanding urban cycling: a data challenge (CDC2013) • The data complexity challenge—new approaches to data harmonisation
2012	• Testing geospatial web services—scientific SDIs • Creating campus applications using free Esri resources (CampusMaps) • 3D web visualization • Complex data mining in a geospatial context • Complexity modeling for urban structure and dynamics • Geographic information retrieval tutorial • Hands-on "open source GIS and WebMapping" • MECHANICITY and GeoDiverCity • Views on the body of knowledge (VoB)
2011	• Co-op strategy workshop/BoK2 • Cartographic support for early warning and crisis management • Higher-dimensional GIS, introducing PC raster 3 • Integrating sensor-web and web-based geoprocessing • Testbed research: testing geospatial and services/persisted testbed • Multi- and interdisciplinary research on spatial knowledge in the light of SII • Spatial thinking • GI research and instruction mixing commercial and open source tools

(continued)

Table 3 (continued)

Year	Workshop subject
2010	• Workshop on geospatial visual analytics: focus on time • Persistent testbed (PTB) workshop • Workshop on movement research: are you in the flow? • Workshop on GI-education mapping
2009	• Grid technologies for geospatial applications • The European qualification framework applicable to the GI domain? • Cross Atlantic workshop on economic value of geoinformation • GI@EarlyWarning • Adaptation in spatial communication • AGILE/EuroSDR/OGC persistent testbed for research and teaching in Europe • Challenges in geospatial data harmonisation

An overview (Table 4) was created of all the session titles from 1998 until 2016 in order to identify mainstream research issues throughout the years. Conclusions should be drawn with care. First, session names are not always representative of all the papers in such a session. Second, the analysis was based on the full programmes for the period 1998–2006 (part a) and on the chapters in the Springer proceedings for the period 2007–2016 (part b).

Themes such as dynamic modeling/spatial processes and spatial modeling/spatiotemporal analysis are persistent, also due to a whole range of subthemes that fit under this umbrella. Data infrastructures, GI policies and interoperability dominated the scene in the period 1998–2007. Also, ontology and semantics, and decision support systems are frequent themes for papers, at least in the period 2003–2016. Some themes come and go, such as urban modeling, GIS and planning and remote sensing, perhaps because the location of the conference activated some nearby members to participate. Other themes re-entered the scene but in a different context, such as visualization (from 2D to 3D). New technologies such as mobile devices and the web contributed to the "re-invention" of older themes (mobile GIS and transportation issues) or new themes (user generated data, discovering knowledge, web services).

6 AGILE—GO! Statements of the Chairs 1998–2017: Appreciating the 20th Anniversary of the Association

We are most grateful to all AGILE chairs for their willingness to contribute to this paper by expressing their feelings within a few lines. Their input provides a significant addition to the understanding of the role AGILE played over the last 20 years.

Table 4 Session themes of AGILE conferences 1998–2016

	a. Based on full conference programmes									b. Based on springer chapters									
Session themes	98	99	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16
Dynamic modelling spatial processes	X		X			X		X				X		X					
Data management	X							X											
Data infrastructures	X	X		X	X	X	X	X	X	X	X				X				
Interoperability	X		X	X	X	X		X	X	X									
Web based visualization		X																	
Data integration		X													X		X		
Environmental modelling		X		X	X		X	X		X									
Emerging technologies		X	X																
GIS and planning		X	X	X															
GT policies			X	X	X	X													
Mobile GIS and transportation			X	X				X	X	X				X	X		X	X	X
Data usability and quality			X	X	X	X	X	X		X					X				
GI databases			X						X										X
Spatial modelling and Analysis			X	X				X			X	X	X	X	X				X
Urban modelling					X	X	X			X					X				
Education					X	X	X	X	X	X									X
Public participation					X														
Location based services					X	X													
Economic issues				X	X		X												
Disaster management						X	X	X											
3D (modelling)						X		X	X	X									
Ontology, semantics						X	X	X	X	X	X		X			X		X	X
Decision support systems						X	X	X	X		X		X			X	X		
Visualisation							X		X			X	X	X	X		X		X
Remote sensing							X	X	X			X				X			X

(continued)

Table 4 (continued)

	a. Based on full conference programmes									b. Based on springer chapters									
Session themes	98	99	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16
Web services								X						X		X	X		
Spatial data processing														X		X	X		
User generated data																X	X	X	X
Discovering knowledge																		X	

Ian Masser

As convenor of the planning group that prepared the proposals for the initial creation of the Association of Geographical Information Laboratories in Europe during 1997 and the first chair of the AGILE Council that was elected at the First AGILE Conference in April 1998 I regard this as one of my most significant achievements of my career. It also gives me great pleasure to see that most of the general principles which were developed by the planning group still feature on the AGILE website and feel proud of the part I played in the initial development of what has become a very successful organisation over the last twenty years.

Mauro Salvemini

In the varied European panorama of sciences and techniques about geography and its applications AGILE provided, from the beginning, an efficient arena to include different approaches giving a large equal opportunity to European researchers looking for sharing a common umbrella. AGILE paved the road to INSPIRE and facilitated its application in member states. AGILE played a unique role to bridge two centuries being conceived since the foundation as one component of a larger discourse of governmental and NG organisations aiming to the European network of geographic information.

Mike Gould

AGILE leaders have made many decisions over the years, both good and bad, but all in an effort to best serve the GI science community and the wishes of the membership. Basing the association on laboratories instead of entire universities was key, as this is where the research happens. Eliminating specific ‘student sessions/papers’ and mixing students and faculty at the conference has proven to be a positive move. The active participation of students at AGILE conferences is very encouraging. The future is bright!

Monica Wachowicz

AGILE represents the success of an association in sharing of ideas, experience and challenges between academics, researchers, industry and wider communities. Its twenty years of existence have effectively demonstrated this exchange in advancing GI Science at national, European and international levels.

Mike Jackson

The ongoing success of AGILE and the pleasure of being part of it lies in its continuing focus on research rather than politics. Through this approach its Conference and core activities have remained stimulating, fresh and rewarding. It has helped integrate the community across age range, seniority, organisation and European country, avoided becoming too pompous and recognised that the most innovative research frequently comes from those at the early stages of their career.

Lars Bernard

AGILE always stayed agile and open: It always followed a lean but efficient management approach, being carried by an active AGILE community. AGILE succeeded in staying open and interested towards new disciplines, approaches, research cultures and practices. This attitude helped AGILE to not only stimulate and fertilize GI science but to also serve findings from GI Science into other communities.

Martin Raubal

AGILE has been and will continue to be a lively community for GI Scientists in Europe. From year to year the conferences demonstrate the diversity of this community—it includes students, senior researchers, administrators, business people, etc. who share a common interest in promoting GI research and teaching for the good of society. All the best to AGILE for the next 20 years and beyond!

Marinos Kavouras

During these 20 years, AGILE has bridged two eras - the relatively small community of those who shaped GI Science in the 80s and 90s and the next generation of scientists developed in the era of the mainstream geospatial technologies without borders. The provocative question: “What is so special about spatial?” has been replaced by the realization that “(almost) everything is spatial”. AGILE has played an important role in the progress of GI Science by putting emphasis on scientific substance and not on GI politics. It has maintained vigor by involving and supporting young researchers in all AGILE activities.

7 Final Remarks

With 20 years of conferences, workshops, joint project activities; AGILE has without doubt proved its “raison d’être”. After all those years, AGILE is still alive. New issues are now being explored, and more importantly, new generations join the yearly events. We are convinced that in the future AGILE will continue to play its role in the European GI arena. Council members are responsible for making sure that yearly events take place and new initiatives are stimulated. However, the backbone of the organization is its members. It is these colleagues involved in GI-research and education who feel the need to meet and co-operate. *They* keep AGILE alive.

Acknowledgements We would like to thank Martin Raubal, Mike Worboys, Tiina Sarjakoski and Frans Rip for their useful suggestions.

References

Craglia M, Gould M, Kuhn W, Toppen F (2001) The AGILE research agenda. In: EC-GIS workshop. European Commission, Potsdam, Germany. http://www.ec-gis.org/Workshops/7ec-gis/papers/html/agile/agile.htm

Masser I (1999) The association of geographic information laboratories in Europe (AGILE), Aims and scope. In: Proceedings of the 4th EC-GIS workshop. Joint Research Centre, Ispra, pp 235–239

The "Springer Series" on AGILE conferences (10 books). Lecture notes in geoinformation and cartography. Springer. ISSN 1836-2246, e-ISSN 1836-2351