

Lecture Notes  
in Geoinformation and Cartography

LNG&C

Tapani Sarjakoski  
Maribel Yasmina Santos  
L. Tiina Sarjakoski *Editors*

# Geospatial Data in a Changing World

Selected papers of the 19th AGILE  
Conference on Geographic Information  
Science

 Springer

# **Lecture Notes in Geoinformation and Cartography**

## **Series editors**

William Cartwright, Melbourne, Australia

Georg Gartner, Wien, Austria

Liqu Meng, Munich, Germany

Michael P. Peterson, Omaha, USA

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>

Tapani Sarjakoski · Maribel Yasmina Santos  
L. Tiina Sarjakoski  
Editors

# Geospatial Data in a Changing World

Selected Papers of the 19th AGILE  
Conference on Geographic Information  
Science



*Editors*

Tapani Sarjakoski  
National Land Survey of Finland  
Finnish Geospatial Research Institute  
Masala  
Finland

L. Tiina Sarjakoski  
National Land Survey of Finland  
Finnish Geospatial Research Institute  
Masala  
Finland

Maribel Yasmina Santos  
Department of Information Systems  
Universidade do Minho  
Guimarães  
Portugal

ISSN 1863-2246                      ISSN 1863-2351 (electronic)  
Lecture Notes in Geoinformation and Cartography  
ISBN 978-3-319-33782-1            ISBN 978-3-319-33783-8 (eBook)  
DOI 10.1007/978-3-319-33783-8

Library of Congress Control Number: 2016939356

© Springer International Publishing Switzerland 2016

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

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

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by Springer Nature  
The registered company is Springer International Publishing AG Switzerland

# Preface

Since 1998, the Association of Geographic Information Laboratories in Europe (AGILE) has promoted academic teaching and research on geographic information at the European level. Its annual conference reflects the variety of topics, disciplines, and actors in this research area. It provides a multidisciplinary forum for scientific knowledge production and dissemination and has gradually become the leading Geographic Information Science Conference in Europe.

For the tenth consecutive year, the AGILE Conference's full papers are published as a book by Springer-Verlag. This year, 48 documents were submitted as full papers, of which 23 were accepted for publication in this volume, after a thorough selection and review process. We congratulate the authors for the quality of their work, and thank them for their contribution to the success of the AGILE Conference and the book series. We also want to use this opportunity to acknowledge the numerous reviewers for providing us with their thorough judgments. Their work was fundamental to select the very best papers, and ultimately for the quality of this volume.

Under the title *Geospatial Data in a Changing World*, this book aims to envision the ways in which GIScience may contribute to the ever-growing need for geospatial data, when resolving the emerging challenges of our society, from climate change to people's mobility.

The scientific papers published in this volume cover a wide range of associated topics. The first part covers the challenges of cognitive and computational aspects on spatial concepts, giving the foundation to understand and perceive our environment in the future. The second part focuses on crowdsourcing and social networks reflecting the change in people's everyday relation and involvement with geospatial data. The third part gives foresights on how to analyze and visualize (big) spatial data to understand spatial phenomena. The fourth part covers pedestrian and vehicle mobility studies, emphasizing people's mobility and the requirements to develop more intelligent solutions for smart cities. The fifth, and the last part, gives insight related to the utilization of new data sources in information retrieval, modelling, and analysis.

Organizing a programme for an international conference and editing a volume of scientific papers takes time, effort and support. The input from the AGILE Council and Committees has been an important asset for us, and we are grateful to all members for their contributions.

We would also like to thank our sponsors for their kind contributions to the 19th AGILE Conference on Geographic Information Science and Springer-Verlag for their willingness to publish the accepted full papers in their academic series, Springer Lecture Notes in Geoinformation and Cartography.

Masala  
Guimarães  
Helsinki  
February 2016

Tapani Sarjakoski  
Maribel Yasmina Santos  
L. Tiina Sarjakoski

# **Organizing Committee**

## **Scientific Programme Committee Chairs**

### **Programme Chair**

Tapani Sarjakoski, Finnish Geospatial Research Institute, NLS, Finland

### **Programme Co-Chair**

Maribel Yasmina Santos, University of Minho, Guimarães, Portugal

L. Tiina Sarjakoski, National Land Survey of Finland, NLS, Finland

## **Local Organizing Committee**

L. Tiina Sarjakoski (Chair), National Land Survey of Finland, NLS, Finland

Heli Ursin, National Land Survey of Finland, NLS, Finland

Risto Kalliola, University of Turku, Finland

Kirsi Virrantaus, Aalto University, Finland

Tuuli Toivonen, University of Helsinki, Finland

Juha Oksanen (Workshop Chair), Finnish Geospatial Research Institute, NLS, Finland

## **Scientific Programme Committee**

Ana Paula Afonso, Universidade de Lisboa, Portugal

Jagannath Aryal, University of Tasmania, Australia

Fernando Bacao, Universidade Nova de Lisboa, Portugal

Marek Baranowski, Institute of Geodesy and Cartography, Poland

Itzhak Benenson, Tel Aviv University, Israel

Lars Bernard, TU Dresden, Germany

Michela Bertolotto, University College Dublin, Ireland  
Ralf Bill, Rostock University, Germany  
Sandro Bimonte, IRSTEA, France  
Thomas Blaschke, University of Salzburg, Austria  
Thomas Brinkhoff, Jade University Oldenburg, Germany  
Pedro Cabral, Universidade Nova de Lisboa, Portugal  
Sven Casteleyn, University Jaume I of Castellón, Spain  
Christophe Claramunt, Naval Academy Research Institute, France  
Serena Coetzee, University of Pretoria, South Africa  
Lex Comber, University of Leeds, UK  
Oscar Corcho, Universidad Politécnica de Madrid, Spain  
Joep Crompvoets, KU Leuven, Belgium  
Isabel Cruz, University of Illinois at Chicago, USA  
Cidalia Fonte, University of Coimbra, Portugal  
Anders Friis-Christensen, European Commission, Joint Research Centre, Italy  
Jerome Gensel, Université Grenoble Alpes, France  
Michael Gould, University Jaume I of Castellón, Spain  
Carlos Granell, University Jaume I of Castellón, Spain  
Henning Sten Hansen, Aalborg University, Denmark  
Lars Harrie, Lund University, Sweden  
Francis Harvey, Leibniz Institute for Regional Geography, Germany  
Roberto Henriques, Universidade Nova de Lisboa, Portugal  
Gerard Heuvelink, Wageningen University, The Netherlands  
Stephen Hirtle, University of Pittsburgh, USA  
Hartwig Hochmair, University of Florida, USA  
Joaquín Huerta, University Jaume I of Castellón, Spain  
Bashkim Idrizi, State University of Tetova, Republic of Macedonia  
Mike Jackson, University of Nottingham, UK  
Bin Jiang, University of Gävle, Sweden  
Didier Josselin, University of Avignon, France  
Risto Kalliola, University of Turku, Finland  
Derek Karssenbergh, Utrecht University, The Netherlands  
Tomi Kauppinen, Aalto University, Finland  
Marinos Kavouras, National Technical University of Athens, Greece  
Karen Kemp, University of Southern California, USA  
Dimitris Kotzinos, Université de Cergy-Pontoise, France  
Werner Kuhn, University of California Santa Barbara, USA  
Patrick Laube, Zurich University of Applied Science, Switzerland  
Robert Laurini, INSA, Lyon, France  
Victor Lobo, Universidade Nova de Lisboa, Spain  
Francisco J Lopez-Pellicer, Universidad Zaragoza, Spain  
Ali Mansourian, Lund University, Sweden  
Bruno Martins, Instituto Superior Técnico, Portugal  
Filipe Meneses, University of Minho, Portugal  
Peter Mooney, National University of Ireland Maynooth, Ireland

Adriano Moreira, University of Minho, Portugal  
João Moura Pires, Universidade Nova de Lisboa, Portugal  
Beniamino Murgante, University of Basilicata, Italy  
Javier Nogueras-Iso, Universidad Zaragoza, Spain  
Juha Oksanen, Finnish Geospatial Research Institute, NLS, Finland  
Toshihiro Osaragi, Tokyo Institute of Technology, Japan  
Frank Ostermann, University of Twente, The Netherlands  
Volker Paelke, Bremen University of Applied Sciences, Germany  
Marco Painho, Universidade Nova de Lisboa, Portugal  
Petter Pilesjö, Lund University, Sweden  
Poulicos Prastacos, FORTH, Greece  
Hardy Pundt, Harz University of Applied Sciences, Germany  
Ross Purves, University of Zurich, Switzerland  
Martin Raubal, ETH Zürich, Switzerland  
Wolfgang Reinhardt, Universität der Bundeswehr Muenchen, Germany  
Claus Rinner, Ryerson University, Canada  
Jorge Rocha, University of Minho, Portugal  
Armanda Rodrigues, Universidade Nova de Lisboa, Portugal  
Anne Ruas, IFSTTAR, France  
Maribel Yasmina Santos, University of Minho, Portugal  
Tapani Sarjakoski, Finnish Geospatial Research Institute, NLS, Finland  
L. Tiina Sarjakoski, National Land Survey of Finland, Finland  
Sven Schade, European Commission, Joint Research Centre, Italy  
Christoph Schlieder, University of Bamberg, Germany  
Monika Sester, Leibniz University Hannover, Germany  
Takeshi Shirabe, Royal Institute of Technology, Sweden  
Maguelonne Teisseire, IRSTEA, France  
Tuuli Toivonen, University of Helsinki, Finland  
Fred Toppen, Utrecht University, The Netherlands  
Nico Van de Weghe, Ghent University, Belgium  
Ron van Lammeren, Wageningen University, The Netherlands  
Jos Van Orshoven, KU Leuven, Belgium  
Danny Vandenbroucke, KU Leuven, Belgium  
Lluís Vicens, University of Girona, Spain  
Luis M. Vilches-Blázquez, Universidad Politécnica de Madrid, Spain  
Kirsi Virrantaus, Aalto University, Finland  
Monica Wachowicz, University of New Brunswick, Canada  
Robert Weibel, University of Zurich, Switzerland  
Stephan Winter, University of Melbourne, Australia  
Bisheng Yang, Wuhan University, China  
F. Javier Zarazaga-Soria, Universidad Zaragoza, Spain  
Alexander Zipf, Heidelberg University, Germany

# Contents

## **Part I Cognitive and Computational Aspects on Spatial Concepts**

<b>A Computational Model for Context and Spatial Concepts . . . . .</b>	<b>3</b>
Juergen Hahn, Paolo Fogliaroni, Andrew U. Frank and Gerhard Navratil	
<b>Inferring Complex Geographical Concepts with Implicit Geometries Using Ontologies: A Case of Peninsulas . . . . .</b>	<b>21</b>
Xiang Zhang, Tinghua Ai and Jantien Stoter	
<b>Question-Based Spatial Computing—A Case Study . . . . .</b>	<b>37</b>
Behzad Vahedi, Werner Kuhn and Andrea Ballatore	
<b>Measuring Space-Time Prism Similarity Through Temporal Profile Curves. . . . .</b>	<b>51</b>
Harvey J. Miller, Martin Raubal and Young Jaegal	
<b>Deriving the Geographic Footprint of Cognitive Regions . . . . .</b>	<b>67</b>
Heidelinde Hobel, Paolo Fogliaroni and Andrew U. Frank	

## **Part II Crowdsourcing and Social Networks**

<b>Android-Based Multi-Criteria Evaluation Approach for Enhancing Public Participation for a Wind Farm Site Selection. . . . .</b>	<b>87</b>
Pece V. Gorsevski and Alberto Manzano Torregrosa	
<b>Presenting Citizen Engagement Opportunities Online: The Relevancy of Spatial Visualization . . . . .</b>	<b>105</b>
Thore Fechner and Christian Kray	
<b>Spatial Data Relations as a Means to Enrich Species Observations from Crowdsourcing. . . . .</b>	<b>123</b>
Stefan Wiemann	

**Cross-Linkage Between Mapillary Street Level Photos and OSM Edits.** . . . . . 141  
Levente Juhász and Hartwig H. Hochmair

**Geo-Privacy Beyond Coordinates.** . . . . . 157  
Grant McKenzie, Krzysztof Janowicz and Dara Seidl

**Part III Analysis and Visualization of (Big) Spatial Data**

**Modelling Spatial Patterns of Outdoor Physical Activities Using Mobile Sports Tracking Application Data** . . . . . 179  
Rusne Sileryte, Pirouz Nourian and Stefan van der Spek

**Estimating the Biasing Effect of Behavioural Patterns on Mobile Fitness App Data by Density-Based Clustering.** . . . . . 199  
Cecilia Bergman and Juha Oksanen

**Enhancing Exploratory Analysis by Summarizing Spatiotemporal Events Across Multiple Levels of Detail.** . . . . . 219  
Ricardo Almeida Silva, João Moura Pires, Maribel Yasmina Santos and Nuno Datia

**Representation and Visualization of Imperfect Geohistorical Data About Natural Risks: A Qualitative Classification and Its Experimental Assessment** . . . . . 239  
Cécile Saint-Marc, Marlène Villanova-Oliver, Paule-Annick Davoine, Cicely Pams Capoccioni and Dorine Chenier

**Part IV Pedestrian and Vehicle Mobility in Smart Cities**

**Conflict in Pedestrian Networks.** . . . . . 261  
Jia Wang, Zena Wood and Mike Worboys

**Personalizing Walkability: A Concept for Pedestrian Needs Profiling Based on Movement Trajectories.** . . . . . 279  
David Jonietz

**Learning On-Street Parking Maps from Position Information of Parked Vehicles** . . . . . 297  
Fabian Bock, Jiaqi Liu and Monika Sester

**Visualizing Location Uncertainty on Mobile Devices: Assessing Users’ Perception and Preferences** . . . . . 315  
Champika Manel Ranasinghe and Christian Kray



**Part V Information Retrieval, Modelling and Analysis**

**Feature-Aware Surface Interpolation of Rooftops Using Low-Density Lidar Data for Photovoltaic Applications . . . . .** 337  
René Buffat

**Critical Situation Monitoring at Large Scale Events from Airborne Video Based Crowd Dynamics Analysis . . . . .** 351  
Alexander Almer, Roland Perko, Helmut Schrom-Feiertag,  
Thomas Schnabel and Lucas Paletta

**Probabilistic Framework for Modelling the Evolution of Geomorphic Features in 10,000-Year Time Scale: The Eurajoki River Case . . . . .** 369  
Jari Pohjola, Jari Turunen, Tarmo Lipping and Ari T.K. Ikonen

**GeoPipes Using GeoMQTT . . . . .** 383  
Stefan Herle and Jörg Blankenbach

**Continuous Generalization of Administrative Boundaries Based on Compatible Triangulations . . . . .** 399  
Dongliang Peng, Alexander Wolff and Jan-Henrik Haurert

**Part I**  
**Cognitive and Computational Aspects on**  
**Spatial Concepts**

# A Computational Model for Context and Spatial Concepts

Juergen Hahn, Paolo Fogliaroni, Andrew U. Frank  
and Gerhard Navratil

**Abstract** A natural language interface can improve human-computer interaction with Geographic Information Systems (GIS). A prerequisite for this is the mapping of natural language expressions onto spatial queries. Previous mapping approaches, using, for example, fuzzy sets, failed because of the flexible and context-dependent use of spatial terms. Context changes the interpretation drastically. For example, the spatial relation “near” can be mapped onto distances ranging anywhere from kilometers to centimeters. We present a context-enriched semiotic triangle that allows us to distinguish between multiple interpretations. As formalization we introduce the notation of contextualized concepts that is tied to one context. One concept inherits multiple contextualized concepts such that multiple interpretations can be distinguished. The interpretation for one contextualized concept corresponds to the intention of the spatial term, and is used as input for a spatial query. To demonstrate our computational model, a next generation GIS is envisioned that maps the spatial relation “near” to spatial queries differently according to the influencing context.

**Keywords** Computational model for context · Context · Spatial concepts · Contextualized concept · Near

---

J. Hahn (✉) · P. Fogliaroni · A.U. Frank · G. Navratil  
Research Group Geoinformation, Technische Universität Wien, Vienna, Austria  
e-mail: hahn@geoinfo.tuwien.ac.at

P. Fogliaroni  
e-mail: paolo@geoinfo.tuwien.ac.at

A.U. Frank  
e-mail: frank@geoinfo.tuwien.ac.at

G. Navratil  
e-mail: navratil@geoinfo.tuwien.ac.at

# 1 Introduction

A fundamental question in the field of Geographic Information Science concerns the development of a natural language interface for GIS (Montello and Freundschuh 2005). In order to establish a natural language interface for GIS, spatial terms (i.e. spatial relations and spatial regions (Montello et al. 2003) have to be mapped onto spatial queries. In this paper we address the mapping of spatial relations. Previous approaches to model spatial terms concluded that their interpretation is mostly context-dependent (Wang 1994; Raubal and Winter 2002; Yao and Thill 2006). For example, the spatial relation “near” can be mapped onto distances ranging from thousands of kilometers (e.g. “the moon is near the Earth”) to a few centimeters (e.g. “the cup is near the milk bottle”).

A central question is always: what is context? In the scope of this work we consider context to be any piece of ancillary or surrounding information that influences the interpretation of a concept of interest. The semiotic triangle (Ogden and Richards 1946) (reviewed in Sect. 2) explains the process of interpretation in a triadic mode, including an object in reality, a concept formed by a cognitive agent, and a term. We introduce an enriched version of the triangle that also includes context, and show how such a modification allows for disambiguating the interpretation of spatial concepts.

A cognitive agent refers to objects in reality by externalizing a context-influenced concept. A concept is, by its very nature, an abstract entity that only exists in the human mind. It therefore cannot be measured in terms of, or categorized by, physical properties. Concepts have been proven (Rosch and Mervis 1975) to be fuzzy, and to include prototypes. This also holds true for spatial concepts, e.g. **downtown**<sup>1</sup> (Montello et al. 2003), **north south** (Montello et al. 2014), **near** (Fisher and Orf 1991; Wang 1994). Prototypes change with the influence of context (Osherson 1999; Aerts and Gabora 2005). For example, a prototypical example for the concept **tree** is different in Sweden and in Greece. We argue that the interpretation of a spatial term relates to the prototype of a concept. To account for the possibility that a concept can inherit multiple prototypes we introduce the notion of *contextualized concepts*. One concept is represented by many contextualized concepts, where each contextualized concept has one prototype and is linked to one context. A contextualized concept is built from grounded observations of reality (Kuhn 2009) observed in a particular context.

Many possible interpretations are narrowed down to a single one by making context explicit for concepts and observations. This resulting interpretation is used as mapping from a spatial term onto a spatial query.

---

<sup>1</sup>Throughout the paper we will use special formatting to indicate when a term is used to denote a **concept**.

In summary the contributions of this paper are:

- enrichment of the semiotic triangle with context
- derivation of an abstract model
- formalization of the abstract model as a computational model

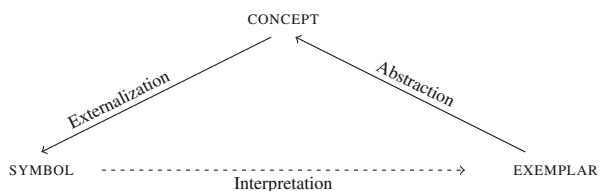
This paper is structured as follows: First, the semiotic triangle is reviewed to show how representations (terms) are connected with concepts that refer to observations in reality in Sect. 2. Next the properties for concepts are pointed out, and we emphasize the influence context has on concepts. Second, the idea and a formalization of it are presented in Sect. 3. Third, this formalization is translated into a computational model in Sect. 4. Fourth, the model is initialized with data, and the usage of the algorithms is demonstrated in mapping the spatial relation “near” according to contexts: walking, driving, and going uphill are mapped into different spatial queries in Sect. 5.

## 2 State of the Art

To achieve mapping from spatial relations or general spatial terms to spatial queries, it is necessary to understand how spatial terms represent reality. Spatial relations are symbols that refer to spatial configurations, such as near or above. The semiotic triangle by Ogden and Richards (1946) is a conceptual model that links symbols (e.g. a word, a drawing, a map, or a gesture), reality, and concepts (see Fig. 1). Each edge of this triangle represents one of the three main phases of representation: abstraction, externalization, and interpretation. On the right corner of the triangle lies physical reality. Physical reality is experienced in the form of exemplars, and abstracted to form a concept in our mind (represented by the top corner of the triangle). When we want to externalize a concept, for example during a communication process, we use a symbol. Symbols are located on the left corner of the triangle, and are successfully interpreted (as are the corresponding concepts) if they are *grounded* (Kuhn 2009) to the subsets of physical reality (the exemplars) that we intended to refer to (indicated by the dotted line).

Possible misunderstandings or misinterpretations arise if the same symbol does not evoke the same typical exemplar in different subjects. One reason is the many-to-many connection between a symbol and the exemplars that it refers to Chandler (2007). For example, in mathematics there is the concept of **neutral element** for a

**Fig. 1** Semiotic triangle from Ogden and Richards (1946)



binary operation. Without specifying the operation (e.g. addition, multiplication) it is not clear which exemplar it refers to (e.g. 0, 1). Another reason is that concepts are formed and adjusted over time from repeated observations of reality (Von Glasersfeld 1995). Since no two persons have identical experiences, the “same” concept can never be completely aligned in two person’s minds.

It was already suggested in the past (Von Glasersfeld 1995; Fisher 2000; Worboys 2003) that context plays a role in aligning concepts. Fisher (2000) studied the case of directional concepts and suggested that one way to avoid (or reduce) misinterpretations is to make explicit the frame of reference (context) that the given directional concepts are embedded in. More generally, Von Glasersfeld (1995) states that successful interpretation is only possible if the context of the speaker and that of the listener are compatible; which means that the speaker and the listener must have experienced exemplars of a concept in “similar” contexts (cf. Weiser and Frank 2013).

## 2.1 (Spatial) Concepts

A concept is an abstract entity that only exists in the human mind; according to Seiler (2001), it is “*primarily a cognitive structure*” that helps us to make sense of the world. The entities from which a concept is derived are called, throughout this work, *instances* or *exemplars*.

According to Freksa and Barkowsky (1996), spatial concepts are all those “*notions that describe spatial aspects of a subset of the world*”. Examples of spatial concepts are **near**, **downtown**, and **lake**. Spatial concepts are central to human cognition (Mark et al. 1999) as they help us to distinguish, categorize, and thus make sense of the physical stimuli we perceive through our senses.

Psychological experiments showed that concepts include prototypes. By using a category–membership verification technique, cognitive psychologist Rosch (1973, 1999) showed that concepts possess a graded structure. Within this structure, a prototype is abstracted from the experienced exemplars (Rosch and Mervis 1975) based on a typicality judgment function. Another modeling approach represents a concept as multiple experienced exemplars (Nosofsky 2011). Both theories share that the membership of an exemplar to a concept is judged within their typicality to the existing concept.

In the field of geographic information science, several studies have previously aimed at characterizing (geo)spatial concepts. Mark et al. (1999) empirically demonstrated that people judge mountains, lakes, and oceans as typical exemplars of the generic concept **geographical feature**. Further studies (Mark et al. 1999) revealed that spatial concepts are typically organized according to a hierarchical structure, and have vague boundaries. For example, it was shown by Smith and Mark (1998) that geographical factors like size or scale induce conceptual hierarchies—as in the case of bodies of water: pond, lake, sea, ocean. Also, it has been shown (Mark and Turk 2003; Mark 1993) that linguistic, cultural, and individual variability influences the

creation and the structuring of spatial concepts. Montello et al. (2014, 2003) investigated the fuzziness of the extension of spatial concepts. They showed that the spatial concept **downtown** Santa Barbara is conceived differently by different subjects.

In the field of geoinformation, multiple approaches have been used to model spatial concepts, such as, for example: qualitative spatial reasoning (Frank 1992), fuzzy sets (Wang 1994; Robinson 2000), multi-valued logic (Fisher 2000; Duckham and Worboys 2001; Worboys 2001), formal concept analysis (Frank 2006), and more. All these formalizations do not account for context explicitly, and scientists concluded that context has a major impact. In contrast, we use context as the base formal drive to determine the interpretation.

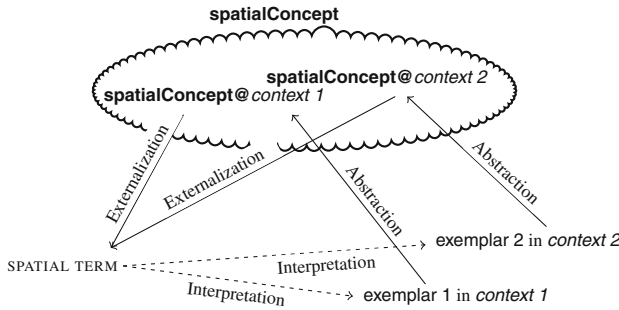
## 2.2 Context and Its Influence on Concepts

According to Kuhn (2005), “*context is an overloaded term and has many aspects. Some of them are relatively easy to handle through domain separation [...]. Others are much harder to deal with [...]*”. Bazire and Brézillon (2005) analyzed 150 different definitions of context collected on the Web. They find that, although different, they all share some common structure, and conclude that the definition of context is highly domain-dependent. This is also true for the spatial domain (Huang et al. 2014).

According to Freksa and Barkowsky (1996), there are three main types of relations that determine the meaning of a concept: (i) relations between a concept and its exemplars, (ii) relations between concept and context, and (iii) relations between concept exemplars and context.

Like any other type of concept, spatial concepts are influenced by context. Several studies have been carried out to study the context–concept influence. Burgio et al. (2010) and Tversky (2003) investigated the influence of context on spatial terminology. Both studies show that context influences spatial descriptions at the level of scale and granularity. Egenhofer and Mark (1995) investigated how different contexts influence the concept **geographic space**. They found, for example, that in a “city” context the interpretation evokes typical exemplars such as streets, buildings, and parks, while in a “country” context these become mountains, lakes, and rivers. Talmy (2003, p.231) argued that the spatial relations “on” and “in” are used for vehicles differently, depending on the existence of a walkway in the vehicle—e.g. on a bus versus in a car. Smith and Mark (1998) showed that the relation “in” in the context “the island is in the lake” means *the island protrudes from the surface of the lake* while in the context “the submarine is in the lake” the interpretation is *the submarine is completely submerged within the corresponding three-dimensional volume*.

Aerts and Gabora (2005) presented a quantum-mechanical model for concepts and influencing contexts. Their model showed that context is also the driving factor in modeling concept combination.



**Fig. 2** Semiotic triangle from Ogden and Richards (1946) used for geographic information science by Kuhn (2005), enriched with context. The exemplars of the same concept, observed in different contexts. If the context is not explicitly reported with symbolic externalization, the symbol can be misinterpreted (many-to-many relation). Misinterpretations vanish if the context is specified, because a one-to-one connection between an exemplar and the symbol is created

### 3 An Abstract Model for Context-Dependent Concepts

In the scope of this work we consider context to be any piece of ancillary or surrounding information that influences the interpretation of a concept of interest. This means that the same concept is possibly associated to different typical exemplars in different contexts.

The core idea is to establish a one-to-one connection between a symbol (with ambiguous semantics) and an observed exemplar based on the context. Context selects from the many interpretations of the symbol a single applicable one—i.e. it reduces a many-to-many relation to a simple one-to-one. This idea is schematized in Fig. 2, which describes the process of abstracting one **spatialConcept**<sup>2</sup> from two experiences (exemplars) observed in two different contexts: *context 1* and *context 2*. Exemplar 1 is experienced in *context 1*, while exemplar 2 is experienced in *context 2*. This generates for the given concept what we call *contextualized concepts*, denoted by **spatialConcept@context 1** and **spatialConcept@context 2**, respectively. Externalizing **spatialConcept** without also giving context does not allow for a definitive interpretation, as the symbol used can refer to many of the exemplars we have experienced. If, conversely, we clearly state that the spatial term is in a particular context, the ambiguity vanishes, and it becomes clear that we intend either exemplar 1 or exemplar 2.

Through the use of contextualized concepts, context structures observed exemplars. The use of contextualized concepts falls into the class of “*compose-and-conquer*” of context uses (Bouquet et al. 2003). This compose-and-conquer approach “takes a context to be a theory of the world that encodes an agent’s perspective of it and that is used during a given reasoning process” (Akman and Surav 1996). Every

<sup>2</sup>In order to remove ambiguity we use special formatting to indicate a *context*, an *exemplar* of a concept, or a concept in a specific context (denoted **concept@context**).



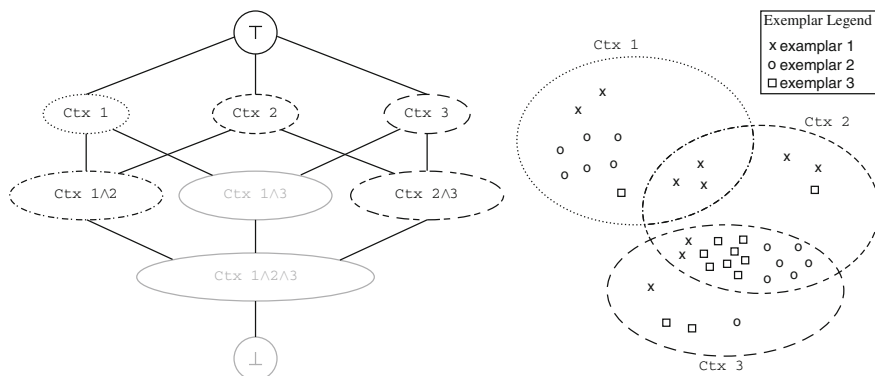
context partitions the mental contents, which is similar to Fauconnier’s idea of mental spaces (Fauconnier 1994).

In order to formalize the idea sketched in Fig. 2 we model context with a lattice structure. In general, a lattice is a partially ordered set under a partial order relation with two binary functions:  $\wedge$  called *meet* and  $\vee$  called *join* (Gratzer 2009). Given two elements  $a$  and  $b$  of the lattice, the function *meet* creates an *infimum* such that  $a \wedge b = \text{inf}\{a, b\}$  meaning that there exists a greatest element that is lower or equal to  $a$  and  $b$ . The function *join* creates a *supremum* of two elements:  $a \vee b = \text{sup}\{a, b\}$  meaning that there exists a least element that is bigger or equal to both elements. A bounded lattice is a lattice with an upper bound element (i.e.  $\top$ ) and a lower bound (i.e.  $\perp$ ) element.

The partial order relation “*is stronger than or equal to*”, denoted  $\leq$ , applies for context. An example for a context lattice is shown in Fig. 3. The  $\top$  element is called *universal context*, meaning the absence of context. The contexts “stronger” than the  $\top$  element are called *basic contexts* (e.g. *ctx 1*, *ctx 2*, and *ctx 3* in Fig. 3) and these are used to derive through the *meet* operation any other context combination (e.g. *ctx 1*  $\wedge$  *2*). The last element (“strongest” context) of the lattice is the  $\perp$  element which is called *empty context* and indicates nonsense—i.e. meaningless context.

Note that not all the infima in the lattice of contexts correspond to contexts that make sense, or that are realizable. In Fig. 3, these contexts are represented in grey. One is *ctx 1*  $\wedge$  *3*, and, consequently, every infimum of this context does not make sense. In Fig. 3 there is only one such infimum: *ctx 1*  $\wedge$  *2*  $\wedge$  *3*. An example for two contexts that do not make sense is included in the example presented in Sect. 5.

The number of contexts (including top and bottom elements) in the lattice obeys the rule  $2^n + 1$ , where  $n$  is the number of basic contexts. Let us look at the lattice as consisting of  $n + 2$  levels: level 0 corresponds to the top element and level  $n + 1$  corresponds to the bottom element. Level  $l$  comprises the lattice elements



**Fig. 3** Relation between observations of exemplars in reality (on the right) and contexts (on the left) for one concept. Contexts are organized in a lattice where infimum contexts corresponds to intersections of the former contexts. Some infima can be impossible in reality, which results in an empty mapping  $\lambda$ . Impossible contexts (*ctx 1*  $\wedge$  *3* and *ctx 1*  $\wedge$  *2*  $\wedge$  *3*) are reported in grey

corresponding to 1-combinations of the base contexts. Then, the number of elements at level  $l$  is equal to  $\binom{n}{l}$  and the total number of elements is  $\sum_{l=1}^n \binom{n}{l} = 2^n - 1$ . This is the number of all possible combinations except the void ones. Counting top and bottom elements as well, we obtain the formula:  $2^n + 1$ , which is in the order of  $O(2^n)$ .

Contexts are linked to concepts via contextualized concepts. One context links to one contextualized concept, which is composed of a set of exemplars. The selection of the exact subset of exemplars is achieved by the mapping  $\lambda$  (Aerts and Gabora 2005). In Fig. 3 the mapping  $\lambda$  is represented by styling the borders of contexts in the lattice and subsets of exemplars for one concept in the same way. The mapping  $\lambda$  can be used to represent a concept in a tabular form. We call this table *observation table* because the exemplars are observed in reality. The connection to reality guarantees that other agents can make the same observations grounded in reality (Kuhn 2009). The columns of this table denote contextualized concepts, the rows denote exemplars. The entries indicate how many times a given exemplar has been observed in a given context. For example, Table 1 represents an observation table for the **spatialConcept** abstracted from observations shown in Fig. 3.

Frequency values for each exemplar in the contexts  $ctx\ 1 \wedge ctx\ 3$  and  $ctx\ 1 \wedge ctx\ 2 \wedge ctx\ 3$  are zero. A zero frequency value reflects that no exemplar was observed. This can occur either in the case of a meaningless context, or if there has been no observation yet. The model does not distinguish between meaningless contexts and not-yet-experienced contexts. It resembles what can also be found in child learning processes (Twaroch and Frank 2005).

The observation frequencies from the observation table are used to calculate the prototypical exemplar for a contextualized concept. As a typicality measure for exemplars, the amount of observations per exemplar is used. The exemplar with the most observations is considered the prototypical exemplar for the contextualized concept. Depending on the context, different typical exemplars can be calculated. For example, consider the data from Table 1, the typical exemplar for the contextualized concept **spatialConcept**@ $\top$  is exemplar 2, and for the **spatialConcept**@*context 3* it is exemplar 3.

The prototypical exemplar of a contextualized concept is used as a mapping from a spatial term onto a spatial query. By making the context explicit, a one-to-one

**Table 1** Observation table for a **spatialConcept** for different contexts according to the example depicted in Fig. 3

<b>Spatial concept</b>	@ $\top$	@ <i>ctx 1</i>	@ <i>ctx 2</i>	@ <i>ctx 3</i>	@ <i>ctx 1</i> $\wedge$ <i>ctx 2</i>	@ <i>ctx 2</i> $\wedge$ <i>ctx 3</i>	@ <i>ctx 1</i> $\wedge$ <i>ctx 3</i>	@ <i>ctx 1</i> $\wedge$ <i>ctx 2</i> $\wedge$ <i>ctx 3</i>
Exemplar 1	10	5	7	3	3	2	0	0
Exemplar 2	13	6	6	7	0	6	0	0
Exemplar 3	12	1	9	10	0	8	0	0

Values indicate how many times an exemplar of a **spatialConcept** (in the rows) has been observed in different contexts (in the columns)

relation between grounded observations and the spatial term is achieved. Having a one-to-one relation, the observations experienced in the same context are selected and used to calculate the prototypical exemplar. The prototypical exemplar is then used as input for a spatial query.

## 4 A Computational Model for Context-Dependent Concepts

The formalization of the previous approach results in a computational model. The implementation includes three parts: the context lattice, the mapping ( $\lambda$ ) from contexts to observations, and the calculation of the prototypical exemplar. The necessary operations and data structures are described with pseudocode and can be implemented in a variety of programming paradigms (e.g. object-oriented, relational algebras, functional). Our implementation using a functional paradigm can be downloaded here: <https://hackage.haskell.org/package/ContextAlgebra>.

Context is implemented as a list of elements:

$$\text{context} : [\text{contextName}_1, \text{contextName}_2, \dots, \text{contextName}_n]$$

An element is an arbitrary data type that supports equality comparison, for simplicity assume that these are names (e.g. character or string). *Basic contexts* include one entry in the list (e.g. [*context 1*]), while *infima contexts* include multiple entries (e.g. [*context 1, context 2*]). The context lattice is implemented as a container for all contexts as well as the empty and universal contexts. The lattice operations MEET<sup>3</sup> and JOIN are implemented as an intersection and union of lists.

A contextualized concept is implemented as a multiset<sup>4</sup> of observations that map to a context in the context lattice. The *observation* data type is realized as a pair consisting of the observed exemplar and a context:

$$\begin{aligned} \text{exemplar} &: \text{exemplarName} \\ \text{observation} &: (\text{exemplar}, \text{context}) \end{aligned}$$

The context is built with the same structure and types as the contexts included in the context lattice which provides the mapping  $\lambda$ . A particular contextualized concept of interest is the **spatialConcept@T** which includes all observations for all contexts. All other contextualized concepts refer to a subset of observations.

The calculation of the prototypical exemplar for a contextualized concept is achieved by the functions: FILTER and COMPUTETYPICALITY.

---

<sup>3</sup>Algorithms are indicated with a small caps typeface.

<sup>4</sup>The multiset is capable of holding the same entry multiple times, in contrast to a set.

The function `FILTER` takes a context ( $ctx$ ) as input parameter and returns a contextualized concept (**spatialConcept@ $ctx$** ). All the observations listed in the **spatialConcept@ $\top$**  are checked, and if one is found whose context coincides with the filter context it is added to **spatialConcept@ $ctx$** . This function relies on the equality operator and on a function `CONTEXT` that returns the context of an observation given in input. All contextualized concepts can possibly be stored for ease of accessibility.

---

**Algorithm 1** Given an exemplar and a contextualized concept, the function `COMPUTETYPICALITY` computes the typicality of the exemplar in the context associated to the contextualized concept.

---

```

1: function COMPUTETYPICALITY(exemplar, spatialConcept@context)
2:   obsForExemplar  $\leftarrow \emptyset$ 
3:   for  $\forall$  observation  $\in$  spatialConcept@context do
4:     if exemplar == EXEMPLAR(observation) then
5:       obsForExemplar  $\leftarrow$  obsForExemplar  $\cup$  observation
6:   return AMOUNT(obsForExemplar) / AMOUNT(spatialConcept@context)

```

---

The function `COMPUTETYPICALITY` takes an exemplar and a contextualized concept as input parameters, and returns the typicality of the given exemplar for the context corresponding to the contextualized concept in input. This is called *contextual typicality* and takes values in the range  $[0, 1]$ . It is computed by counting the number of exemplars equal to the one given, and by dividing this number by the number of elements in the contextualized concept. This function relies on the equality operator for exemplars (denoted  $==$ ), the `AMOUNT()` function to enumerate exemplars, and on the function `EXEMPLAR()` returning the exemplar of an observation given in input. `COMPUTETYPICALITY` is further used to calculate the prototypical exemplar of a contextualized concept.

## 5 Case Study: Mapping the Spatial Relation “near” onto Spatial Queries

There exist many scenarios where context plays an important role, and to which the presented model could therefore be applied. Some of these include:

- Detection of landmarks has to be context-aware because landmarks depend on context (e.g. night or day (Winter et al. 2005)).
- Different map layers can be displayed depending on context, e.g. the request “I need a map to find the pub” intends a city street map, in contrast to the request “I need a map for hiking” where hiking paths should be included.
- Interpreting spatial relations has to make use of context for mapping to a metric distance.

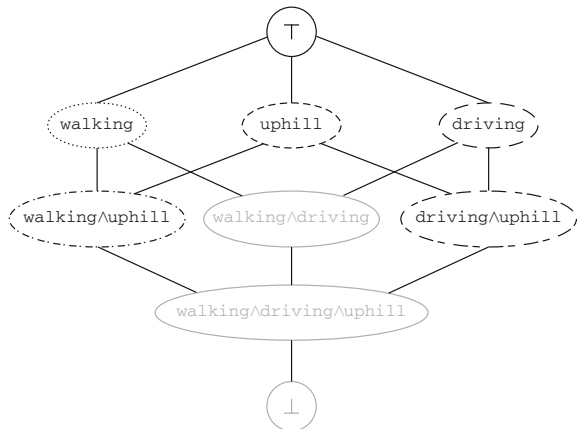
In the following section we apply the model as a case study of the spatial relation “near”, and show how such a concept can be encoded in spatial queries according to context. The aim is to show how “near” is mapped onto different metric distances according to the influencing context.

### 5.1 The Case Study of “near”

In general, spatial terms are influenced by many contexts, e.g. weather, mobility, time of the day, neighborhood, transportation mode, terrain. In this case study the reasonable set of exemplary basic contexts are the following: *walking*, *driving*, and *uphill*. These contexts can be derived by information obtained from sensors commonly available in modern smartphones. For example, the difference between *walking* and *driving* can be derived from speed data computable either from GNSS or accelerometers, while the *uphill* context can be derived by matching position and elevation information. The whole lattice of contexts that can possibly influence the interpretation of **near** is generated by recursively executing the function MEET on the available contexts—shown in Fig. 4. Note that the context *walking*  $\wedge$  *driving* obtained by combining the basic contexts *walking* and *driving* is not realizable, as one cannot drive and walk at the same time. This is indicated by the grey border. Accordingly, any infima deriving from this context (in this example *walking*  $\wedge$  *driving*  $\wedge$  *uphill* and  $\perp$ ) are also not realizable.

The corresponding observation table is given in Table 2. The reported observations are an educated guess driven by common sense and by the results presented by Wallgrün et al (2014), who evaluated the interpretation of **near** in a corpus of web documents. How observations can be reliably collected is not the focus of this work, however in Sect. 6 we outline some possibilities that could lead to future work.

**Fig. 4** Bounded context lattice for the case study of “near”. Grey boundaries indicate impossible contexts

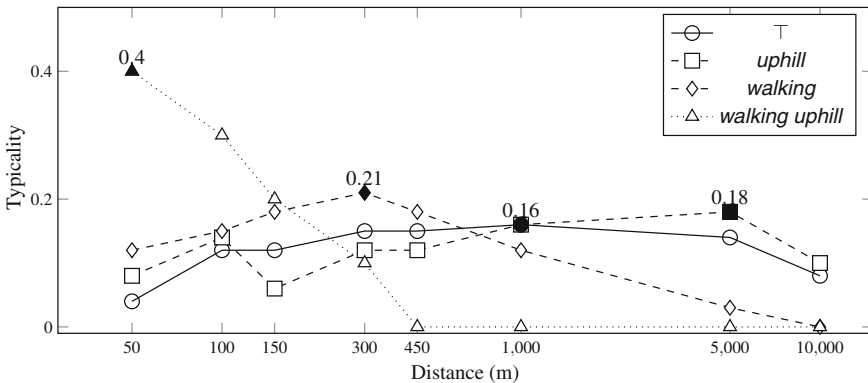


**Table 2** Observation table for the concept **near** for different contexts

<b>near</b> (m)	@ $\top$	@ <i>walking</i>	@ <i>uphill</i>	@ <i>driving</i>	@ <i>walking</i> $\wedge$ <i>uphill</i>	@ <i>driving</i> $\wedge$ <i>uphill</i>	@ <i>walking</i> $\wedge$ <i>driving</i>	@ <i>walking</i> $\wedge$ <i>driving</i> $\wedge$ <i>uphill</i>
50	4	4	4	0	4	0	0	0
100	10	5	7	3	3	2	0	0
150	10	6	3	4	2	1	0	0
300	13	7	6	6	1	5	0	0
450	13	6	6	7	0	6	0	0
1000	14	4	8	9	0	7	0	0
5000	12	1	9	10	0	8	0	0
10000	7	0	5	7	0	5	0	0
<b>Sum</b>	83	33	48	46	10	34	0	0

The last row shows the sums of observations in different contexts, and is provided to compute exemplar typicality

Given the context lattice and the observation table, typicality values and prototypical exemplars for the contextualized concepts are generated through the functions FILTER and COMPUTETYPICALITY (introduced in Sect. 4) as follows. The observation sets for each contextualized concept (**near@ctx**) are obtained by executing the function FILTER(*ctx*) for each context *ctx* in the lattice. The contextual typicality values are then computed by executing the function COMPUTETYPICALITY(**e**, **near@ctx**) for each *ctx* in the lattice and each observed exemplar **e** in **near@ $\top$** . The exemplar with the highest contextual typicality is selected as the prototypical exemplar. Typicality values for the contextualized concepts we will be using in our example are plotted in Fig. 5, where the prototypical exemplars are colored in black.



**Fig. 5** Contextual typicalities for the concept **near** in different contexts; different *line* styles denote different context as reported in the legend. Points with filled markers indicate prototypical exemplars, and thus the interpretation for the concept in a given context

## 5.2 Examples for the Mapping of “near” onto Spatial Queries

Imagine a next-generation personalized geographic information system (PersonalizedGIS (Abdalla et al. 2013)) installed on a user’s smartphone. The GIS part has access to classical geographic information (particularly, points of interest and elevation data). The personalized part is an implementation of the computational model presented in Sect. 4. Also, imagine that the system has seen the same situations as its user, in the sense that the observations given in Table 2 match with a high level of precision the concept **near** in the user’s mind.

The user attends a conference in Lisbon (Portugal) and needs to find a restaurant near his or her current position. The user asks the PersonalizedGIS: “*Please show all the restaurants near me*”. Imagine a natural language processing algorithm that extracts the spatial relation “near” as well as the influencing contexts for such inputs. The result for this input is “near” and no influencing context. The absence of context indicates that every observation from the observation table (Table 2) has to be considered, which is represented by the contextualized concept **near@T**. The prototypical exemplar for **near@T** is 1000 m (conduct Fig. 5) which is included as a metric value in the following (pseudo SQL) spatial query by the PersonalizedGIS:

```
SELECT coords FROM restaurants
WHERE distance(actual_coords, coords) <= 1000 m;
```

Assume further that the smartphone with the PersonalizedGIS is equipped with an accelerometer that detects the mode of transportation. Now the user asks the above query while moving with the smartphone. The accelerometer detects the motion “walking”, which prompts the PersonalizedGIS to influence **near** by *walking*. So, rather than retrieving the prototypical exemplar for **near@T**, it retrieves the prototypical exemplar for **near@walking**. The prototypical exemplar is 300 m which can be used in the spatial query shown above by the PersonalizedGIS.

The PersonalizedGIS can narrow down the interpretation even further by traversing the context lattice automatically. Assume the function `GETSTRONGERCONTEXTS` for the context lattice outputs all infima for a given context. For this example the function is executed with `GETSTRONGERCONTEXTS(walking)` which outputs the infima: *walking*  $\wedge$  *uphill* and *walking*  $\wedge$  *driving*. The *walking*  $\wedge$  *driving* context is nonsense and is not taken into further account. The *walking*  $\wedge$  *uphill* context is used to narrow down the interpretation of the spatial term. For both contexts (**near@walking**  $\wedge$  *uphill* and **near@walking**) prototypical exemplars are 50 m and 300 m. These metric values are used as input to a refined spatial query. The refined query the PersonalizedGIS then executes, retrieving all those restaurants that are closer than 300 m, but excluding those that are uphill in respect to the current location (`actual_elevation <= rest.elevation`), unless they are closer than 50 m:

```

SELECT rest.* FROM (
  SELECT coords, elevation FROM restaurants
  WHERE distance(actual_coords, coords) <= 300 m
) AS rest WHERE
actual_elevation <= rest.elevation
AND distance(actual_coords, rest.coords) <= 50 m

```

## 6 Conclusions and Outlook

In this paper a computational model to map spatial terms onto spatial queries is introduced. In the review of the semiotic triangle, the problem of the many-to-many relation of symbols to objects in reality is identified as a main problem for computational models. We argue that context establishes a one-to-one connection between symbols and objects in reality. Our formalization is inspired by a quantum-mechanical approach presented by Aerts and Gabora (2005). The computational model integrates context and connects it to a concept underlying the externalized spatial term. In an envisioned next-generation GIS, the computational model is used to map the spatial term “near” onto different spatial queries dependent on context.

The envisioned GIS for the spatial relation “near” draws upon a set of observations that were assumed to be given. This is an important aspect that must be addressed in future work. In a realistic scenario the contexts can be derived from smartphone sensors. For example, the contexts: walking, driving, biking, etc. can be detected through accelerometer data or a mix of sensors, provided that ranges for the sensor values are detected that correspond to different contexts. Another mechanism that remains to be solved is aligning the observation base with the observations in the mind of a user. Feedback from the user can be used to gradually align the observations with the concepts in a user’s mind, as for example: “Was this distance near for you?”. It remains an open question how to get a user properly involved in such a mechanism. Perhaps via some sort of gamification process?

A more theoretical direction for future work concerns the investigation of the relations between the model presented in this paper—especially the distributions that exemplars take in a given context—and fuzzy membership functions (Zadeh 1965). Can the model be reinterpreted with classical fuzzy set theory? Would this add some benefits to operations and inferences that can be made when considering several contexts? Some previous work that addressed the problem of modeling concepts like **near** and **far** with fuzzy membership functions is presented by Wang (1994). Wang finds that **near** cannot be opportunely represented with a unique membership function. Rather, he suggests that more functions must be conceived as context information changes.

The mutual influence of several (contextualized) concepts warrants further investigation. Some previous work about concept combination for GIS is presented, for



example, by Hahn and Frank (2014) where thematic maps are selected on the basis of context.

Finally, for real usage of the model in applications it would be necessary to determine which contexts must be considered that can effectively influence a spatial concept.

## References

- Abdalla A, Weiser P, Frank AU (2013) Design principles for spatio-temporally enabled pim tools: A qualitative analysis of trip planning. In: Vandenbroucke D, Bucher B, Crompvoets J (eds) Geographic information science at the heart of Europe, Lecture notes in geoinformation and cartography. Springer, pp 323–336. doi:[10.1007/978-3-319-00615-4\\_18](https://doi.org/10.1007/978-3-319-00615-4_18)
- Aerts D, Gabora L (2005) A theory of concepts and their combinations i. *Kybernetes* 34(1/2):167–191. doi:[10.1108/03684920510575799](https://doi.org/10.1108/03684920510575799)
- Akman V, Surav M (1996) Steps toward formalizing context. *AI Mag* 17(3):55. doi:[10.1609/aimag.v17i3.1231](https://doi.org/10.1609/aimag.v17i3.1231)
- Bazire M, Brézillon P (2005) Understanding context before using it. In: Dey A, Kokinov B, Leake D, Turner R (eds) Modeling and using context, Lecture notes in computer science, vol 3554. Springer, Berlin, pp 29–40. doi:[10.1007/11508373\\_3](https://doi.org/10.1007/11508373_3)
- Bouquet P, Ghidini C, Giunchiglia F, Blanzieri E (2003) Theories and uses of context in knowledge representation and reasoning. *J Pragmatics* 35(3):455–484. doi: [10.1016/S0378-2166\(02\)00145-5](https://doi.org/10.1016/S0378-2166(02)00145-5)
- Burigo M, Coventry K (2010) Context affects scale selection for proximity terms. *Spat Cogn Comput* 10(4):292–312. doi:[10.1080/13875861003797719](https://doi.org/10.1080/13875861003797719)
- Chandler D (2007) *Semiotics: the basics*. Routledge
- Duckham M, Worboys M (2001) Computational structure in three-valued nearness relations. In: Montello D (ed) Spatial information theory, Lecture notes in computer science, vol 2205. Springer, Berlin, pp 76–91. doi:[10.1007/3-540-45424-1\\_6](https://doi.org/10.1007/3-540-45424-1_6)
- Egenhofer MJ, Mark DM (1995) Naive geography. In: Frank A, Kuhn W (eds) Spatial information theory a theoretical basis for GIS, Lecture notes in computer science, vol 988. Springer, Berlin, pp 1–15. doi:[10.1007/3-540-60392-1\\_1](https://doi.org/10.1007/3-540-60392-1_1)
- Fauconnier G (1994) *Mental spaces: aspects of meaning construction in natural language*. Cambridge University Press
- Fisher PF (2000) Sorites paradox and vague geographies. *Fuzzy Sets Syst* 113(1):7–18. doi:[10.1016/S0165-0114\(99\)00009-3](https://doi.org/10.1016/S0165-0114(99)00009-3)
- Fisher PF, Orf TM (1991) An investigation of the meaning of near and close on a university campus. *Comput Environ Urban Syst* 15(1–2):23–35. doi:[10.1016/0198-9715\(91\)90043-D](https://doi.org/10.1016/0198-9715(91)90043-D)
- Frank AU (1992) Qualitative spatial reasoning about distances and directions in geographic space. *J Vis Lang Comput* 3(4):343–371. doi:[10.1016/1045-926X\(92\)90007-9](https://doi.org/10.1016/1045-926X(92)90007-9)
- Frank AU (2006) Distinctions produce a taxonomic lattice: are these the units of mentalese? In: Bennete B, Fellbaum C (ed) *Formal ontology in information systems*, vol 150. IOS Press, pp 27–38
- Freksa C, Barkowsky T (1996) On the relation between spatial concepts and geographic objects. Geographic objects with indeterminate boundaries, pp 109–121
- Gratzer G (2009) *Lattice theory: first concepts and distributive lattices*. Courier Corporation
- Hahn J, Frank AU (2014) Select the appropriate map depending on context in a hilbert space model (scop). In: Atmanspacher H, Haven E, Kitto K, Raine D (eds) *Quantum interaction*, Lecture notes in computer science, vol 8369. Springer, Heidelberg, pp 122–133. doi:[10.1007/978-3-642-54943-4\\_11](https://doi.org/10.1007/978-3-642-54943-4_11)

- Huang H, Hahn J, Claramunt C, Reichenbacher T (eds) (2014) Proceedings of the 1st international workshop on context—awareness in geographic information services (CAGIS 2014 ). Eigenverlag, Wien. [http://publik.tuwien.ac.at/files/PubDat\\_232845.pdf](http://publik.tuwien.ac.at/files/PubDat_232845.pdf)
- Kuhn W (2005) Geospatial semantics: why, of what, and how? In: Spaccapietra S, Zimányi E (eds) *Journal of Data Semantics III, Lecture notes in computer science*, vol 3534. Springer, Heidelberg, pp 1–24. doi:[10.1007/11496168\\_1](https://doi.org/10.1007/11496168_1)
- Kuhn W (2009) Semantic engineering. In: Navratil G (ed) *Research trends in geographic information science, Lecture notes in geoinformation and cartography*. Springer, Heidelberg, pp 63–76. doi:[10.1007/978-3-540-88244-2\\_5](https://doi.org/10.1007/978-3-540-88244-2_5)
- Mark DM (1993) Toward a theoretical framework for geographic entity types. In: Frank A, Campari I (eds) *Spatial information theory a theoretical basis for GIS, Lecture notes in computer science*, vol 716. Springer, Heidelberg, pp 270–283. doi:[10.1007/3-540-57207-4\\_18](https://doi.org/10.1007/3-540-57207-4_18)
- Mark DM, Turk AG (2003) Landscape categories in yindjibarndi: ontology, environment, and language. In: Kuhn W, Worboys M, Timpf S (eds) *Spatial information theory. Foundations of geographic information science, Lecture notes in computer science*, vol 2825. Springer, Heidelberg, pp 28–45. doi:[10.1007/978-3-540-39923-0\\_3](https://doi.org/10.1007/978-3-540-39923-0_3)
- Mark DM, Freksa C, Hirtle SC, Lloyd R, Tversky B (1999a) Cognitive models of geographical space. *Int J Geogr Information Science* 13(8):747–774. doi:[10.1080/136588199241003](https://doi.org/10.1080/136588199241003)
- Mark DM, Smith B, Tversky B (1999b) Ontology and geographic objects: An empirical study of cognitive categorization. In: Freksa C, Mark D (eds) *Spatial information theory. Cognitive and computational foundations of geographic information science, Lecture notes in computer science*, vol 1661. Springer, Heidelberg, pp 283–298. doi:[10.1007/3-540-48384-5\\_19](https://doi.org/10.1007/3-540-48384-5_19)
- Montello DR, Friendschuh S (2005) Cognition of geographic information. A research agenda for geographic information science, pp 61–91
- Montello DR, Goodchild MF, Gottsegen J, Fohl P (2003) Where’s downtown?: behavioral methods for determining referents of vague spatial queries. *Spat Cogn Comput* 3(2–3):185–204. doi:[10.1080/13875868.2003.9683761](https://doi.org/10.1080/13875868.2003.9683761)
- Montello DR, Friedman A, Phillips DW (2014) Vague cognitive regions in geography and geographic information science. *Int J Geogr Inf Sci* 28(9):1802–1820. doi:[10.1080/13658816.2014.900178](https://doi.org/10.1080/13658816.2014.900178)
- Nosofsky RM (2011) The generalized context model: an exemplar model of classification. *Formal approaches in categorization*, pp 18–39
- Ogden CK, Richards (1946) *The meaning of meaning*. Harcourt, Brace and World, New York
- Osherson DN (1999) On the adequacy of prototype theory as a theory of concepts Daniel N, Osherson and Edward E. Smith. *Concepts: core readings*, p 261
- Raubal M, Winter S (2002) Enriching wayfinding instructions with local landmarks. In: Egenhofer M, Mark D (eds) *Geographic information science, Lecture notes in computer science*, vol 2478. Springer, Heidelberg, pp 243–259. doi:[10.1007/3-540-45799-2\\_17](https://doi.org/10.1007/3-540-45799-2_17)
- Robinson V (2000) Individual and multipersonal fuzzy spatial relations acquired using human-machine interaction. *Fuzzy Sets Syst* 113(1):133–145. doi:[10.1016/S0165-0114\(99\)00017-2](https://doi.org/10.1016/S0165-0114(99)00017-2)
- Rosch E (1973) On the internal structure of perceptual and semantic categories. In: Moore TE (ed) *Cognitive development and the acquisition of language*. Academic Press, Oxford, p 308
- Rosch E (1999) *Principles of categorization. Concepts: core readings*, pp 189–206
- Rosch E, Mervis CB (1975) Family resemblances: studies in the internal structure of categories. *Cogn Psychol* 7(4):573–605
- Seiler TB (2001) *Begreifen und Verstehen: Ein Buch über Begriffe und Bedeutungen*. Wiss.-HRW eK, Allg
- Smith B, Mark DM (1998) Ontology with human subjects testing. *Am J Econ Sociol* 58(2):245–312
- Talmy L (2003) *Toward a cognitive semantics*, vol 1. MIT press
- Tversky B (2003) Navigating by mind and by body. In: Freksa C, Brauer W, Habel C, Wender K (eds) *Spatial cognition III, Lecture notes in computer science*, vol 2685. Springer, Heidelberg, pp 1–10. doi:[10.1007/3-540-45004-1\\_1](https://doi.org/10.1007/3-540-45004-1_1)

- Twaroch F, Frank A (2005) Sandbox geography—to learn from children the form of spatial concepts. In: *Developments in spatial data handling*. Springer, Heidelberg, pp 421–433. doi:[10.1007/3-540-26772-7\\_32](https://doi.org/10.1007/3-540-26772-7_32)
- Von Glasersfeld E (1995) *Radical Constructivism: a Way of Knowing and Learning*. Stud Math Educ Ser: 6 ERIC
- Wallgrün JO, Klippel A, Baldwin T (2014) Building a corpus of spatial relational expressions extracted from web documents. In: *Proceedings of the 8th workshop on geographic information retrieval*, ACM, New York, NY, USA, GIR' 14, pp 6:1–6:8. doi:[10.1145/2675354.2675702](https://doi.org/10.1145/2675354.2675702)
- Wang F (1994) Towards a natural language user interface: an approach of fuzzy query. *Int J Geogr Inf Syst* 8(2):143–162. doi:[10.1080/02693799408901991](https://doi.org/10.1080/02693799408901991)
- Weiser P, Frank AU (2013) Cognitive transactions—a communication model. In: Tenbrink T, Stell J, Galton A, Wood Z (eds) *Spatial information theory*, Lecture notes in computer science, vol 8116. Springer, pp 129–148. doi:[10.1007/978-3-319-01790-7\\_8](https://doi.org/10.1007/978-3-319-01790-7_8)
- Winter S, Raubal M, Nothegger C (2005) Focalizing measures of salience for wayfinding. In: Meng L, Reichenbacher T, Zipf A (eds) *Map-based mobile services*. Springer, Heidelberg, pp 125–139. doi:[10.1007/3-540-26982-7\\_9](https://doi.org/10.1007/3-540-26982-7_9)
- Worboys MF (2001) Nearness relations in environmental space. *Int J Geogr Inf Sci* 15(7):633–651. doi:[10.1080/13658810110061162](https://doi.org/10.1080/13658810110061162)
- Worboys MF (2003) Communicating geographic information in context. *Foundations of geographic information science*, pp 33–45
- Yao X, Thill JC (2006) Spatial queries with qualitative locations in spatial information systems. *Comput Environ Urban Syst* 30(4):485–502. doi:[10.1016/j.compenvurbsys.2004.08.001](https://doi.org/10.1016/j.compenvurbsys.2004.08.001). <http://www.sciencedirect.com/science/article/pii/S0198971504000523>. *Geographic Information Retrieval (GIR)*
- Zadeh LA (1965) Fuzzy sets. *Inf Control* 8(3):338–353. doi:[10.1016/S0019-9958\(65\)90241-X](https://doi.org/10.1016/S0019-9958(65)90241-X)

# Inferring Complex Geographical Concepts with Implicit Geometries Using Ontologies: A Case of Peninsulas

Xiang Zhang, Tinghua Ai and Jantien Stoter

**Abstract** Ontology-driven concept inference has the merit of high flexibility and transparency. Users can composite and reuse atomic primitive concepts and relationship to interpret complex geographical concept without the need to retouch or even know the technical details. The major issue that we are focusing on is the implicit geometry problem. That is, the geometries corresponding to some primitive concept defining the complex geographic concept are missing or not fully represented in a spatial database, making it impossible to inferring the high-level semantics of the objects. This paper combines terminological/assertional inference (for general logic reasoning) and spatial operations (for making implicit geometries explicit), therefore enabling an ontology-driven inference of complex concepts that can handle cases where some concept has no explicit geometries. In the end, the concept of peninsula is used to demonstrate the proposed methodology.

**Keywords** Part-whole relations • Hierarchical structure • Terminological reasoning • Implicit semantics • Delaunay triangulation

---

X. Zhang (✉) · T. Ai

School of Resource and Environmental Sciences, Wuhan University,  
Wuhan, China

e-mail: xiang.zhang@whu.edu.cn

T. Ai

e-mail: tinghua\_ai@163.net

X. Zhang · T. Ai

Key Laboratory of Geographic Information System, Ministry of Education,  
Wuhan, China

X. Zhang · T. Ai

Key Laboratory of Digital Mapping and Land Information Application Engineering,  
National Administration of Surveying, Mapping and Geo-Information, Wuhan, China

J. Stoter

Faculty of Architecture and the Built Environment, 3D GeoInformation,  
TU Delft, Delft, Netherlands

e-mail: j.e.stoter@tudelft.nl

## 1 Introduction

Inferring geographical concepts is concerned with finding the most specific semantics of the spatial objects. For instance, it is possible to infer house types (e.g. terraced house) or land-use types from bare geometry data (Lüscher et al. 2008). This is typically useful when users want to retrieve specific types of geographic features from a spatial database or via a spatial-enabled search engine.

However, information in current spatial databases typically covers basic topographic features and primitive semantics which are nonetheless not sufficient for the increasing need for complex spatial queries and semantic interoperation across information communities (Kuhn 2005; Klien 2007). Those queried spatial information usually includes complex geographic concepts (e.g. floodplain, terraced/semi-detached house) that commonly appear in natural languages, but that are not explicitly available in spatial databases. One challenge is thus concerning the semantic enhancement of geospatial data.

This paper takes one step further. We claim that, besides the lack of explicit semantics, the geometries required directly or indirectly for the spatial inference tasks are not always explicitly available in spatial databases, too. This brings about new challenges, e.g., triggering spatial algorithms during the logic inference to detect the implicit geometry. The implicit geometry issue in ontology-driven spatial data interpretation is a common issue in spatial information retrieval and semantic interoperation (Bennett et al. 2008), and may eventually decline the usefulness of the above-mentioned semantic enhancement practice.

This paper follows an ontology-driven approach to concept inference. We address the implicit geometry issue by combining the terminological inference with spatial operations. Specifically, certain algorithms are triggered on-demand to make implicit geometries explicit, before the reasoning proceeds with inferred semantics. Section 2 reviews related work. Section 3 is the core of the paper and describes the method in detail. Section 4 shows the feasibility of the method by applying it to the interpretation of peninsulas from spatial databases. The paper ends with discussion and conclusions (Sect. 5).

## 2 Related Work

### 2.1 Algorithmic Approach to Concept Inference

Many of the earlier spatial data interpretation techniques were motivated by the need of complex spatial information retrieval tasks. A basic assumption is that a lot of implicit information can be drawn from the geometric, topological and semantic relations encoded in spatial data (Sester 2000). Some of the techniques are based on graph theory, pattern recognition and statistical approaches, while some others

adapt methods from spatial data mining (e.g. Regnauld 1996; Christophe and Ruas 2002; Steiniger et al. 2008).

However, these approaches consist of algorithms where the knowledge is hard-coded, and they can hardly be reused in a different context. Hence we consider them in the class of algorithmic approach. Furthermore, Lüscher et al. (2008) argued that it is doubtful whether the comprehensive interpretation of more general, higher-level geographic concepts can be accomplished by the purely algorithmic approach.

## 2.2 *Ontology-Driven Spatial Data Interpretation*

In automated map generalization, interpretation of hidden semantics is closely related to ‘*data enrichment*’ (Neun et al. 2008). Traditionally, data enrichment techniques were developed and tightly coupled into sophisticated algorithms for specific tasks; see for example (Regnauld 1996; Christophe and Ruas 2002; Steiniger et al. 2008). This algorithm-driven approach was recognized by Sester (2000) and Lüscher et al. (2007) to have various weaknesses in applications where knowledge have to be made explicit.

In contrast, ontology-based approaches to the interpretation of geographic concepts have been increasingly adopted in recent years, where the interpretation is viewed as building a formal knowledge base for a domain on which reasoning processes are applied. The viewpoint is supported by recent developments in artificial intelligence (Baader et al. 2003; Möller and Neumann 2008). In the spatial domain, Lüscher et al. (2008) proposed an ontology-based model for urban pattern recognition. Later, they re-implemented the concepts using supervised Bayesian network (Lüscher et al. 2008), where the recognition process was manually translated from the ontology. This approach does not use the reasoning capability underlying ontologies. Similarly, Thomson and Béra (2007, 2008) showed the use of concept hierarchy to represent geographical concepts, and recommended *Description Logics* (DLs), a knowledge-formalism from artificial intelligence (Baader et al. 2003), but they did not implement the process. Nevertheless, we found that DLs is insufficient for the inference of geographical concepts because many concepts can only be recognized with algorithms that detect the spatial and/or part-whole relations between spatial objects. Hence to enable *spatio-terminological reasoning*, the inference process should be enriched with spatial computations.

In automated reasoning, the notion of spatio-terminological reasoning was proposed by Haarslev et al. (1994, 1998) aiming at integrating spatial calculation with logic-based reasoning process. The notion looks promising and we decide to follow this approach, but we will focus more in this paper on the issue of implicit geometries and how can it be integrated with the spatio-terminological inference.

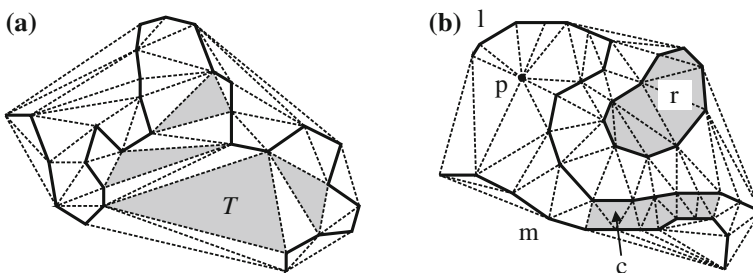
### 3 Method

The idea of our proposed method is as follows. Our method firstly formalizes the primitive concepts and relationships by a set of generic spatial operations (including algorithms), and then formalizes complex concepts with an ontological language. This formalization declares the complex concept as a formal structure of primitive concepts. In the next step the complex concepts are automatically inferred using the spatially enriched reasoning techniques. This approach is highly flexible since the generic spatial operations can be reused when inferring different geographical concepts by only altering the knowledge defined in the formal language (knowledge driven rather than algorithm driven).

#### 3.1 Generic Algorithms to Detect Primitive Relations and Implicit Geometries

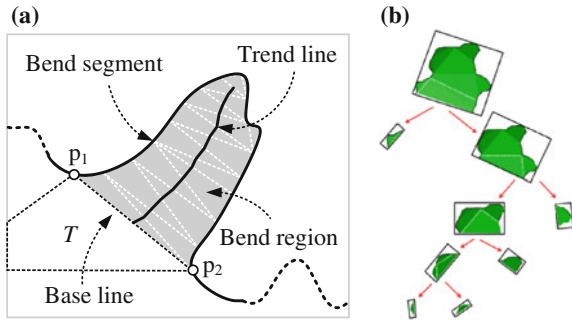
Here we outline a triangulation-based data structure, on top of which proximity relations and implicit geometries (especially the part-whole geometries) can be obtained on the structure (Fig. 1). For example, it is easy to use this structure to model spatial proximity: any two objects that are connected by a triangle edge are immediate neighbors (e.g.  $l$  and  $p$  in Fig. 1b); narrow part (sub-region with implicit geometries) between two parallel roads  $l$  and  $m$  can be identified also on this structure ( $c$  in Fig. 1b). For a more detailed explanation to the structure and its operations, one may refer to Ai (2006).

Due to the relevance to the subject matter of this paper, we describe in more detail the shape descriptor and algorithm concerning a bend structure. The descriptors are derived with Delaunay triangulation based algorithms (Fig. 2a). The *bend segment* is the curve between  $p_1$  and  $p_2$ ; the link from  $p_1$  to  $p_2$  defines the *base line* which can be used to depict the mouth of the bend; the extent of the *bend region* (gray area in Fig. 2a) is enclosed by the bend segment and the base triangle  $T$ . The *trend line* is



**Fig. 1** Triangulation-based structure that supports structural knowledge **a** sub-region and shape description for a meandering line; **b** spatial relations and narrow region detection

**Fig. 2** The descriptors of bend structure (a), which can be used to describe the shape of micro bends at different levels of hierarchy (b)



derived through triangulating and skeletonizing (Zhang et al. 2008; Ai et al. 2014) the bend region. The trend line of a bend region reflects the main orientation of the bend region which can be measured by the trend line pointing from the base to the end of the bend region. Properties like compactness and elongation can be derived based on these descriptors. Note that with the hierarchical bend structure described in Ai et al. (2014), any bend in the structure can be characterized by these descriptors (Fig. 2b) so that we can find needed ones in the hierarchy.

Since the above-mentioned operations are not built-in functions in spatial databases, we implemented them as database extensions so that they can be used together with those built-in functions.

### 3.2 Formalizing Geographical Concepts with Description Logics

While there are many knowledge formalisms (e.g. frames, semantic networks, rules), we decided to use *Description Logics* (Baader et al. 2003) as knowledge formalism for the description of concepts and relationships in the geospatial domain. DLs provide abstract syntax, which forms the foundation of many ontology languages such as the Web Ontology Language (OWL<sup>1</sup>). Note that the abstract syntax of DLs is used in the paper for clarity and OWL, a concrete syntax of DLs, is employed for implementation purposes (see Sect. 4).

To understand the DL syntax used, we briefly introduce the basics of the syntax. The syntax consists of *concepts* (unary predicates), *roles* (binary relations) and restrictions on roles. A role links an individual in a domain to an individual or property in a range which is also called *role filler*. The following axioms (a.k.a TBox) show a possible description of knowledge in the spatial domain:

<sup>1</sup><http://www.w3.org/TR/owl-guide/>.



$$\mathit{SpatialObject} \sqsubseteq_C \mathit{TopConcept}(\mathcal{T}) \quad (1)$$

$$\mathit{SpatialRelation} \sqsubseteq_R \mathit{SpatialObject} \times \mathit{SpatialObject} \quad (2)$$

$$\mathit{Building} \sqsubseteq_C \mathit{SpatialObject} \quad (3)$$

$$\mathit{hasNearbyNeighbor} \sqsubseteq_R \mathit{SpatialRelation} \quad (4)$$

$$\begin{aligned} \mathit{ClusteredBuilding} &\equiv_C \mathit{Building} \\ &\sqcap (\geq 2 \mathit{hasNearbyNeighbor}.\mathit{Building}) \end{aligned} \quad (5)$$

Axiom (1) expresses that *SpatialObject* is subsumed by the top concept (which is never subsumed by other concepts), where  $\sqsubseteq_C$  is the *concept subsumption* construct. Similarly, Axiom (2) describes a *role subsumption* ( $\sqsubseteq_R$ ). The role *SpatialRelation* has a *domain*: *SpatialObject* and a *range*: *SpatialObject* (an object mapped to another object). Since *hasNearbyNeighbor* is subsumed by *SpatialRelation*, this implies the former inherits the domain and range of the latter. Axiom (5) describes the concept *ClusteredBuilding* with the *concept definition* construct ( $\equiv_C$ ), expressing that a clustered building is a building which has at least two nearby neighbors. The statement also implies that *ClusteredBuilding* is subsumed by *Building*. The *intersection* construct ( $\sqcap$ ) is used when composing a complex concept from different atomic ones. In Axiom (5), atomic concepts are *Building* and ( $\geq 2 \mathit{hasNearbyNeighbor}.\mathit{Building}$ ). The latter statement can be seen as an *anonymous concept*, which restricts the intension of *ClusteredBuilding* in Axiom (5). In the *anonymous concept*, the *unqualified number restriction* construct ( $\geq$ ) is used to specify the cardinality of the role: *hasNearbyNeighbor.Building*, meaning that each clustered building must have at least 2 nearby neighbors which are instances of *Building*.

Other constructs like the *role definition* ( $\equiv_R$ ), the *existential quantifier* ( $\exists$ ), the *universal quantifier* ( $\forall$ ) and the *negation* ( $\neg$ ) will also be used in the remaining sections to define high-level cartographic concepts and complex roles. Details of all notations, their semantics and interpretations refer to Baader et al. (2003).

### 3.2.1 Reasoning with Description Logics

DLs can be used to describe the knowledge bases (KB) for a concrete domain. A knowledge base consists of a set of terminological axioms (TBox) and a set of assertional axioms (ABox). A TBox forms a priori knowledge of the domain by defining concepts and their relations (e.g. Axioms (1)–(5)), while an ABox describes the known facts about the world. An example ABox may look like:

*Building*(*a*) - concept assertion  
*hasNearbyNeighbor*(*a*, *b*) - role assertion

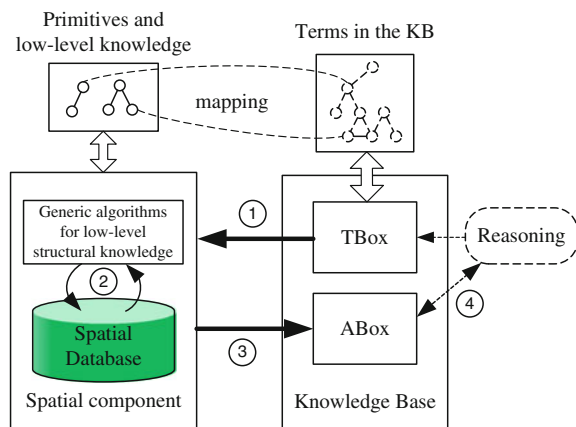
where *a*, *b* are data instances, and the ABox asserts that *a* is a building and *b* is *a*'s nearest neighborhood. Reasoning with KB is an important ability of DL systems. It can be used to infer implicit knowledge (semantics) hidden in TBoxes and ABoxes, and make them explicit (Baader et al. 2003). The inference capabilities that are concerned in this paper are *Realization* (i.e. finding the most specific concept of which an object is an instance) and *Retrieval* (i.e. finding all objects that are instances of a given concept).

The interpretation task can be viewed as a *realization* process, which aims at finding the most specific concept of which an individual is an instance (Möller and Neumann 2008). However, the interpretation of complex geo-concepts from spatial databases cannot be addressed simply based on the logic-based reasoning. Take the interpretation of *ClusteredBuilding* for example. If *a*, *b*, and *c* are asserted as instances of *Building* in an ABox, the *realization* will not work as the reasoner does not know the relation between the three objects. This therefore calls for a spatio-terminological reasoning.

### 3.3 Concept Interpretation as Reasoning Over Knowledge Bases

To address the spatial related reasoning problem, we adopt the notion of spatio-terminological reasoning to integrate spatial algorithms with logic-based automated reasoning process. However we implement the notion in a different way than Haarslev et al. (1994). In the alternative integration the spatial functionalities are loosely coupled with DL systems. The design is shown in Fig. 3.

**Fig. 3** The proposed design of spatio-terminological reasoning process

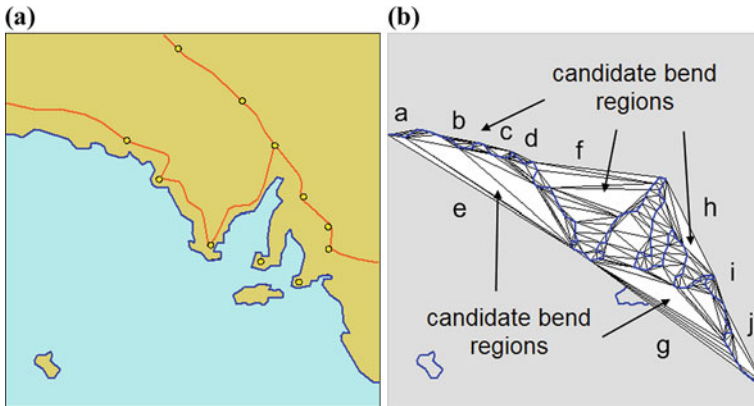


The spatio-terminological reasoning process in our design has two general components, namely the spatial component and the DL system (including a knowledge base and a reasoning engine). The basic idea of this spatio-terminological reasoning is to design a workflow to control the communication between the spatial component and the DL system to facilitate the automatic interpretation process. In Fig. 3 ‘Terms in the KB’ is directly connected to the TBox in the knowledge base. It represents the defined structures of any high-level concept of interest. ‘Spatial component’ consists of a spatial database and a set of generic operations (algorithms) to detect low-level structural knowledge. Here, we assume that the basic operations such as geometric, topological predicates and non-spatial queries are available in the spatial database. Hence the spatial component is sufficient to detect the low-level knowledge, including primitive entities, low-level properties and structural relationships. The vocabulary of the spatial component is marked by ‘Primitives and low-level knowledge’. The ‘Reasoning’ component is a standard reasoner which is responsible for the consistency of both the TBox and the ABox and other reasoning services.

The automated interpretation of a high-level concept is decomposed into four major steps (see also Fig. 3):

1. **Mapping:** all Parent Concepts (*PC*) which subsume the high-level concept, all Roles (*R*) and Role Fillers (*RF*) appeared in the axioms of the concept definition are firstly identified. Then *PC*, *R* and *RF* are mapped to the vocabulary provided by the spatial component. This results in a list of database concepts and spatial relationships needed for the next step;
2. **Spatial processing:** according to the list resulted from step (1), database objects that are instances of *PC* and *RF* are retrieved. Then, the identified spatial relationships are tested with spatial operations (either built-in predicates or enriched algorithms) between the objects belonging to *PC* and the objects belonging to *RF*. The test subjects are always the objects from *PC*. The number of object pairs for each testing can be reduced using the range in each role specified by DLs;
3. **Assertion:** all the retrieved database objects (*O*) in the last step are asserted as the instances of *PC* and *RF* and then added to the ABox. In the case of role assertion, the object pairs that pass the relationship testing in the last step are added to the ABox as new role assertions. After this, ABox becomes ABox’, which is well prepared for the *realization* step;
4. **Reasoning:** by *realization* service, the high-level concept (hidden semantics) is inferred automatically with the ABox’ and the TBox. The interpretation process is by then terminated.

Note that, step (1) and step (3) are the communications between spatial component and the DL system, whereas step (2) and step (4) are carried out within the spatial component and the DL system respectively.



**Fig. 4** **a** Example data set in Australia; **b** candidate bend regions generated for the interpretation task

## 4 Case Study: Interpreting Peninsulas from Spatial Databases

A prototype encompassing the spatial component and DL system was implemented to validate the proposed methodology. Besides the spatial operations available in the database, the generic algorithms that detect the proximity relationships, generate parts from wholes, and forming wholes from parts were implemented on top of a Delaunay triangulation (DT) based model. We used Pellet<sup>2</sup> as the underlying DL reasoner, which can be readily accessed by its OWL API.<sup>3</sup> The API was used for practical reasons: it enables programmers to define OWL-based knowledge bases (though they can be defined using an ontology editor like Protégé<sup>4</sup>), to access or modify concepts/roles axioms in a TBox, and to add known facts (e.g. *Sea(a)*, *Neighbor(a, b)*) to or remove them from an ABox. Moreover, one can invoke almost all automated reasoning services available in the underlying reasoner via the API.

### 4.1 Setting the Scene

A simple test dataset of coastal area is depicted in Fig. 4a, where peninsulas, harbors, bays, mainland, island, sea, cities are visible to human beings on the map. People usually asks, e.g., which cities/places are inside a peninsula? Here peninsula is understood as a region with implicit boundaries. Therefore, features like peninsulas

<sup>2</sup><http://clarkparsia.com/pellet/>.

<sup>3</sup><http://owlapi.sourceforge.net/>.

<sup>4</sup><http://protege.stanford.edu/>.

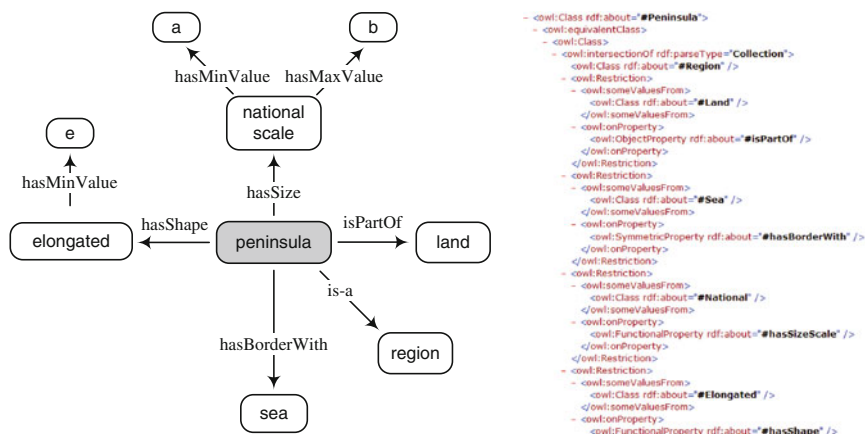


Fig. 5 A schematic diagram of the ontology of peninsulas (left) and its OWL description (right)

should be firstly retrieved before one can answer such questions. We applied the proposed method to identify instances of concepts like ‘peninsula’ and ‘bay’.

## 4.2 Initializing the Knowledge Base

We first built up the initial TBox using Protégé and check its consistency. Atomic concepts in the case consist of *Region*, *Land*, *Island*, *Sea*, and *City*; complex ones include *Peninsula*, *Bay*, *CityInPeninsula*, and *OtherCity* which are formed based on the atomic concepts, attributes, and their relationships. Then, a mapping table was established so that the declared concepts can be linked to database classes and spatial operations. Finally, all the spatial objects in the database are added as initial known facts to the ABox. This is done by first searching the objects in the spatial database according to the classes appeared in the mapping table, and then add the identifiers of these objects to the ABox, using OWL API. Example assertions are *Sea(Australia\_sea)* and *Land(Australia\_land)*.

*Peninsula* is defined based on explanations in Wikipedia<sup>5</sup>: “[...] it is a region, with shape and size restrictions, contained by land and sharing common boundary with sea.” Figure 5 shows its ontology and OWL document.

As a proof-of-concept, ‘*Peninsula*’ is defined in DLs in Listing (6). However, we are by no means to say that this is the only way to define peninsulas. For instance,

<sup>5</sup>[http://en.wikipedia.org/wiki/List\\_of\\_peninsulas](http://en.wikipedia.org/wiki/List_of_peninsulas).

a peninsula needs not to be at the national scale. We add this constraint here only to demonstrate the formalism and the inference process.

$$\begin{aligned}
 \mathit{Peninsula} \equiv_c \mathit{Region} & \\
 \sqcap \exists \mathit{isPartOf.Land} & \\
 \sqcap \exists \mathit{hasBorderWith.Sea} & \\
 \sqcap \exists \mathit{hasShape.Elongated} & \\
 \sqcap \exists \mathit{hasSize.NationalScale} &
 \end{aligned} \tag{6}$$

#### 4.2.1 Characterization of Peninsula Candidates

The inference starts by parsing the OWL document of *Peninsula* using the API. The program informs the spatial component that only one parent concept (i.e. *Region*) is needed for the subsequent reasoning. However, only ‘*Austrilia\_sea*’ and ‘*Austrilia\_land*’ regions are available in the database (Fig. 4a), which fail to pass the *isPartOf.Land* test. Therefore, we generated the candidate bend regions using the algorithm described in Sect. 3.1 and fed the detected sub-regions  $\{a, b, c, d, e, f, g, h, i, j\}$  (Fig. 4b) into the ABox.

After the high-level concept is formalized with DLs, the next step is the detection of the required low-level knowledge. The formal definition of Peninsula (Listing (6)) informs the spatial component that knowledge of topology, shape, and size are needed. The knowledge on these three concepts is detected as follows.

*hasShape* is characterized using Elongation ratio (i.e. length (trend line)/length (base line)) based on the bend descriptors introduced in Sect. 3.1. Two qualitative descriptors are derived here: *Elongated* (Elongation ratio  $\geq 0.6$ ) and *Flattened* (Elongation ratio  $< 0.6$ ). These reflect only roughly the shape property of peninsulas as well as bays, rather than definitive values.

*hasSize* is determined by the size of bend region. Since the size of any peninsula in reality has a lower and an upper bound, *NationalScale*, a value restriction, is specified tentatively as  $\text{size} \in [80, 250] \times 10^2 \text{ km}^2$ . Likewise, *Local*:  $[0, 20]$ , *Regional*:  $[20, 80]$  and *Global*:  $[250, -]$  are specified as tentative values characterizing different sizes.

*isPartOf* and *hasBorderWith* were mapped to topological operations (e.g. contain and meet) in the spatial database. Such relationships are tested between the detected bend regions and other objects specified in the range of the relationships, and the new role assertions are formed (e.g. *isPartOf(e, Austrilia\_sea)*).

After the above knowledge was tested for all bend regions and between the regions and other individuals, the asserted knowledge was added to the ABox via OWL API. Finally, the automated interpretation was launched by invoking the *realization* service in Pellet, also through the API. This process is fully automated. The size and shape values of all the candidate regions are displayed in Table 1.

**Table 1** Characteristics (Char.) of the bend regions derived using the descriptors mentioned in Sect. 4.2.1 (L-Local; R-Regional; N-National; G-Global; E-Elongated; F-Flattened)

Instance	a	b	c	d	e	f	g	h	i	j
Size ( $\times 10^2$ km <sup>2</sup> )	5.7	9.6	11.7	3.8	554	486	691	246	15.6	72.8
Char.	L	L	L	L	G	G	G	N	L	R
Shape	0.2	0.5	0.4	0.4	0.3	0.8	0.9	1.0	0.6	0.3
Char.	F	F	F	F	F	E	E	E	E	F

## 4.2.2 Refined Results

The automated interpretation identifies  $h$  as the only peninsula from the candidate regions. This result is not very promising as we also observe other peninsulas in Fig. 4a. There are several reasons for the misinterpretations. First, the use of crisp condition such as *NationalScale* is too restrictive. Second, the inference did not make use of all bend regions detected at different levels of detail.

In this experiment, the inference made use of all bend regions in the hierarchy (using the technique in Ai et al. 2014). The inference traverses the hierarchy from the root until it finds the largest bend regions that satisfy conditions defined in Listing (6). Some bend regions that originally had branches are decomposed into separated bend regions (e.g. the bend  $g$  and  $h$  in Fig. 4b). The shape and size of these generated sub-regions are shown in Table 2 (some small bends are excluded for clarity). The refined result was more satisfactory. Regions  $\{f1, h1, h2\}$  are now recognized as peninsulas and regions:  $\{g1, g2\}$  are recognized as bays (bay differs from peninsula only in that it is part of sea and adjacent to land; its formal definition is not shown due to limited space). The results of automated interpretation are visualized in Fig. 6. As a consequence, the described interpretation task also identifies instances of *CityInPeninsula* and *OtherCity*.

To get more insight into the reasoning process we can look at different states of the knowledge base during the inference process (visualized in Fig. 7 Protégé with Ontoviz plug-in). For clarity, only *subsumption* and *instance-of* relationships are displayed.

**Table 2** The characteristics of the sub-regions refined from regions  $f$ ,  $g$ , and  $h$ 

Instance <sup>1</sup>	f		g			h	
	f1	f2	g1	g2	g3	h1	h2
Size	187	12	223	81	106	96	54
Char.	N	L	N	N	N	N	R
Shape	0.9	0.7	2.3	1.9	0.5	2.6	0.9
Char.	E	E	E	E	F	E	E

<sup>1</sup>Instances obtained using hierarchical bend structure, listing only the changed instances

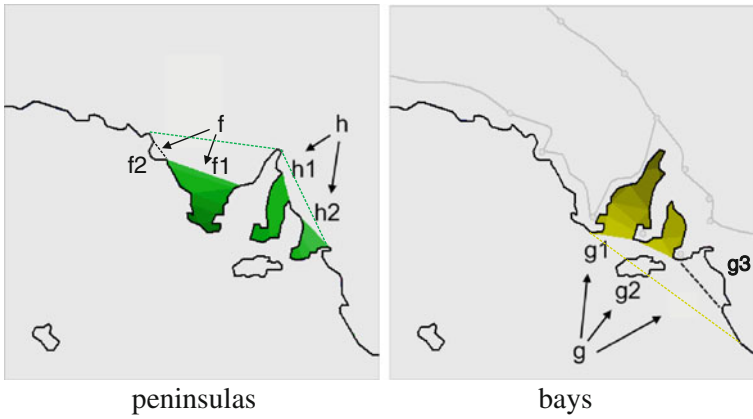


Fig. 6 Visualized results of the interpreted instances

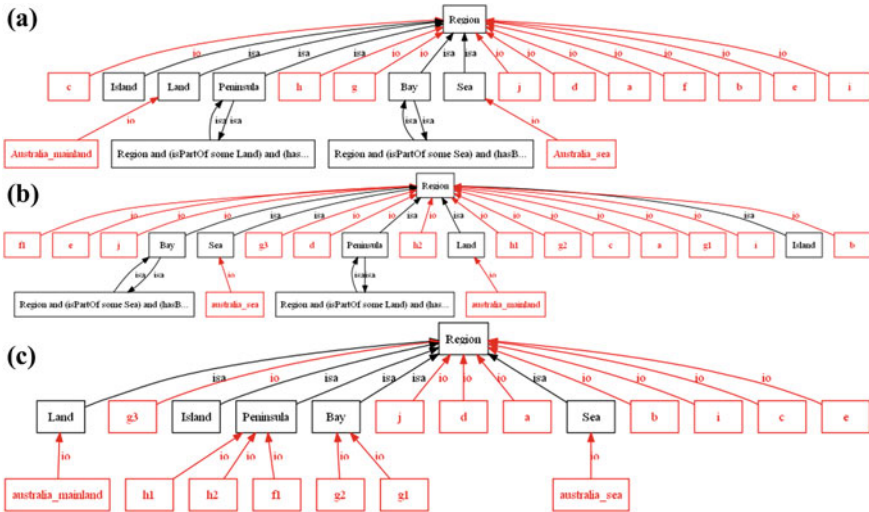


Fig. 7 Different states of the knowledge base (‘isa’—subclass of; ‘io’—instance of) **a** individuals and their asserted types in initial state; **b** the state after refinement to candidate bend regions; **c** the state after automated reasoning—most specific concepts of instances inferred

The bend regions are originally instances of the concept *Region*. After the *realization* process, some of them are inferred as instances of more specific concepts such as *Peninsula* and *Bay* (Fig. 7c).



## 5 Main Findings and Outlook

In this paper we proposed a method to automatically interpret complex geographical concepts from low-level knowledge. A practical spatio-terminological reasoning process was designed and implemented by enhancing existing DL reasoners with spatial functionalities. Combining the spatial and terminological components and applying them to the example of peninsula showed the potential of this methodology.

The use of *Description Logics* enables a more transparent modeling and better maintenance of spatial knowledge. A domain expert or knowledge engineer does not need to dive into the implementation level in order to revise the condition for inferring a different concept. The ultimate goal is that one can find a set of atomic spatial operations that are reusable and composable in flexible ways, so that query over any complex geographic concept can be achieved by chaining those atomic operations and primitive facts, as described in the knowledge base. But the harder question is whether there exists such a set of atomic spatial operations, and how to semantically annotate the operations such that they can be found and matched to foster a fully automated inference process. Further research along this line of thought should also contribute to the endeavor of semantic web and service chaining. In addition, future work will see how the proposed approach performs with real world datasets.

In addition, the proposed method does not consider the uncertainty in the knowledge modeling process. Uncertainty is unavoidable even in our toy example, where crisp thresholds, such as shape (*Elongated*) and size (*NationalScale*) in the peninsula case, are limiting factors and should be address in the future. We have to adopt the workaround because the inference tools available to us at the moment were not capable of making uncertainty reasoning.

In the past few years, there are many proposals that aim to extend DLs to be able to handle uncertainty (Haarslev et al. 2006). The uncertainty can be handled at the language, knowledge base, and reasoning levels. The underlying reasoning model can be based on Fuzzy logic, probabilistic theory or others (Baader et al. 2003; Haarslev et al. 2006). This trend is also taken by the W3C working group on uncertainty reasoning,<sup>6</sup> and may become the foundation for the next generation web (semantic web). Although uncertainty reasoning using DLs has been proposed in recent years (e.g. Carvalho et al. 2010), there is still a lack of tools for practical use. Upon the perfection of such inference techniques, geospatial information retrieval would benefit most from it as geographic concepts are inherently vague.

**Acknowledgments** We thank the three anonymous reviewers for their helpful comments. The work is supported by the National Natural Science Foundation of China (Grant No. 41301410 and No. 41531180) and the National High Technology Research and Development Program of China (Grant No. 2015AA1239012).

---

<sup>6</sup><http://www.w3.org/2005/Incubator/urw3/XGR-urw3-20080331/>.

## References

- Ai T (2006) A spatial field representation model based on Delaunay triangulation. *Acta Geodaetica Cartogr Sin* 35(1):71–76
- Ai T, Zhou Q, Zhang X, Huang Y, Zhou M (2014) A simplification of Ria coastline with geomorphologic characteristics preserved. *Mar Geodesy* 37:167–186
- Baader F, Calvanese D, McGuinness DL, Nardi D, Patel-Schneider PF (2003) *The description logic handbook: theory, implementation and applications*. Cambridge University Press
- Bennett B, Mallenby D, Third A (2008) An ontology for grounding vague geographic terms. In: Eschenbach C, Grüninger M, (eds) *Formal ontology in information systems—Proceedings of the fifth international conference*. FOIS-08, Saarbrücken, Germany, pp 280–293
- Carvalho RN, Laskey KB, Costa PCG (2010) PR-OWL 2.0—Bridging the gap to OWL semantics. In: *Proceedings of the 9th international semantic web conference*, Shanghai, China, p 12
- Christophe S, Ruas A (2002) Detecting building alignments for generalisation purposes. In: Richardson DE, van Oosterom P (eds) *Advances in spatial data handling (10th international symposium on spatial data handling)*, Springer, Berlin Heidelberg, New York, pp 419–432
- Haarslev V, Möller R, Schröder C (1994) Combining spatial and terminological reasoning. In: Nebel B, Dreschler-Fischer L (eds) *Advances in artificial intelligence*, proceedings of the 18th German annual conference on artificial intelligence, LNAI 861, pp 142–153
- Haarslev V, Lutz G, Möller R (1998) Foundations of spatioterminological reasoning with description logics. In: Cohn A, Schubert L, Shapiro S (eds.) *Principles of knowledge representation and reasoning*, proceedings of the sixth international conference (KR'98), pp 112–123
- Haarslev V, Pai H, Shiri N (2006) Uncertainty reasoning in description logics: a generic approach. In: *Proceedings of the 19th international FLAIRS conference*. AAAI Press, pp 818–823
- Klien E (2007) A rule-based strategy for the semantic annotation of geodata. *Trans GIS* 11 (3):437–452
- Kuhn W (2005) Geospatial semantics: why, of what, and how? *J Data Semant III*, LNCS 3534:1–24
- Lüscher P, Burghardt D, Weibel R (2007) Ontology-driven enrichment of spatial databases. In: *The 10th ICA workshop on generalisation and multiple representation*, Moscow
- Lüscher P, Weibel R, Mackaness W (2008) Where is the terraced house? On the use of ontologies for recognition of urban concepts in cartographic databases. In: *Proceedings 13th international symposium on spatial data handling*, pp 449–466
- Möller R, Neumann B (2008) Ontology-based reasoning techniques for multimedia interpretation and retrieval. In: Kompatsiaris Y, Hobson P (eds) *Semantic multimedia and ontologies*, pp 55–98
- Neun M, Burghardt D, Weibel R (2008) Web service approaches for providing enriched data structures to generalization operators. *Int J Geogr Inf Sci* 22(2):133–165
- Regnauld N (1996) Recognition of building clusters for generalization. In: Kraak M-J, Molenaar M (eds) *Advances in GIS research II: proceedings of the seventh international symposium on spatial data handling*. Taylor & Francis, London, 4B, pp 1–14
- Sester M (2000) Knowledge acquisition for the automatic interpretation of spatial data. *Int J Geogr Inf Sci* 14(1):1–24
- Steiniger S, Lange T, Burghardt D, Weibel R (2008) An approach for the classification of urban building structures based on discriminate analysis techniques. *Trans GIS* 12(1):31–59
- Thomson MK, Béra R (2007) Relating land use to the landscape character: toward an ontological inference tool. In: *Proceedings GIS research UK 15th annual conference (GISRUK-2007)*, Maynooth, Ireland

- Thomson MK, Béra R (2008) A methodology for inferring higher level semantic information from spatial databases. In: Proceedings of the GIS research UK 16th annual conference (GISRUK 2008), Manchester, UK
- Zhang X, Ai T, Stoter J (2008) The evaluation of spatial distribution density in map generalization. In: International archives of the photogrammetry, remote sensing and spatial information sciences (ISPRS Congress 2008), vol XXXVII. Part B2, Beijing, pp 181–188

# Question-Based Spatial Computing—A Case Study

Behzad Vahedi, Werner Kuhn and Andrea Ballatore

**Abstract** Geographic Information Systems (GIS) support spatial problem solving by large repositories of procedures, which are mainly operating on map layers. These procedures and their parameters are often not easy to understand and use, especially not for domain experts without extensive GIS training. This hinders a wider adoption of mapping and spatial analysis across disciplines. Building on the idea of core concepts of spatial information, and further developing the language for spatial computing based on them, we introduce an alternative approach to spatial analysis, based on the idea that users should be able to ask questions about the environment, rather than finding and executing procedures on map layers. We define such questions in terms of the core concepts of spatial information, and use data abstraction instead of procedural abstraction to structure command spaces for application programmers (and ultimately for end users). We sketch an implementation in Python that enables application programmers to dispatch computations to existing GIS capabilities. The gains in usability and conceptual clarity are illustrated through a case study from economics, comparing a traditional procedural solution with our declarative approach. The case study shows a reduction of computational steps by around 45 %, as well as smaller and better organized command spaces.

**Keywords** Spatial computing · Core concepts · Question-based analysis · Abstract data types

---

B. Vahedi (✉) · W. Kuhn · A. Ballatore  
Center for Spatial Studies, Department of Geography,  
University of California at Santa Barbara (UCSB), Santa Barbara, CA, USA  
e-mail: behzad@geog.ucsb.edu

W. Kuhn  
e-mail: werner@ucsb.edu

A. Ballatore  
e-mail: aballatore@spatial.ucsb.edu

© Springer International Publishing Switzerland 2016  
T. Sarjakoski et al. (eds.), *Geospatial Data in a Changing World*,  
Lecture Notes in Geoinformation and Cartography,  
DOI 10.1007/978-3-319-33783-8\_3

# 1 Introduction

Geographic Information Systems (GIS), whether proprietary or open source, have evolved into large and complex toolboxes that require expert technical knowledge to be applied to real-world spatial problems. This evolution toward greater complexity happened primarily because of the growing demand for spatial computing in business and government operations. The demand for more functionality has mostly been addressed through “creeping featurism”, i.e., incremental extensions to existing tools rather than a comprehensive approach to designing the next generation of spatial computing systems (Tomlinson 2007).

As a result, scientists and decision makers can often only benefit from GIS after prolonged training. Moreover, domain competence tends to get separated from technological prowess, producing disjointed and potentially sub-optimal solutions to problems. It is then often impossible or too time consuming to iterate solution procedures or explore alternatives.

GIS users and application programmers who wish to ask *questions* using spatial data are forced to write *procedures* to generate and manipulate maps that, hopefully, will answer the questions. Domain questions are often lost in translation during the formulation and execution of these procedures. We believe that an approach based on the idea of asking spatial questions will create new and greater opportunities for programmers to develop applications and eventually allow users to interact with spatial computing systems in a way that is more interrogative and, thus, more natural and productive. Clearly, the challenge to address is to define the vocabulary in terms of which such questions can be asked and answered.

This paper proposes a novel approach to address this challenge and, more generally, the fragmented landscape of spatial computing. To do so, it utilizes the core concepts of spatial information presented earlier (Kuhn 2012; Kuhn and Ballatore 2015) and Abstract Data Types (ADTs) built on them to propose a new style of computational solution to spatial problems, in terms of spatial *questions* to be asked rather than *procedures* to be executed. The approach is illustrated by a case study from economics, in which a domain expert evaluates economic activity in China by assessing nocturnal luminosity around roads. All source code is available online.<sup>1</sup>

The remainder of this paper is organized as follows. Section 2 explains the idea of question-based computing and the differences to procedural approaches. Section 3 presents a summary of related work and is followed by Sect. 4 on the notion of core concepts of spatial information and the pertinent computations. The case study is presented and its Python-based implementation is shown in Sect. 5, before concluding with a discussion of results and ongoing research (Sect. 6).

---

<sup>1</sup><https://github.com/spatial-ucsb/ConceptsOfSpatialInformation>.

## 2 Question-Based Spatial Computing

We posit that users want to find *answers to spatial questions* when they use GIS. These questions might stem from different domains, ranging from economics to ecology and public health, but they usually reflect some form of spatial thinking and reasoning used by humans (Gao and Goodchild 2013). Yet, GIS are not designed in a way that would allow users to find and understand answers without going through procedures that are often complicated and difficult to assemble. GIS users do not really want to assemble chains of procedures, but have to learn to do so, often with difficulty, because there is no alternative. Application developers, in turn, find it hard to develop question-answering tools, because the computations available to them are not organized around contents and questions, but around layers, data formats, and historically grown commands and parameters.

A system organized around questions rather than procedures could be beneficial for end users since it would be closer to their needs and thus easier to understand. Users of such a system would not need to worry about the procedures and their sometimes numerous input parameters, often resulting from implementation details. The ideal situation could be compared to the realm of relational databases where Edgar F. Codd in the 1970s defined an algebra with a well-founded semantics for asking questions about the data stored in relational databases (Codd 1970). He defined tables as the “core concept” of data in databases and then defined five “core computations” on tables, namely selection, projection, Cartesian product, set union, and set difference (Codd 1970). These operations form the Relational Algebra that is the foundation for the Structured Query Language (SQL). SQL allows users to access data by `SELECTing` attributes `FROM` tables `WHERE` some conditions hold. This is a purely declarative approach, focusing on questions and answers rather than (at the user level) on procedures to operate on data. We are seeking a similar approach for GIS, enabling users to ask questions without concerning themselves about particular spatial data structures and their procedures.

To achieve this longer term goal, we propose an abstraction on GIS information contents, borrowing the concept of Abstract Data Types (ADT) from computer science (Liskov and Zilles 1974) and applying it to the user level. An ADT has been defined as a “class of objects whose logical behavior is defined by a set of values and a set of operations” (Dale and Walker 1996). ADT specifications capture a wide range of implementations, but users of the ADT only need to understand the specification, rather than the implementation, to know how instances of the type behave (Guttag and Horning 1978). Relational algebra can be seen as a user-level ADT for databases. Applying this idea to GIS, we use the core concepts of spatial information as classes of objects and define core computations as operations capturing their behavior. These classes and operations can then be used to ask and answer spatial questions; each concept has a set of specific computations, answering questions and defining its behavior.

### 3 Related Work

To the best of our knowledge, the realization of a question-based approach to spatial computing with user-level abstract data types has no direct predecessors or competitors, although there are a number of related initiatives and technologies that deserve discussion.

The communities of the Semantic Web and of Applied Ontology<sup>2</sup> have produced innovative ways to express and query spatial information, mainly based on the paradigm of Linked Data (Kuhn et al. 2014). Analytical capabilities, however, are not normally available with these approaches. Geospatial ontologies specify key application domains and data, e.g., land cover or hydrology (Ahlqvist 2005; Ames et al. 2012), but without the level of generalization sought here. The deeper semantic questions, concerning what contents users want to talk about and manipulate at a GIS user interface, have hardly been addressed so far in the form of implemented approaches. More generally, there is still no commonly accepted classification of the types of spatial (or even just geographic) information—contrasting with the vast literature on data types and structures to store that information.

Standardization has mainly focused on service-based data exchange, such as through the Geography Markup Language GML,<sup>3</sup> to some extent abandoning the original aspiration of defining spatial computing services independently of encoding formats (Kottman 2001). The idea of a software-independent essential model of geospatial computing (Cook and Daniels 1994) was not pursued, and the language of geographic information standards became one of software technologies and data transfers, rather than one of spatial information contents, questions, and computations. While this may have been a good choice for existing GIS technology vendors and user communities, it resulted in voluminous standards with heavy doses of jargon and acronyms, severely limiting outreach to and adoption by new communities.

One reason for this situation is the lack of theory on how to structure spatial computations in order to serve actual application needs. Attempts at devising a theoretical basis for GIS operations (Tomlin 1990b; Albrecht 1998) produced useful groupings of GIS functions, but have not attained a theoretical power and simplicity comparable to Codd's relational algebra for databases. Thus, the obligation to improve the situation is primarily that of researchers, not of vendors and standards bodies.

Relational algebra itself, although efficient for tabular data, is not sufficient for spatial data, since the nature and structure of spatial data is different from that of tables. Spatial databases have been an essential component in GIS and related information systems for two decades now, providing efficient data structures to index and query spatial data. However, from the user's perspective, systems such as PostGIS,<sup>4</sup>

---

<sup>2</sup><http://semantic-web-journal.org>—All URLs cited in this article were accessed on December 4, 2015.

<sup>3</sup><http://www.opengeospatial.org/standards/gml>.

<sup>4</sup><http://postgis.net>.

Oracle Spatial,<sup>5</sup> and Spatial Hadoop<sup>6</sup> offer only spatial extensions of SQL, forcing their users to formulate their questions in the language of tables, which does not sufficiently capture the spatial nature of the data. Specialized graph and array databases, such as SciDB,<sup>7</sup> can be used for spatial computing, but focus on one particular view of data, without providing the necessary additional abstractions for asking spatial questions independently of data structures. The same comment applies to the spatial extensions of the statistical package R (Lovelace and Cheshire 2014). Finally, spatial computing libraries such as GDAL<sup>8</sup> provide specific raster and vector views of contents. While they are powerful for complex data manipulations and format conversions, these tools tend to lock users into particular representational choices and burden them with implementation details.

By contrast, our mission is to enable question-based spatial computing through information content and quality abstractions. These abstractions are provided by the core concepts of spatial information and the computations on them, as summarized in the next section.

## 4 Core Concepts and Computations for Spatial Information

Core concepts of spatial information have been defined as concepts to interpret spatial data and to bridge the gap between spatial thinking and spatial computing (Kuhn and Ballatore 2015). They initially started as ten concepts and were later reduced to the following seven:

1. **Location**—The idea of locating something relative to something else, applicable to instances of the following four content concepts.
2. **Field**—A property with a value for each position in space and time.
3. **Object**—An individual that has properties and relations with other objects.
4. **Network**—A set of objects (nodes) linked by a binary relation (edges).
5. **Event**—Something that happens and involves fields, objects, and/or networks as participants.
6. **Granularity**—The level of detail in some spatial information.
7. **Accuracy**—The correspondence of some spatial information with what is considered a true state of affairs.

Location is a base concept, applicable to the next four concepts, which in turn can be seen as “content concepts.” Granularity and accuracy are “quality concepts” that can be applied to the base and content concepts, in order to set and assess the quality of

---

<sup>5</sup><http://www.oracle.com/database/big-data-spatial-and-graph>.

<sup>6</sup><http://spatialhadoop.cs.umn.edu>.

<sup>7</sup><http://www.paradigm4.com>.

<sup>8</sup><http://www.gdal.org>.



spatial information (see Kuhn 2012). The core concepts provide a conceptual foundation for spatial computing, which is still evolving. However, merely translating GIS commands and data structures into core concept terms would not yield useful results. Rather, GIS applications should be reorganized around questions posed in terms of core concepts.

For this purpose, we define a set of core computations for each core concept. These core computations should form a minimal yet powerful and complete set of operations that are applicable across different application domains. Defining a large, open set of operations would go against the central philosophy of the core concepts and would be similar to existing, bloated GIS command sets. Core computations allow for extension by combining with computations of other core concepts and they get implemented through existing GIS functions; this allows for a small set of computations. The definition of core computations is ongoing research, and the initial set formalized in Kuhn and Ballatore (2015) has changed. A new field operations, called *restrict domain* and a new granularity operation, called *coarsen*, will be introduced in Sect. 5.2.

Defining core concepts and core computations in this way creates Abstract Data Types (ADTs) for GIS users. This kind of data abstraction will enable users to formulate spatial questions to be answered by existing GIS. The real world entities captured in a GIS application can be seen as instances of the different concepts, which then suggests computations to be applied in GIS projects.

Spatial questions may be answered directly with one or a set of core computations, but in most cases they need to be decomposed into simpler questions, each of which could be answered by one or a set of operations. This process can be seen as a case of *problem decomposition*, which in existing GIS platforms has to be performed in terms of data structure manipulations. By contrast, with the core concepts, users can ask questions and find answers by using a smaller set of computations organized around spatial concepts, not their representations. This difference is the essence of our approach to question-based computing.

The goal of our work is not to reinvent or replace existing GIS or any other spatial computing platforms, but to make the computations easier to discover, understand, and use. From an object-oriented implementation perspective, core concepts and core computations can act as wrappers for existing GIS and library functions. To achieve this goal, we specify and test the core computations in a Python implementation, articulating our approach through a real-world case study.

## 5 A Case Study from Economics

MIT economist Matt Lowe provides a striking recent example of the difficulties and frustrations that a domain scientist encounters when using GIS (Lowe 2014). Lowe uses nocturnal luminosity observed by satellites as a proxy measure of development and welfare, noting that light density has a fairly strong correlation with local economic activity.

He proposes computations on four pieces of spatial information: (1) satellite imagery that captures global nocturnal luminosity, (2) a global map of gas flares, (3) the boundaries of all the countries in the world, and (4) roads in China. Gas flares are of industrial origin and outside of Lowe's economic scope of interest, so they are intentionally excluded from the luminosity analysis. As for the data layers, the first is imported as raster data and the following three are imported as vector data.

Ultimately, Lowe wants to answer the spatial question: *What was the level of economic activity near roads in China in 1994, based on nighttime luminosity as a proxy?* In his description of the ArcGIS procedure he has used, and based on the ArcPy scripting library, Lowe repeatedly puzzles over the amount of complex details involved in executing this seemingly simple spatial analysis (Lowe 2014). In order to address the reasons for this complexity and then a remedy, we first present Lowe's procedural approach and then our alternative approach based on core concepts driven by spatial questions.

## 5.1 Solution with Procedural GIS

The procedural steps that Lowe uses to answer his spatial question are as follows:

1. Erase gas flares from the data of countries
2. Calculate the average luminosity
3. Clip the luminosity data (output of 2) to the output of step 1
4. Extract the data in China from the output of step 3
5. Create a buffer around the China roads
6. Clip the output of step 4 to the extent of buffered roads
7. Clip the luminosity data to the extent of the output of step 6
8. Create a grid
9. Apply a zonal mean operator to each grid cell to calculate mean luminosity

This recipe of nine steps is a typical example of the procedural approach currently necessitated by spatial computing: in order to obtain an answer to a question, users have to apply a sequence of steps on specific data formats, which are not trivial to locate in the complex toolboxes of modern GIS.

An examination of Lowe's analysis illustrates how users without substantial expertise have difficulty finding, understanding, and correctly applying GIS procedures. Some of Lowe's analysis choices reveal confusions or inefficiencies that a GIS user might face. For instance, to determine the luminosity inside China, excluding the gas flares, he first erases the gas flares from the global map of countries, then clips the luminosity data of the world to the output of the previous process, and again clips the resulting data set to the extent of China. Whereas first clipping the global map of countries to the extent of China and then clipping the luminosity data to the extent of China would have been a more expedient technique.



**Fig. 1** The final result of Lowe's ArcPy procedure

The final results of Lowe's procedure is shown in Fig. 1. Dark areas have no luminosity, while light areas have high luminosity. Cross-hatched areas are outside of the road buffer and have been excluded from analysis.

## 5.2 *Solution Based on Spatial Questions*

In order to demonstrate the potential of core concepts for simplifying spatial computing, we propose an alternative approach to answer Lowe's question. The main question is decomposed into three smaller questions that can be answered with one or two computations each. This captures the idea of question-based computing, in which computations are designed to answer specific spatial questions and are therefore more "declarative" than procedural. Figure 2 shows the ArcPy scripts developed by Lowe on the left with the alternative solution based on core concepts, also written in Python, on the right.

We utilize Python for several reasons. Python is popular among application programmers and users, has the capacity to implement ADTs, and is interoperable with existing GIS tools and libraries. For the back-end implementation of our approach, we work with ArcPy to keep consistency with the original solution and to compare

ArcPy	Core Concepts
<pre> ArcPy  # Get China less gas flares polygon arcpy.Select_analysis("countries_nogas", "china1.shp",     "\NAME\" = 'China'") # Average two satellites for 1994 outRaster = (Float("F101994")+Float("F121994"))/2 outRaster.save("FX1994")  # Use buffer tool and roads to make polygon of China # close to roads, then clip china1 to this arcpy.Buffer_analysis("a2010_final_proj", "roadbuff.shp", "0.5 DecimalDegrees", "FULL", "ROUND", "ALL", "") arcpy.Clip_analysis("H:/Research/Data/Lights/china1.shp",     "H:/Research/Data/Lights/roadbuff.shp", "china2.shp", "")  # Clip each lights raster to extent of china2 rasterList = arcpy.ListRasters("F**") for raster in rasterList:     arcpy.Clip_management(raster, "-179.9999 -90.0 180.0 83.62741", "G"+str(raster[1:]),     "H:/Research/Data/Lights/china2.shp", "",     "ClippingGeometry")  # Create grid to extent of one of new light rasters arcpy.CreateFishnet_management("ch_grid.shp", "73.55416 18.15416", "73.5541 28.15416", "0.1", "0.1", "0", "0",     "134.77916 53.5625", "NO_LABELS", "G101992", "POLYGON") arcpy.RasterToPolygon_conversion("G101992", "G101992p.shp",     "NO_SIMPLIFY", "Value")  # Process: Clip grid to perimeter of polygon arcpy.Clip_analysis("H:/Research/Data/Lights/ch_grid.shp",     "H:/Research/Data/Lights/G101992p.shp", "china_grid.shp",     "")  # Zonal statistics on each year rasterList = arcpy.ListRasters("G**") for raster in rasterList:     arcpy.gp.ZonalStatisticsAsTable_sa("H:/Research/Data/Lights/ china_grid.shp", "FID", raster,     "1"+str(raster[5:])+".dbf", "DATA", "MEAN")                 </pre>	<pre> Core Concepts  # load input data china_boundary = MakeObject('data/china.shp') china_lights_1 = MakeField('data/lights_1994a.tif')     .restrict_domain(china_boundary, 'inside') china_lights_2 = MakeField('data/lights_1994b.tif')     .restrict_domain(china_boundary, 'inside') gas_flares = MakeObject('data/china_flares.shp') roads = MakeObject('data/china_roads.shp')  # What is the luminosity in year 1994 in China, excluding gas flares? average_luminosity = china_lights_1     .local(china_lights_2, 'average') luminosity_noGas = average_luminosity     .restrict_domain(gas_flares, 'outside')  # What is the luminosity within 0.5 degrees from roads? roads_buffered = buffer(roads, 0.5, 'Decimal Degrees') luminosity_around_roads = luminosity_noGas     .restrict_domain(roads_buffered, 'inside')  # What is the mean luminosity in a 0.1 by 0.1 degree area? final_results = luminosity_around_roads.coarsen(0.1,0.1)                 </pre>

**Fig. 2** Comparison between the procedural and core concepts solutions to the China luminosity study (Lowe 2014)

on a fair and relevant basis. In fact, the computations used here, act as wrapper for ArcPy functions.

Before starting the actual computations, the preliminary step is always to provide the required input data. This process is conventionally done by loading data layers as raster or vector data into the chosen GIS. This perspective tends to lock users into conceptualizations associated with the data formats, limiting the set of available computations to those defined for each data type and contributing to unnecessary format conversions. For instance, in order to clip the grid to the extent of the luminosity data in China, Lowe had to convert the luminosity raster to vector data. The decision to convert from raster to vector is encouraged by the chosen GIS framework and not inherent (in fact rather contrary) to the question.

Alternatively, core concepts allow users to decide which concept is most appropriate for framing the problem and interpreting the data sets, given the questions to be answered. For instance, a user could treat the Chinese road data as a network or instead as set of road objects. The only computation that needs to be performed on the road data involves determining which areas lie within a certain distance from roads. Therefore, the most appropriate (simplest) choice is to treat the roads as a set of objects to which a buffer operation can be applied. Thus, the road data are interpreted and subsequently loaded as a set of objects.

Similarly, China and the gas flares are each interpreted as objects and the two raster files containing the luminosity data for 1994 as fields. The following Python code is used to load these data. In Lowe’s solution, these operations were performed in a pre-processing stage.

```
china_boundary = makeObject('data/china.shp')
china_lights_1 = makeField('data/lights_1994a.tif')
                  .restrict_domain(china_boundary, 'inside')
china_lights_2 = makeField('data/lights_1994b.tif')
                  .restrict_domain(china_boundary, 'inside')
gas_flares = makeObject('data/china_flares.shp')
roads = makeObject('data/china_roads.shp')
```

A considerable number of steps applied by Lowe (four out of nine) are spent on limiting the spatial coverage (extent) of his data. He uses “select by attribute” in the fourth step and “clipping” in the third, sixth, and seventh steps. Through the core concepts, limiting coverage is a much easier task achieved by restricting the domain of a field. By definition, each field has a domain for which it is defined (Kuhn and Ballatore 2015) and when creating a new instance of a field, its domain is inferred from the extent of the loaded data. Since clipping in fact only reduces the domain of a field, it is conceptually easier to restrict the domain by the new boundary, avoiding the need to modify or create any values. Our `restrict_domain` computation restricts the domain of a field in this way. In fact, there is not even a need for any computations in such a domain restriction, as one can just add or subtract spatial extents to the domain definition. In the second and the third commands of the above code, the domain of the fields (that are being created from a tif file) are restricted to the inside of the boundary of China, at the time of creation.

Once the data have been interpreted and loaded as core concept instances, we can start asking spatial questions. Lowe’s overall question is naturally subdivided into three simpler questions:

- What is the average luminosity in 1994 in China, excluding gas flares?
- What is the luminosity within 0.5° from roads?
- What is the mean luminosity at a coarser (0.1 by 0.1°) granularity?

In order to answer the first question, the luminosity in the year 1994 is averaged from two satellite data sources for that year. Then the domain of this field gets restricted to the region outside of the gas flares. As the luminosity data have been interpreted as fields, a local field operation can be used for averaging the values. Local operations are well-known from map algebra (Tomlin 1990a). To restrict the domain of the resulting field to the outside of gas flares, the `restrict_domain` function is applied.

```
average_luminosity = china_lights_1
                    .local(china_lights_2, 'average')
luminosity_noGas = average_luminosity
                  .restrict_domain(gas_flares, 'outside')
```

Answering the second question requires determining the luminosity near roads (within a 0.5° buffer). Buffering is a core computation on objects, requiring objects and the buffer distance as input. After applying it, the domain of the luminosity field can be reduced to the buffered region.

```
roads_buffered = roads.buffer(0.5, 'DecimalDegrees')
luminosity_around_roads = luminosity_noGas
    .restrict_domain(roads_buffered, 'inside')
```

To address the third question, the luminosity field can simply be coarsened. Lowe defined a vector grid to aggregate the raster into larger cells. This approach is computationally expensive and necessitates a raster to vector conversion on the luminosity data. Our alternative solution uses the concept of granularity, as a quality concept applicable to all content concepts, including fields. Two core computations for granularity, *coarsen* and *refine*, allow for decreasing and increasing, respectively, the granularity of fields, objects, networks, and events (here shown for the case of fields):

```
final_results = luminosity_around_roads.coarsen(0.1, 0.1)
```

To do the same process in ArcPy, Lowe had to undertake a multi-step process: convert the luminosity raster to a polygon layer, clip the luminosity polygons to the extent of the grid, and then apply zonal statistics.

### 5.3 Implementation

To implement the core computations, we have constructed an hierarchy of abstract and concrete classes that leverages the classes defined by Kuhn and Ballatore (2015) and existing spatial computation platforms (currently only ArcPy). Abstract class constructors generate subclass instances to appropriately type and handle the loaded data. This removes common but unnecessary choices and parameters from user commands. The CcField abstract class, for example, points to the GeoTiffField subclass (hidden from the user), an instance of which is generated when a user loads “.tiff” data as a field. Within the same GeoTiffField subclass, functions for retrieving the domain and for reading in raster data allow `local` to execute and optimize the map algebra local operation in a number of ways (see Sect. 5.2).

### 5.4 Comparison and Discussion

Our core concept solution is not replicating the step-wise procedure that Lowe had to undertake, which was chaining together tools operating on data formats. It is also an improvement over a standalone use of ArcPy, which restricts users to fragmented step-wise operations on raster and vector data. As shown, the procedure used by Lowe contained nine analytical steps, some consisting of multiple operations. In

contrast, our method based on core concepts has only three analytical steps, with five operations in total, thus being 45 % shorter overall. A further improvement is that our computations should be simpler to understand, as they contain fewer parameters, the user does not have to deal with format conversions, and the steps follow a logical order established by the three sub-questions. We also expect overall command spaces to end up smaller and more usefully organized by the core concepts, though the current state of our work cannot yet demonstrate this effect.

## 6 Conclusions

This paper proposed a new approach to spatial computing based on asking questions rather than performing data manipulation procedures. Inspired by the notion of Abstract Data Types (ADT) from computer science, and leveraging the idea of the core concepts of spatial information and the language designed based on them, our approach allows for answering spatial questions by using core computations defined for each core concept.

In a first attempt of putting this approach into practice, we used a case study from economics, in which a domain expert had used ArcGIS (through its scripting language, ArcPy), to determine aggregated nocturnal luminosity in China near roads. We answered this spatial question by interpreting the input data in terms of core concepts and performing core computations on them. We see the case study as a realistic scenario to test the idea of question-based spatial computing and the underlying core concepts and computations for spatial information. Despite the encouraging results, this case study is only an initial step, and more real-world scenarios are needed to demonstrate the power and to identify the potential weaknesses of our approach.

Furthermore, a spatial question can either be answered directly by a core computation or the user decomposes it into smaller, answerable questions. Formulating a method to decompose spatial questions into smaller units will be pursued in future research, while further developing the core computations. We believe that this is an ambitious but achievable goal that will expand the range of spatial questions answerable through the core concepts and computations, without the need to grow their number.

This research also helped us extend and improve the formal specifications of core concepts and modify the computations defined on them. For instance, the granularity concept did not have a rigid formal specification beforehand, but now has a clearer definition and two core computations (refine and coarsen). The method proposed in this paper ends up being remarkably shorter than the published GIS solution and appears easier to understand. It uses just over half the number of computations used by Lowe.

Our ADTs were implemented in Python, currently using the ArcPy library, because of Python's interoperability with existing GIS tools and its growing usage within and outside GIScience. The core computations implemented for each concept are not yet complete, and defining new operations as well as modifying existing

ones continues, based on additional case studies. Also, the method used in this paper has its limitations, as some of the operations were not fully compatible with ArcPy, requiring several workarounds. This problem will be addressed in future work leveraging multiple platforms beyond ArcPy, including commercial and open source GIS. Another line of future research will study transitions from one core concept to another, and their implementation as constructor functions.

We foresee our research to eventually lead to an open source Application Programming Interface (API) for spatial computing, based on core concepts and computations and dispatching to commercial or open source spatial computing platforms. This API will provide higher-level access to existing commands on these platforms and thus support a broader community of users, ranging from GIS experts to domain users or application programmers without technical GIS training.

**Acknowledgments** We gratefully acknowledge the contributions of Thomas Hervey, Sara Lafia, Michael Wang, and others at the UCSB Center for Spatial Studies for helping shape and refine this idea and its implementation. We also acknowledge Professors Rich Wolski and Chandra Krintz from the Computer Science department at UCSB, who have been challenging us to apply the question-based approach to this kind of case study. We thank the anonymous reviewers for their insightful comments, which led to improvements in the paper.

## References

- Ahlqvist O (2005) Using uncertain conceptual spaces to translate between land cover categories. *Int J Geogr Inf Sci* 19(7):831–857
- Albrecht J (1998) Universal analytical GIS operations: a task-oriented systematization of data structure-independent GIS functionality. In: Onsrud H, Craglia M (eds) *Geographic information research: transatlantic perspectives*. Taylor and Francis, pp 577–591
- Ames DP, Horsburgh JS, Cao Y, Kadlec J, Whiteaker T, Valentine D. Hydrodesktop: web services-based software for hydrologic data discovery, download, visualization, and analysis. *Environ Model Softw* 37
- Codd EF (1970) A relational model of data for large shared data banks. *Commun ACM* 13(6):377–387
- Cook S, Daniels J (1994) *Designing object systems*, vol 135. Prentice Hall, Englewood Cliffs
- Dalee N, Walker HM (1996) *Abstract data types: specifications, implementations, and applications*. Jones and Bartlett Learning
- Gao S, Goodchild MF (2013) Asking spatial questions to identify GIS functionality. In: 2013 fourth international conference on computing for geospatial research and application (COM. Geo). IEEE, pp 106–110
- Guttag JV, Horning JJ (1978) The algebraic specification of abstract data types. *Acta Informatica* 10(1):27–52
- Kottman C (2001) White paper on trends in the intersection of GIS and IT. Open GIS Consortium
- Kuhn W (2012) Core concepts of spatial information for transdisciplinary research. *Int J Geogr Inf Sci* 26(12):2267–2276
- Kuhn W, Ballatore A (2015) Designing a language for spatial computing. In: Bacao F, Santos MY, Painho M (eds) *AGILE 2015: geographic information science as an enabler of smarter cities and communities*. Springer, Berlin, pp 309–326
- Kuhn W, Kauppinen T, Janowicz K (2014) Linked data—A paradigm shift for geographic information science. In: *Geographic information science*. Springer, pp 173–186



- Liskov B, Zilles S (1974) Programming with abstract data types. In: ACM sigplan notices, vol 9. ACM, pp 50–59
- Lovelace R, Cheshire J (2014) Introduction to visualising spatial data in R
- Lowe M (2014) Night lights and ArcGIS: a brief guide. [Online; Accessed Nov-2015] <http://economics.mit.edu/files/8945>
- Tomlin CD (1990a) A map algebra. Harvard Graduate School of Design
- Tomlin DC (1990b) Geographic information systems and cartographic modeling. Prentice Hall, Englewood Cliffs
- Tomlinson RF (2007) Thinking about GIS: geographic information system planning for managers. ESRI, Inc

# Measuring Space-Time Prism Similarity Through Temporal Profile Curves

Harvey J. Miller, Martin Raubal and Young Jaegal

**Abstract** Space-time paths and prisms based on the time geographic framework model actual (empirical or simulated) and potential mobility, respectively. There are well-established methods for quantitatively measuring similarity between space-time paths, including dynamic time warping and edit-distance functions. However, there are no corresponding measures for comparing space-time prisms. Analogous to path similarity, space-time prism similarity measures can support comparison of individual accessibility, prism clustering methods and retrieving prisms similar to a reference prism from a mobility database. In this paper, we introduce a method to calculate space-time prism similarity through temporal sweeping. The sweeping method generates temporal profile curves summarizing dynamic prism geometry or semantic content over the time span of the prism's existence. Given these profile curves, we can apply existing path similarity methods to compare space-time prisms based on a specified geometric or semantic prism. This method can also be scaled to multiple prisms, and can be applied to prisms and paths simultaneously. We discuss the general approach and demonstrate the method for classic planar space-time prisms.

**Keywords** Activity space · Time geography · Similarity · Accessibility

---

H.J. Miller (✉) · Y. Jaegal  
The Ohio State University, Columbus, USA  
e-mail: miller.81@osu.edu

Y. Jaegal  
e-mail: jaegal.1@osu.edu

M. Raubal  
ETH Zurich, Zurich, Switzerland  
e-mail: mraubal@ethz.ch

# 1 Introduction

Space-time paths and prisms based on the time geographic framework model actual and potential movement, respectively, of an object through geographic space with respect to time (Hägerstrand 1970). *Path similarity measures* quantitatively assess the resemblance between two space-time paths, with greater similarity indicating mobility and activity patterns that are more alike. Path similarity can provide insights into dynamic phenomena such as traffic congestion and crime (Yuan and Raubal 2014), identify similar patterns of environmental exposure (Briggs 2005; Sinha and Mark 2005), and analyze collective movement patterns (Gudmundsson et al. 2012). Other applications of path similarity measures include path clustering (finding groups of similar paths), path aggregation (forming composite representative paths), and mobile objects database (MOD) queries to find paths that resemble a reference path.

The *space-time prism* (STP) is the envelope of all possible space-time paths between two anchor locations with known departure and arrival times respectively. STPs are measures of space-time path uncertainty when a moving object's locations are undersampled with respect to time (Pfoser and Jensen 1999). They are also measures of individual accessibility and exposure within an environment and have been widely applied in human and ecological science (Espeter and Raubal 2009; Long and Nelson 2012). As extensions of space-time paths, STP similarity indicates similar patterns of accessibility and potential exposure within an environment. As with path similarity, we may wish to measure STP similarity to cluster or aggregate prisms as well as search for corresponding prisms within a MOD. However, no methods for measuring STP similarity exist.

This paper develops a time-based approach to measuring STP similarity. Our approach sweeps a STP with respect to time, recording its geometric and/or semantic properties at discrete moments. This allows us to construct one-dimensional temporal profile curves that summarize the properties with respect to time. These curves can be compared visually or quantitatively using existing path similarity measures. We restrict our attention to classic planar STPs in this paper but discuss how to extend these measures to other prism types, including network time prisms (Kuijpers and Othman 2009; Miller 1991) and field-based prisms (Miller and Bridwell 2009).

The next two sections of this paper provide background to our methods. Section 2 reviews existing approaches to measuring path similarity while Sect. 3 describes geometric and semantic properties of STPs that can be measured analytically. Section 4 presents our generic method for generating temporal profile curves. Section 5 provides an example using planar STPs. Section 6 identifies future research steps and Sect. 7 concludes the paper with a brief summary.

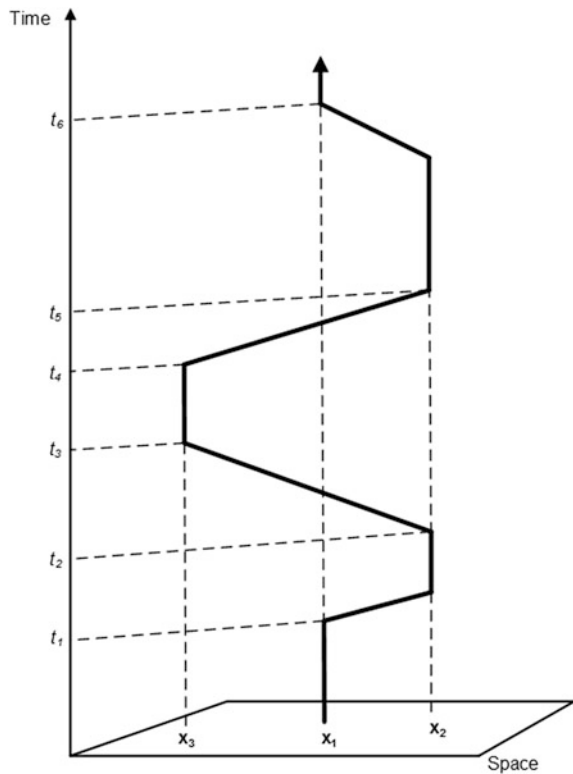
## 2 Space-Time Paths and Path Similarity

The space-time path represents a history of individual movement with visited locations, and sequence of the movement. Figure 1 shows a space-time path in two-dimensional space and time corresponding to a person moving among three locations with corresponding departure and arrival times at locations. The vertical line segments represent the duration that the person stayed at the same location for a certain amount of time. This integrated view of space and time provides an effective visual environment to understand human movement, activities and accessibility in space and time (Pred 1977).

In addition to visualization there is a wide range of analytical descriptions and summaries for space-time paths. Basic path measures include both *moment-based descriptors* (such as the time, location, direction and speed at any moment) and *interval-based descriptors* (such as the minimum, maximum and mean speed, the distribution and sequence of speeds and directions, and the geometric shape of the path over some time interval) (Andrienko et al. 2008).

*Path similarity measures* allow quantitative comparisons among space-time paths, particularly with respect to geometric similarity in space-time and with

**Fig. 1** A space-time path in two-dimensional space and time



respect to semantics. Geometric similarity captures the resemblance of the mobile objects in terms of their patterns in space with respect to time. Semantic similarity (Janowicz et al. 2011) refers to the characteristics and activities of the moving object, such as commuting, shopping and socializing (in the case of humans) (Raubal et al. 2004), or foraging and migration (in the case of animals). Similarity measures also support *path clustering and aggregation methods* for identifying synoptic spatio-temporal patterns from large collections of mobile objects (Long and Nelson 2013).

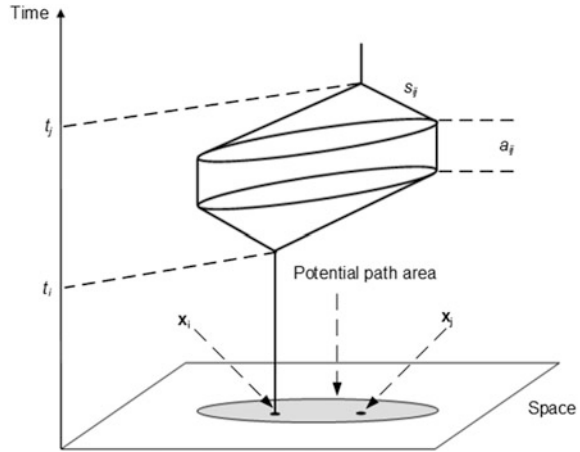
Two types of path similarity measures are shape-based measures and time-based measures (Yuan and Raubal 2014). *Shape-based measures* focus on the geometry of the paths; these include the average Euclidean distance between corresponding locations on the paths, and the Hausdorff distance or maximum of the minimum distances between the paths. *Time-based measures* take into account the temporal aspects of paths by considering them as multidimensional time series data. They include synchronized Euclidean distance, Fréchet distance, dynamic time warping, longest common subsequences, and edit-distance functions. Fréchet distance is the shortest of the set of closest distances that connects the objects moving along their paths at any speed without backtracking.

The time-based measures discussed above focus on path geometry including time stamps of points and order, but more general time-based measures such as dynamic time warping, longest common subsequences and edit-distance functions can capture both geometry and semantics. Required is some coding that translates path geometry and/or semantics into an ordered and exhaustive sequence of states with their time durations (see, for example, Dodge et al. 2012). *Dynamic time warping* measures the similarity between two sequences or trajectories by local stretching or compressing to match the time series and compute the sum of the paired distances (Yuan and Raubal 2012). *Longest common subsequence (LCSS)* methods measure similarity based on the length of the longest common subsequence in a set of sequences (Nanni et al. 2008). *Edit-distance functions* generalize LCSS: these measure similarities between sequential patterns based on the cost of the insertion, deletion and substitution operations required to transform one sequence into the other. Such functions can also account for spatial and temporal information in their cost functions (Yuan and Raubal 2014).

### 3 Analytical Space-Time Prisms

A space-time prism (STP) is the envelope of all possible paths in space with respect to time between two anchoring locations and corresponding departure and arrival times, subject to a maximum travel speed and any stationary activity time. STP parameters are  $\{\mathbf{x}_i, \mathbf{x}_j, t_i, t_j, s_{ij}, a_{ij}\}$  where  $\mathbf{x}_i, \mathbf{x}_j$  are the first and second anchor locations with associated departure and arrival times  $t_i, t_j$  respectively,  $s_{ij}$  is the maximum travel speed and  $a_{ij}$  is the stationary activity time (if any). The spatial

**Fig. 2** A planar space-time prism



footprint of a STP is the *potential path area* (PPA): in general, this is an ellipse with the two anchors as foci. Figure 2 illustrates a general STP.

Although it is the intersection of simple objects such as cones and cylinders (see Burns 1979), it is difficult to analytically describe the entire STP. However, it is easy to describe its spatial extent at a moment in time; this can serve as the basis for a wide range of prism analytics. At a moment in time  $t \in (t_i, t_j)$ , the spatial extent of a STP (denoted by  $Z_{ij}(t)$ ) is the intersection of three convex spatial sets: (i) the *future disc*  $f_i(t)$  comprising all locations that can be reached from the first anchor by time  $t_i + t$ ; (ii) the *past disc*  $p_j(t)$  encompassing all locations at time  $t$  that can reach the second anchor by time  $t_j - t$ ; and, (iii) the *potential path area*  $g_{ij}$  that constrains the prism locations to account for any stationary activity time:

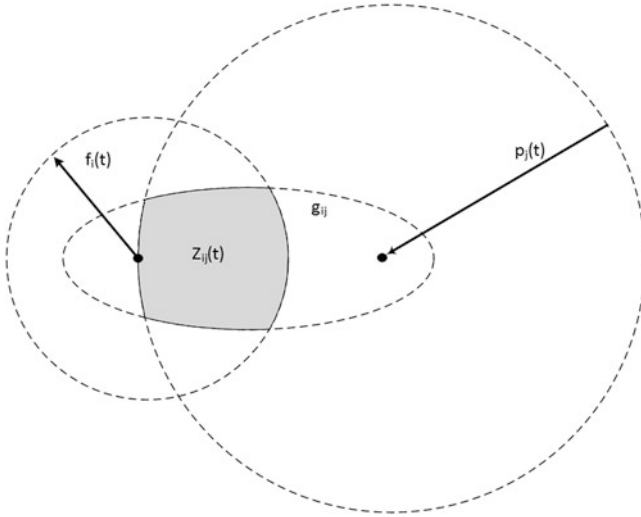
$$Z_{ij}(t) = \{f_i(t) \cap p_j(t) \cap g_{ij}\} \tag{1}$$

$$f_i(t) = \{\mathbf{x} \mid \|\mathbf{x} - \mathbf{x}_i\| \leq (t - t_i)s_{ij}\} \tag{2}$$

$$p_j(t) = \{\mathbf{x} \mid \|\mathbf{x} - \mathbf{x}_j\| \leq (t_j - t)s_{ij}\} \tag{3}$$

$$g_{ij} = \{\mathbf{x} \mid \|\mathbf{x} - \mathbf{x}_i\| + \|\mathbf{x} - \mathbf{x}_j\| \leq (t_j - t_i - a_{ij})s_{ij}\} \tag{4}$$

Figure 3 provides an illustration. This definition of the STP is not limited to two-dimensional space. In one-dimensional space, the sets described by Eqs. (2)–(4) are line segments. In two-dimensional space, the discs are circles and the geo-ellipse is an ellipse. In three-dimensional space, the discs are spheres and the geo-ellipse is a spheroid. There are scalable methods for calculating these objects and their intersections. Also, the intersection geometry of a STP at a moment in time never requires finding the intersection of all three spatial sets since the future and past disc change size and can be enclosed by the other two sets for part of the prism’s existence. This will be discussed in more detail below (Miller 2005).



**Fig. 3** Analytical construction of a STP at a moment in time

In addition to the prism boundaries, the prism interior also has *intrinsic* and *extrinsic* properties of interest. Intrinsic properties relate to the prism itself. The probabilities of the object visiting different locations within the prism interior are not equal: the object is more likely to visit locations near the axis connecting the two anchors relative to locations near the prism boundaries since there are more possible paths through the former rather than the latter (Winter and Yin 2010a). We can simulate the visit probability distribution within a prism interior using simple random walks or Brownian Bridges methods, truncated to account for STP constraints (Song and Miller 2014). Alternatively, the prism visit probability distribution at a moment in time can be approximated using a clipped bivariate normal distribution (Winter and Yin 2010b).

Related to the visit probability distribution within the prism interior is the distribution of possible speeds at each location at a moment in time. The speed distribution at a given location and time within the prism is Markovian in the sense that the history that precedes it is irrelevant: the location and remaining time determines the possible speeds. Locations near the prism interior tend to have a wider range of possible speeds while locations near the boundaries are more constrained, with locations on the prism boundary constrained to only the maximum speed.

Extrinsic properties of the prism interior relate to type of activity locations, resources, environmental features and other individuals encompassed by the spatial region. This external content describes the possible activities and experiences for the individual and therefore supports derivation of accessibility semantics. At a moment in time these properties comprise a two-dimensional spatial distribution that can be described using spatial statistics. Point pattern measures including density-based approaches or distance-based measures, such as the K function that is based on counting points within a series of distances of each point (O'Sullivan and

Unwin 2010), can be applied to a spatial distribution of activity locations such as restaurants or movie theaters. With entropy measures such as spatial entropy, one can quantitatively determine the uncertainties about the structure of two-dimensional spatial distributions (Batty 2010). Furthermore, spatial autocorrelation can be measured for these distributions, for example, through indices such as Moran’s I, which is usually applied to areal units representing environmental or epidemiological data (O’Sullivan and Unwin 2010).

Other properties of a prism include its relationship with other space-time paths and prisms. Calculating the binary query if a path lies in a prism at a given moment in time only requires testing if a point (the path at a moment in time) lies within a disc, ellipse, or a disc–ellipse intersection. We also can easily test if a prism-prism intersection exists at a given moment in time by evaluating a small set of linear inequalities. More complicated but still tractable is solving for the intersection region at a given moment in time: this requires solving for the intersection of two, three, or four simple spatial sets based on the prisms’ morphologies at that moment. The worse-case for two prisms is a four-set intersection involving two discs and two ellipses (Miller 2005). Finally, it is also possible to solve for the Euclidean, Hausdorff and other distances between two prisms at a moment in time since these are simple spatial sets.

#### 4 Measuring Space-Time Prism Similarity Using Temporal Profile Curves

The basic idea behind our method is to reduce the dimensionality of the space-time prism by sweeping it with respect to time and summarizing its geometric and/or semantic properties at discrete moments in time using the methods outlined in general above. This generates a temporal profile curve for the given attribute. These

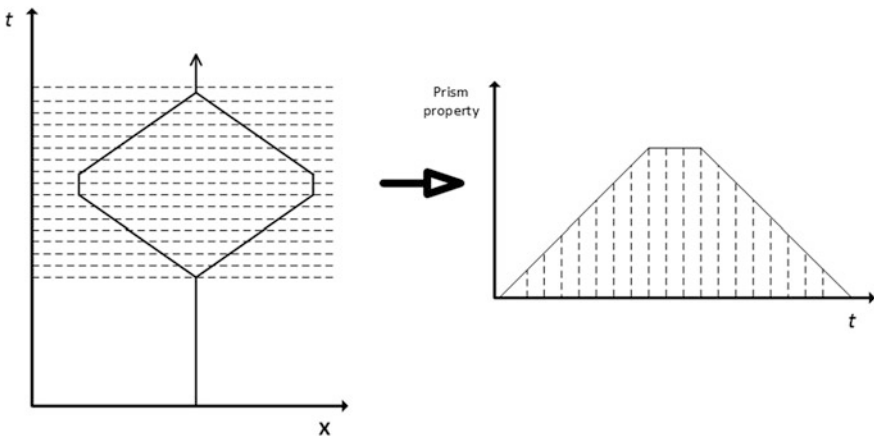


Fig. 4 Temporally sweeping a space-time prism



profile curves can be compared visually. We can also apply existing path similarity measures to determine their resemblance based on the chosen parameter, as well as apply clustering and aggregation methods based on these curves. Figure 4 illustrates the core idea for a single STP. It is also possible to sweep multiple STPs simultaneously to compare their properties within the same time frame. We could also sweep multiple STPs independently and compare the prisms after normalizing the time horizons of each prism.

As noted above, the intersection geometry of a STP at a moment in time never requires finding the intersection of all three spatial sets since the future and past disc change size and may be enclosed by the other two sets during subintervals of the prism’s existence. With a general prism as in Fig. 2 we only need to solve (in the following order) the future disc alone, the intersection of the future disc with the potential path ellipse, the potential path ellipse alone, the intersection of the potential path ellipse with the past disc intersection and finally the past disc alone (Miller 2005). Figure 5 illustrates these subintervals for a general prism. The temporal subinterval boundaries are:

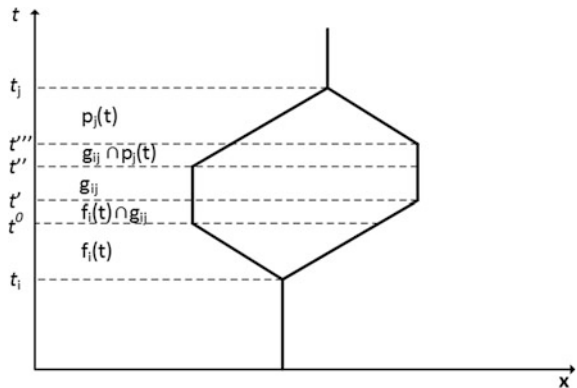
$$t^0 = \frac{(t_i + t_j - t_{ij}^* - a_{ij})}{2} \tag{5}$$

$$t' = \frac{(t_i + t_j - t_{ij}^*)}{2} \tag{6}$$

$$t'' = \frac{(t_i + t_j + t_{ij}^*)}{2} \tag{7}$$

$$t''' = \frac{(t_i + t_j + t_{ij}^* + a_{ij})}{2} \tag{8}$$

**Fig. 5** Temporal subintervals for analytically calculating planar STPs at a moment in time



where  $t_{ij}^*$  is the minimum travel time between the anchors. If stationary activity time  $a_{ij}$  is zero then Eqs. (5) and (8) are irrelevant and the prism simplifies to only three subintervals that require solving for the future disc, future disc-past disc intersection and the past disc, respectively. Forming these sets and intersections are simple operations that can be performed using standard buffering and overlay techniques available in most GIS software shortcuts. In the examples below, we calculate these intersections and their properties at discrete moments in time using off-the-shelf overlay tools available in ArcGIS 10.1.

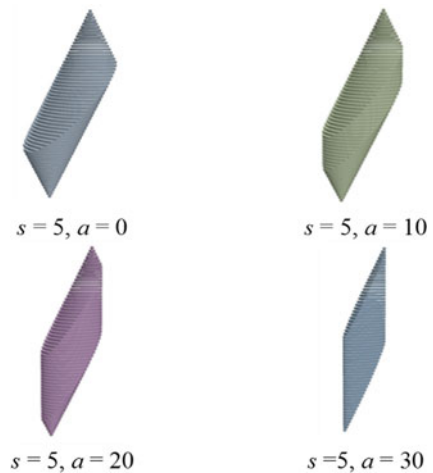
### 5 Examples

This section illustrates the temporal sweeping method for planar STPs. We first demonstrate the technique for generating temporal profile curves for summarizing geometric prism properties, specifically, prism area. We then demonstrate the technique for summarizing prism semantics for an empirical example.

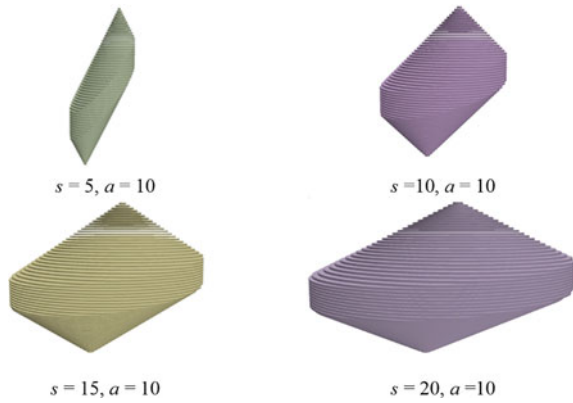
**Geometric similarity.** Figure 6 provides four prisms with varying activity times but speed limits fixed at  $s = 5$ . Figure 7 provides four prisms with varying speed limits but with stationary activity times fixed at  $a = 10$ . Stationary activity times refer to minimum times required for immobile activity, such as shopping or dining. All prisms have anchors at  $(0, 0)$  and  $(100, 100)$  and time budgets of 60. Also note that the prism with  $s = 5$  and  $a = 10$  is the same in both figures.

Figures 8 and 9 provide the corresponding temporal profile curves for the prism areas in Figs. 6 and 7, respectively. As expected, lower activity times and higher speeds correspond to larger prism areas and shifts in the locations of the curves positively with respect to the y-axis. Changes in STP speed limit have a larger impact on the locations of the curves relative to changes in activity time (note the

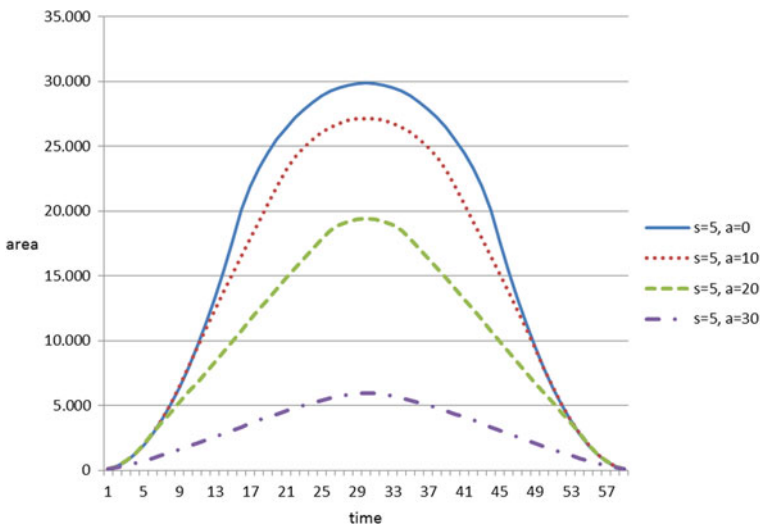
**Fig. 6** Four prisms with varying activity times



**Fig. 7** Four prisms with varying speed limits



difference in y-axis scale between the two curves). This corresponds to the more dramatic changes in STP size evident from comparing Figs. 6 and 7, as well as time geographic theory that suggests relaxing speed limits has a bigger impact on accessibility than reducing stationary activity time (see Burns 1979). However, in addition to shifts in the curve locations there are changes in the curve morphology. Figure 8 indicates that a lower activity time results in the shapes of the curves to become more rounded. In contrast, Fig. 9 indicates that higher speed limits correspond to the curves becoming more peaked. These differences in profile curve locations and morphology can be exploited when measuring prism similarity or performing related analysis such as prism clustering and aggregation.



**Fig. 8** Profile curves for the varying activity time prisms in Fig. 6

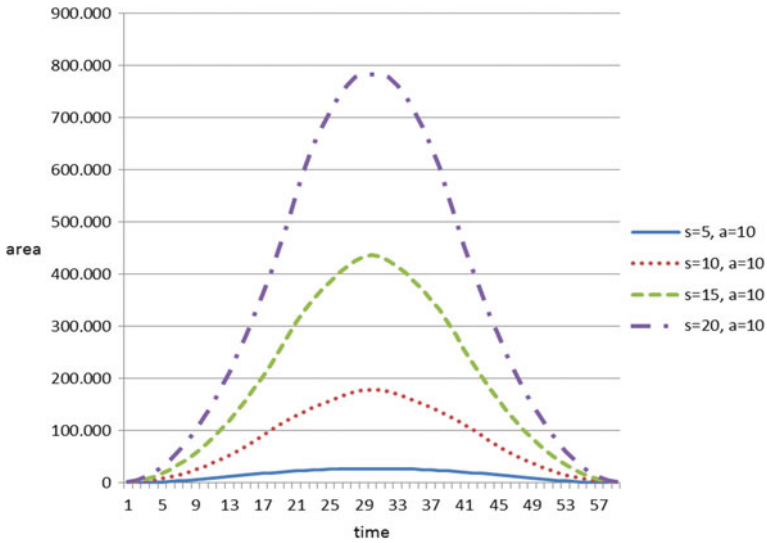


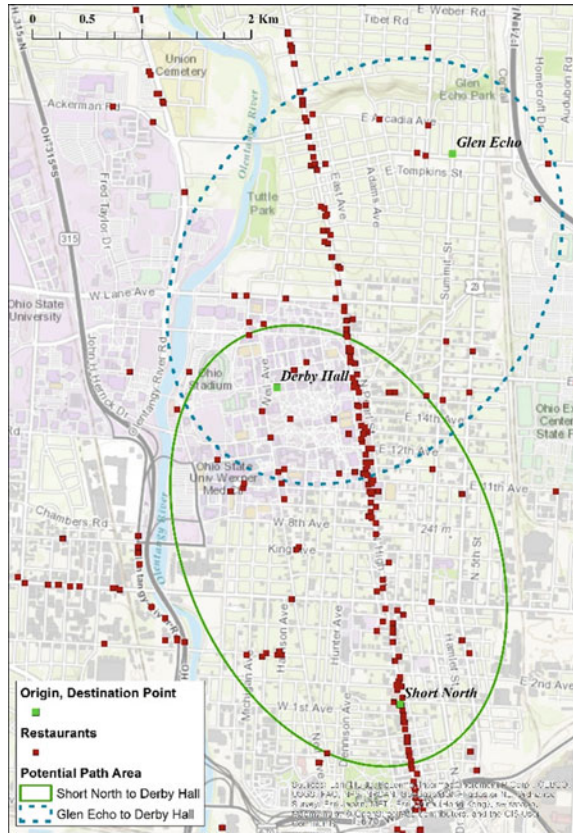
Fig. 9 Profile curves for the varying speed limit prisms in Fig. 7

Table 1 provides a summary of inter-curve distances calculated via the dynamic time warping (DTW) procedure in R using the Sakoe-Chiba band window with a maximum time deviation between matched pairs of 4 (Giorgino 2009; Sakoe and Chiba 1978). We normalized these distances using the symmetric2 procedure in R: this gives a higher weight to diagonal transition relative to horizontal or vertical transitions (see Giorgino 2009). As Table 1 indicates, the DTW distances distinguish between STPs with different activity times and speed limits. However, distances between STPs with different speed limits are greater than STPs with different activity times due to the greater effect of speed limit than activity time on STP area. This is not necessarily true of all STP geometric properties; an open research question is to identify and assess STP geometric properties with respect to changes in the STP parameters (see Burns 1979). Also, DTW captures differences in curve locations better than curve morphology; another open research question is determining curve similarity measures that can capture morphological differences.

Table 1 DTW inter-curve distances

Speed limit changes	Activity time changes			
	s = 5, a = 0	s = 5, a = 10	s = 5, a = 20	s = 5, a = 30
s = 5, a = 10	528.4	–	2,501.2	9,441.9
s = 10, a = 10	51,092.3	53,026.6	58,132.1	65,820.8
s = 15, a = 10	145,217.3	147,151.6	152,264.4	160,036.4
s = 20, a = 10	277,258.9	279,193.2	284,306.2	292,114.5

**Fig. 10** Example for semantic similarity of two prisms (i) Short North to Derby Hall; (ii) Glen Echo to Derby Hall



**Semantic similarity.** Figure 10 provides an empirical example for calculating semantic similarity: walking to Derby Hall on The Ohio State University campus from two different neighborhoods in Columbus, Ohio, USA. The Short North neighborhood to the south is a denser urban setting while Glenn Echo to the northeast is a more suburban neighborhood. Figure 10 shows the potential path areas for two prisms with origin anchors in the centers of the Short North and Glen Echo neighborhoods but with a common destination anchor at Derby Hall. Both prisms reflect a speed limit of 6.4 kph (a brisk walk) with a total time budget of one hour, and stationary activity time of 30 min for stopping at a restaurant. The map also shows the location of all restaurants in the respective areas.

Figures 11 and 12 show the prisms' profile curves for two semantic properties: the restaurant density (relative to prism area at each moment in time) and the average nearest neighbor ratio calculated as the observed average distance divided by the expected average distance based on a null model of complete spatial randomness. In Fig. 12, values less than 1 indicate spatial clustering while values greater than 1 indicate spatial dispersion.

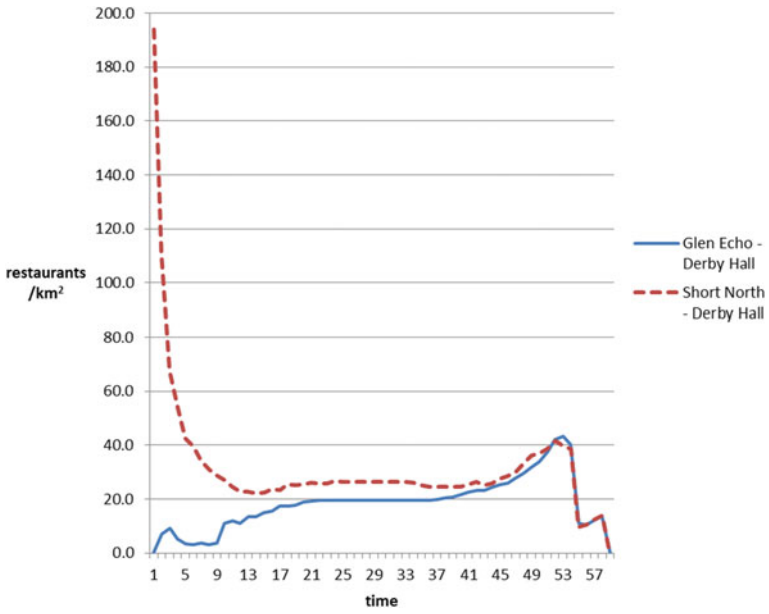


Fig. 11 Prism profile curves for restaurant density

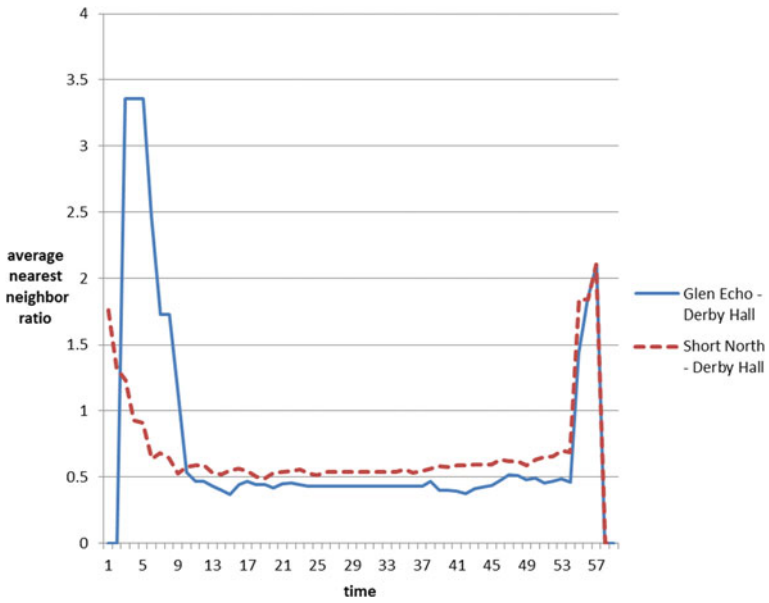


Fig. 12 Prism profile curves for average nearest neighbor ratio

Figures 11 and 12 suggest that the profile curves distinguish well between the semantic content of the two prisms. The profile curve for the Short North prism suggests an accessible environment with initially a high density of spatially clustered restaurants that becomes less dense and clustered with time and movement towards the destination. In contrast, the profile curve for the Glen Echo prism suggests an accessible environment with initially a low density of spatially dispersed restaurants that becomes denser and more clustered with time and movement towards the destination. In the last stages of both prisms the profile curves converge as the two prisms converge on the common destination. However, except for an early spike in the Glenn Echo nearest neighbor curve, the nearest neighborhood profiles appear more similar than the density profiles for both prisms. Normalized DTW distances for the profile curves of 10.1 and 0.2 for density and nearest neighbor, respectively, support this qualitative result.

## 6 Future Steps

The preliminary analysis in this paper suggests the value of the temporal profile curves for distinguishing among STPs. With respect to geometry, the STP area profiles indicate that changes in prism morphology can be reflected in changes in both the locations and shape of the profile curves. The semantic example of restaurant density and spatial clustering also generated profile curves that distinguish between the content of prisms. A next step is to explore different geometric and semantic STP indicators, such as shape measures (e.g., compactness versus elongation), physical properties such as average speed, and spatial statistics describing prism content at a moment in time, and assess their effectiveness at distinguishing among STPs under different conditions.

As noted earlier in this paper, similarity measures can facilitate the analysis of prism collections by supporting summarization methods. A next step after determining appropriate geometric and semantic indicators is to develop scalable STP clustering and aggregation methods. STP clustering is more straightforward than STP aggregation since the latter requires procedures for generating a composite STP that reflects the common properties of the disaggregated STPs. A simple approach for space-time paths is to use vector averaging: treat each segment of the polyline as a vector and find the average of the corresponding vectors (see Kobayashi and Miller 2014). Additional investigation is required to determine aggregation methods for the more complex case of STPs.

We restrict our attention to classic planar STPs in this paper. A longer term research task is to develop similar methods for the important case of network time prisms (NTPs). These methods can be based on graph theoretical measures of the spatial footprint at each moment in time in a NTP. Another possibility is to calculate geometric properties for the NTP using its trapezoidal and triangular regions in space and time (see Kuijpers and Othman 2009). With respect to NTP semantics, the profile curves can exploit address-matched and other network-referenced data

combined with network-based spatial analytical methods (Okabe and Sugihara 2012). Another future step is to develop methods for the more complex geometry of field-based prisms where speeds vary continuously in space. These methods can exploit properties of the lattice approximation for field-based prisms (Miller and Bridwell 2009).

## 7 Conclusion

This paper develops an approach to measuring space-time prism (STP) similarity in a manner similar to methods for measuring path similarity. Our method reduces the dimensionality of a STP by temporally sweeping it to generate one-dimensional profile curves that summarize changes in geometric and/or semantic properties with respect to time. We demonstrate this approach using the example of prism area under varying activity times and speed limits, as well as the semantic content for an empirical example of a travel and activity episode. Preliminary results suggest that this approach is promising: the locations and morphologies of the profile curves reflect changes in prism geometry and semantics, and differences between the curves can be summarized effectively using distance measures such as Dynamic Time Warping. We outline several next steps to continue this research, including determining effective geometric and semantic indicators for STPs, developing STP clustering and aggregation methods that exploit these profile curves and distance measures, and extending these methods to other types of prism, such as network time prisms and field-based prisms.

## References

- Andrienko N, Andienko G, Pelekis N, Spaccapietra S (2008) Basic concepts of movement data. In: Giannotti F, Pedreschi D (eds) *Mobility, data mining and privacy*. Springer, Heidelberg, pp 15–38
- Batty M (2010) Space, scale, and scaling in entropy maximizing. *Geogr Anal* 42:395–421
- Briggs D (2005) The role of GIS: coping with space (and time) in air pollution exposure assessment. *J Toxicol Environ Health Part A* 68:1243–1261
- Burns LD (1979) *Transportation, temporal and spatial components of accessibility*. Lexington Books, Lexington
- Dodge S, Laube P, Weibel R (2012) Movement similarity assessment using symbolic representation of trajectories. *Int J Geogr Inf Sci* 26:1563–1588
- Espeter M, Raubal M (2009) Location-based decision support for user groups. *J Location Based Serv* 3:165–187
- Giorgino T (2009) Computing and visualizing dynamic time warping alignments in R: the dtw package. *J Stat Softw* 31:1–24
- Gudmundsson J, Laube P, Wölle T (2012) Computational movement analysis. In: Kresse W, Danko DM (eds) *Springer handbook of geographic information*. Springer, Berlin, pp 423–438
- Hägerstrand T (1970) What about people in regional science? *Pap Reg Sci Assoc* 24:1–12



- Janowicz K, Raubal M, Kuhn W (2011) The semantics of similarity in geographic information retrieval. *J Spat Inf Sci* 2:29–57
- Kobayashi T, Miller HJ (2014) Exploratory visualization of collective mobile objects data using temporal granularity and spatial similarity. In: Cervone G, Lin J, Waters N (eds) *Data mining for geoinformatics: methods and applications*. Springer, pp 127–154
- Kuijpers B, Othman W (2009) Modeling uncertainty of moving objects on road networks via space-time prisms. *Int J Geogr Inf Sci* 23:1095–1117
- Long JA, Nelson TA (2012) Time geography and wildlife home range delineation. *J Wildl Manage* 76:407–413
- Long JA, Nelson TA (2013) A review of quantitative methods for movement data. *Int J Geogr Inf Sci* 27:292–318
- Miller HJ (1991) Modeling accessibility using space-time prism concepts within geographical information systems. *Int J Geogr Inf Syst* 5:287–301
- Miller HJ (2005) A measurement theory for time geography. *Geogr Anal* 37:17–45
- Miller HJ, Bridwell SA (2009) A field-based theory for time geography. *Ann Assoc Am Geogr* 99:49–75
- Nanni M, Kuijpers B, Körner C, May M, Pedreschi D (2008) Spatio-temporal data mining. In: Giannotti F, Pedreschi D (eds) *Mobility, data mining and privacy*. Springer, pp 267–296
- Okabe A, Sugihara K (2012) *Spatial analysis along networks: statistical and computational methods*. Wiley
- O'Sullivan D, Unwin D (2010) *Geographic information analysis*, 2nd edn. Wiley, Hoboken
- Pfoser D, Jensen CS (1999) Capturing the uncertainty of moving-object representations. In: Güting RH, Papadias D, Lochovsky F (eds) *Advances in spatial databases: 6th international symposium (SSD'99)*, vol 1651. Springer Lecture Notes in Computer Science, Berlin, pp 111–131
- Pred A (1977) The choreography of existence: comments on Hagerstrand's time-geography and its usefulness. *Econ Geogr* 53:207–221
- Raubal M, Miller HJ, Bridwell S (2004) User-centered time geography for location-based services. *Geografiska Annaler B* 86(4):245–265
- Sakoe H, Chiba S (1978) Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans Acoust Speech Signal Process* 26:43–49
- Sinha G, Mark DM (2005) Measuring similarity between geospatial lifelines in studies of environmental health. *J Geogr Syst* 7:115–136
- Song Y, Miller HJ (2014) Simulating visit probability distributions within planar space-time prisms. *Int J Geogr Inf Sci* 28:104–125
- Winter S, Yin ZC (2010a) The elements of probabilistic time geography. *Geoinformatica* 15:417–434
- Winter S, Yin ZC (2010b) Directed movements in probabilistic time geography. *Int J Geogr Inf Sci* 24:1349–1365
- Yuan Y, Raubal M (2012) Extracting dynamic urban mobility patterns from mobile phone data. In: Xiao N, Kwan M-P, Goodchild M, Shekhar S (eds) *Geographic information science—seventh international conference, GIScience 2012, Columbus, Ohio, USA, Sept 18–21 2012, Proceedings*. Springer, Berlin, pp 354–367
- Yuan Y, Raubal M (2014) Measuring similarity of mobile phone user trajectories: a spatio-temporal edit distance method. *Int J Geogr Inf Sci* 28:496–520

# Deriving the Geographic Footprint of Cognitive Regions

Heidelinde Hobel, Paolo Fogliaroni and Andrew U. Frank

**Abstract** The characterization of *place* and its representation in current Geographic Information System (GIS) has become a prominent research topic. This paper concentrates on places that are cognitive regions, and presents a computational framework to derive the geographic footprint of these regions. The main idea is to use Natural Language Processing (NLP) tools to identify unique geographic features from User Generated Content (UGC) sources consisting of textual descriptions of places. These features are used to detect on a map an initial area that the descriptions refer to. A semantic representation of this area is extracted from a GIS and passed over to a Machine Learning (ML) algorithm that locates other areas according to semantic similarity. As a case study, we employ the proposed framework to derive the geographic footprint of the *historic center of Vienna* and validate the results by comparing the derived region against a historical map of the city.

**Keywords** Geographic information retrieval · Cognitive regions · User generated content · Natural language processing · Machine learning · Semantic similarity

---

H. Hobel (✉)

Doctoral College Environmental Informatics, Vienna University of Technology,  
SBA Research, Vienna, Austria  
e-mail: hobel@geoinfo.tuwien.ac.at

P. Fogliaroni · A.U. Frank

Department for Geodesy and Geoinformation, Vienna University of Technology,  
Vienna, Austria  
e-mail: fogliaroni@geoinfo.tuwien.ac.at

A.U. Frank

e-mail: frank@geoinfo.tuwien.ac.at

# 1 Introduction

The characterization of *place* and its representation within Geographic Information System (GISs) are becoming prominent research topics in the field of geographic information science (Gao et al. 2013; Goodchild 2011; Scheider and Janowicz 2014; Kuhn 2001). The notion of place is strictly related to people’s conceptualization of space, and may correspond to different things, e.g. points of interest, geographic regions (Montello 2003), or settings (Schatzki 1991) (i.e., aggregations of spatial features).

A place is generally regarded as a region of space that is homogeneous with respect to certain criteria. We adopt the taxonomy for geographic regions proposed in Montello (2003) and focus on the category of so-called *cognitive regions*: conceptual regions derived by people as they experience the world. The geographic interpretation of cognitive regions may (and usually does) differ slightly among several individuals, as shown, for example, with the cognitive regions of downtown Santa Barbara (Montello et al. 2003), and Southern and Northern California and Alberta (Montello 2014).

In this paper we propose a novel approach to derive the geographic footprint (i.e. the location and extension) of a cognitive region from User Generated Content (UGC) sources containing textual descriptions of places. We argue that this type of UGC is a valuable knowledge base to derive an approximated geographic footprint of a cognitive region from, as it contains the conceptualizations that several people have of that region. In particular, we focus on a special type of cognitive regions: those that are conceptualized as homogeneous areas in terms of the activities they allow to be performed. To derive the geographic footprint of these regions we propose a novel framework that employs Natural Language Processing (NLP) tools to extract from textual descriptions of a place a set of named geographic features. These are used to detect on a map an initial area that the descriptions refer to, and to retrieve the activities one can perform in it. These activities provide a simplified semantic representation of the cognitive region of interest that is passed over to a Machine Learning (ML) algorithm to extend the initial area by locating other areas offering similar opportunities.

To the best of our knowledge, this is the first computational approach that exploits NLP and ML techniques based on the categorical attributes to derive an approximation of the geographic footprint of cognitive regions. As a case study we used the suggested framework to derive the geographic footprint of the cognitive region *historic center of Vienna*. Indeed, while the *historic center of Vienna* is clearly a concept that is widely referred to by people and has even a dedicated entry in Tripadvisor,<sup>1</sup> it is generally not retrievable from current GISs, at least at the time of writing this paper. As a preliminary evaluation we compared the derived area with a historic map of Vienna dating back to 1850, and we found that the two mostly coincide. Meanwhile,

---

<sup>1</sup><http://www.tripadvisor.at/>.

we are designing a questionnaire to assess the quality of the derived region as done in Montello et al. (2003). A pilot investigation showed that the region we derived matches very well with the conceptualizations of the subjects interviewed so far. We plan to publish the final results of this more detailed evaluation in a further paper.

The remainder of this paper is structured as follows. In Sect. 2, we review related work in the fields of place representation, Geographic Information Retrieval and Natural Language Processing. We present the framework in Sect. 3 and discuss the implementation and the results for the case study in Sect. 4. Section 5 concludes the paper, also discussing limitations of the presented approach and sketching future work.

## 2 Related Work

In this section, we discuss related work in the fields of Place and Vague Regions and Geographic Information Retrieval and Natural Language Processing.

### 2.1 *Place and Activities*

The notion of place plays a relevant role in everyday life (Winter et al. 2009; Winter and Truelove 2013). In the field of geographic information science, different research directions have emerged which investigate the representation of places (Goodchild 2011), classify and categorize various forms of places (Schatzki 1991), and model places according to their relations to activities and affordance theory (Jordan et al. 1998). According to Schatzki (1991): “[...] places are defined by reference to human activity”. Such a statement is supported by further research (Alazzawi et al. 2012; Montello et al. 2003; Rösler and Liebig 2013; Scheider and Janowicz 2014) implying that place semantics are closely related to activities.

In Schatzki (1991) it is argued that places organize into settings, local areas, and regions. This general notion of a hierarchical structuring of space is relatively undisputed and supported by findings of other researchers (Couclelis and Gale 1986; Freundsuh and Egenhofer 1997; Montello 1993; Richter et al. 2013). More specifically, Schatzki (1991) distinguishes two types of settings: those demarcated by barriers (e.g. apartment building), and those identified by bundles of activities that occur in them (e.g. playing in a park, shopping at a mall). Recently, the idea of equipping next-generation geographic search engines and recommendation systems with models that view places as aggregated entities has been receiving increasing attention (Ballatore 2014; Hobel et al. 2015).

## 2.2 *Geographic Information Retrieval and Natural Language Processing*

Geographic Information Retrieval (GIR) is a specialization of traditional information retrieval supported by geographic knowledge bases that enables the retrieval of geographic information and geotagged objects. The respective tools enable the identification and disambiguation of place names, the mapping of place names onto spatial features and vice versa, and the derivation of place semantics. Regarding the latter, the literature is mainly focused on the identification and classification of places (Tversky and Hemenway 1983; Smith and Mark 2001) and on the automatic generation of ontologies (Popescu et al. 2008).

To enhance the capabilities of the next generation of geographic search engines, different approaches are currently being pursued to facilitate the retrieval of geo-related content. Applications range from the conceptualization of space into a metric space algebra (Adams and Raubal 2009), to the contextualization of unstructured text (Adams et al. 2015; Adams and McKenzie 2012) to relate concepts to places, to the development of content-rich knowledge bases and vocabularies (Ballatore 2015), and to semantic similarity measures for geographic terms (Ballatore et al. 2013).

Interesting approaches of automatically mapping spatial content is pursued in different fields. Jones et al. (2008) focused on modeling vague regions by statistical density surfaces and mining place descriptions in natural language to infer the approximate region. Grothe and Schaab (2009) exploited freely available georeferenced photographs to derive the geographic footprint of imprecise regions by using Kernel Density Estimation and Support Vector Machines. Cunha and Martins (2014) derived imprecise regions by exploiting machine learning for interpolating from a set of point locations. Lüschner and Weibel (2013) concentrated on using characteristics of topographical data to delineate regions.

The current focus on similarity measures for geographic terms (Ballatore 2015; Ballatore et al. 2014) is further proof that there is an interest in the disambiguation of places and place descriptions. One of the goals is to prepare shared and universally accepted vocabularies to facilitate the interpretation and the resolution of spatial requests. For instance, if the task is to search for a place where *one can get something to eat*, there are more possible matches than just restaurants. Coffee shops, pubs, or even supermarkets may also fulfill the requirements of the request.

The availability of mature Natural Language Processing (NLP) tools (Manning et al. 2014) allows for advanced processing of textual spatial descriptions (Chang et al. 2015, 2014; Chang 2014; Coyne and Sproat 2001) where tokenization and part-of-speech taggers are used to automatically break text into meaningful symbols—a selection of Part-of-Speech Tags (POST) is shown in Table 1. Two recent interesting approaches are presented in Alazzawi et al. (2012) and McKenzie et al. (2013). The former builds upon current state-of-the-art NLP to extract spatial activities from unstructured text; the latter presents a model to derive user similarity from spatial topics they discuss on social media.

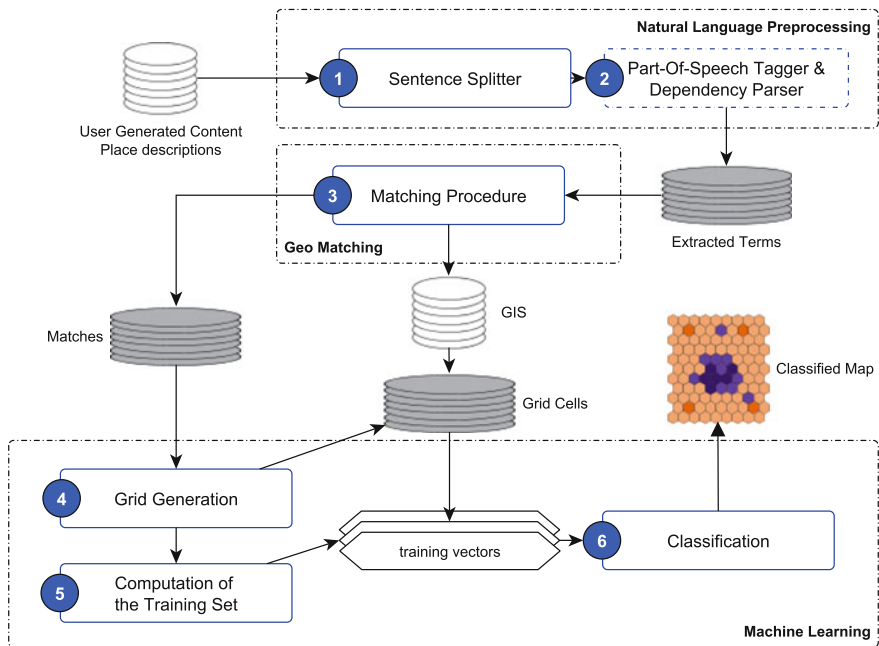
**Table 1** A selection of part-of-speech tags (POST) (Santorini 1990)

POST		POST	
Tag	Definition	Tag	Definition
CC	Coordinating conjunction	DT	Determiner
IN	Preposition or subordinating conjunction	WRB	Adverb
JJ	Adjective	WP	Pronoun
NN	Noun, singular or mass	TO	To
NNP	Proper noun, singular	VB	Verb

### 3 Deriving the Geographic Footprint of Cognitive Regions

In the following, we outline a processing workflow (see Fig. 1) to derive the geographic footprint of a given cognitive region from textual descriptions of that region. Details of the single steps involved are given in further sections.

The proposed approach relies on two types of data sources (depicted as white databases in the figure): (i) a User Generated Content (UGC) database containing



**Fig. 1** Schematic illustration of the proposed workflow to derive the geographic footprint of cognitive regions

textual descriptions of a given cognitive region, and (ii) a Geographic Information System (GIS).

The workflow consists of three main stages labeled in Fig. 1 as *Natural Language Preprocessing*, *Geo Matching*, and *Machine Learning*, respectively. First, the textual descriptions undergo a natural language processing phase in order to extract from them a set of nouns referring to geographic features. In the next step, this set is compared to the geonames available in the spatial database in order to assign each a location on the map. Finally, a grid of regular cells is superimposed onto the map and the cells containing at least one of the geographic features mentioned in the textual descriptions are selected. These, together with a different set of cells selected randomly from the grid as counterexamples, are used as training samples for a machine learning algorithm that categorizes all other cells according to the activities they allow. As a result, each cell is associated to either of the two training sets, unless too little information is known about it—in which case it is marked as “unclassified”.

### 3.1 Natural Language Preprocessing

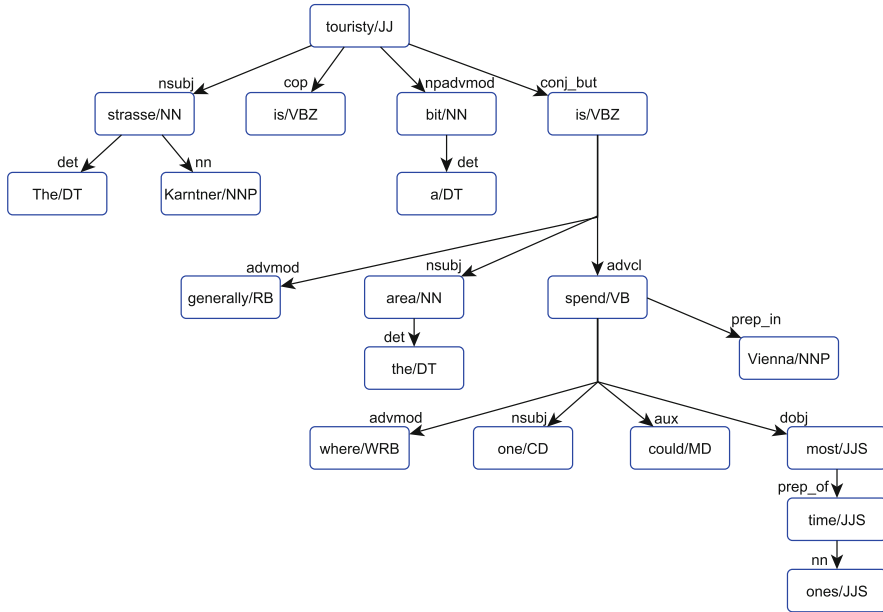
The natural language preprocessing stage relies on the Stanford CoreNLP Natural Language Processing Toolkit (Manning et al. 2014). More specifically, it relies on three of the tools it provides: the *sentence splitter*, the *part-of-speech tagger*, and the *dependency parser*.

The sentence splitter tokenizes each UGC description into sentences (step 1 in Fig. 1) that are passed over to the part-of-speech tagger and the dependency parser (step 2 in Fig. 1). The tagger classifies every word in a sentence according to its syntactical class, e.g. noun (NN), verb (VB), adjective (JJ) (see Table 1 for a more complete list of syntactical classes and tags). The parser generates a so-called dependency tree whose nodes denote the syntactical class of each word in a sentence, with edges representing the hierarchical structure of grammatical relations between the words. For example, given the sentence “*The Karntner Strasse<sup>2</sup> is a bit touristy, but generally the area is where one could spend most of one’s time in Vienna.*”, the part-of-speech tagger and the dependency parser produce the tree shown in Fig. 2. Note that each term is also lemmatized, i.e. it is transformed into its base form.

Given a dependency tree, it is easy to extract from it the set  $\mathcal{N}$  of common and proper nouns—tagged NN and NNP, respectively. Possibly, this set contains any reference to geographic features contained in the textual description that we are interested in locating on the map. Since the name of a geographic feature may be a compound noun (e.g. Kärntner Straße, St. Stephen’s Cathedral), we need to further process the set of nouns before trying to match them with geonames available in the geographic database.

---

<sup>2</sup>The correct spelling in German language is Kärntner Straße. This comment has been retrieved from the web and is purposely reported in its original, wrongly spelled, form.



**Fig. 2** The dependency tree generated by the Stanford’s part-of-speech tagger and dependency parser (Manning et al. 2014) for the sentence “The Karnener Strasse is a bit touristy, but generally the area is where one could spend most of one’s time in Vienna”

---

### Algorithm 1 Finding candidate compound geonames

---

**Input**

$\mathcal{N} = \{\text{nouns in UGC descriptions}\}$ ,  
 $x = \text{maximum number of words making up a compound geoname}$

**Output**  $\mathcal{C} = \{\mathcal{C}_n, n \in \mathcal{N}\} = \{\text{candidate geonames for each noun } n \text{ in } \mathcal{N}\}$

- 1: **procedure** COMPOUNDGEONAMES
  - 2:    $\mathcal{C} \leftarrow \emptyset$
  - 3:   **for all**  $n \in \mathcal{N}$  **do**
  - 4:      $\mathcal{C}_n \leftarrow \emptyset$
  - 5:      $\mathcal{D} \leftarrow \{n\} \cup \text{RetrieveDependencies}(n, x)$
  - 6:     **for all**  $d \in 2^{\mathcal{D}}$  **do**
  - 7:        $\mathcal{C}_n \leftarrow \mathcal{C}_n \cup \{\text{PermutationsOf}(d)\}$
  - 8:      $\mathcal{C} \leftarrow \mathcal{C} \cup \mathcal{C}_n$
- 

We propose the procedure reported in Algorithm 1 that, given the set of nouns  $\mathcal{N}$ , produces a set  $\mathcal{C}$  consisting of simple and compound nouns that we refer to as *candidate geonames*. For each noun  $n \in \mathcal{N}$  we access again the dependency tree to retrieve other nouns that, together with  $n$ , might make up a compound noun. This is done through the function *RetrieveDependencies*( $n, x$ ) (line 5) which, starting from the node corresponding to  $n$ , traverses the tree upwards (towards the root) and downwards (towards the leaves) and retrieves up to  $x \in \mathbb{N}$  other nouns in both directions. These nouns, together with  $n$ , are stored in the set  $\mathcal{D}$ . The final set  $\mathcal{C}_n$  of candidate



compound nouns associated to  $n$  consists of all possible permutations of the power-set of  $\mathcal{D}$  (lines 6–7). The complete set of candidate geonames consists of the union (line 8) of all such sets of candidate geonames:  $\mathcal{C} = \bigcup_{n \in \mathcal{N}} \mathcal{C}_n$ .

In our example, from the dependency tree in Fig. 2 we derive:

$$\mathcal{N} = \{\textit{Strasse, Karntner, bit, area, time, Vienna}\}$$

And for the noun  $n = \textit{Karntner}$  we have:

$$\mathcal{D}_{\textit{Karntner}} = \{\textit{Karntner, Strasse}\}$$

$$\mathcal{C}_{\textit{Karntner}} = \{\emptyset, \textit{Karntner, Strasse, Karntner Strasse, Strasse Karntner}\}$$

Note that in this case the number  $x$  of dependencies to be retrieved does not influence the sets of candidate compound names, as far as  $x > 0$ .

### 3.2 Geographic Matching

This stage does not rely on any external tool. The objective is trying to match every candidate geoname obtained in the previous stage against a unique feature in the geographic database according to name comparison (step 3 in Fig. 1). The result is a set  $\mathcal{G}$  that, for each noun  $n \in \mathcal{N}$ , contains at most one geographic feature from the database: the one whose name best matches the candidate geonames for  $n$  (i.e., in  $\mathcal{C}_n$ ). This implies that we also discard nouns referring to categorical features (e.g. street, square), as our final goal is to pinpoint an initial area on the map that the textual descriptions refer to.

---

#### Algorithm 2 Geographic matching

---

**Input**

$\mathcal{N} = \{\textit{nouns in UGC descriptions}\},$

$\mathcal{C} = \{\mathcal{C}_n, n \in \mathcal{N}\} = \{\textit{candidate geonames for each noun } n \textit{ in } \mathcal{N}\},$

$\varepsilon = \textit{threshold}$

**Output**  $\mathcal{G} = \{\textit{matched geonames}\}$

```

1: procedure GEOMATCHING
2:    $\mathcal{G} \leftarrow \emptyset$ 
3:   for all  $n \in \mathcal{N}$  do
4:      $\mathcal{D} \leftarrow \textit{patternMatch}(n)$ 
5:      $(\underline{p}, \underline{d}) \leftarrow (\textit{nil}, +\infty)$ 
6:     for all  $(c, p) \in \mathcal{C}_n \times \mathcal{D}$  do
7:        $d \leftarrow \textit{Levenshtein}(c, p.\textit{name})$ 
8:       if  $d \leq \varepsilon \cdot \textit{WordsIn}(c) \wedge d < \underline{d}$  then
9:          $(\underline{p}, \underline{d}) \leftarrow (p, d)$ 
10:    if  $\underline{p} \neq \textit{nil} \wedge \textit{IsUnique}(\underline{p})$  then
11:       $\mathcal{G} \leftarrow \mathcal{G} \cup \{\underline{p}\}$ 

```

---

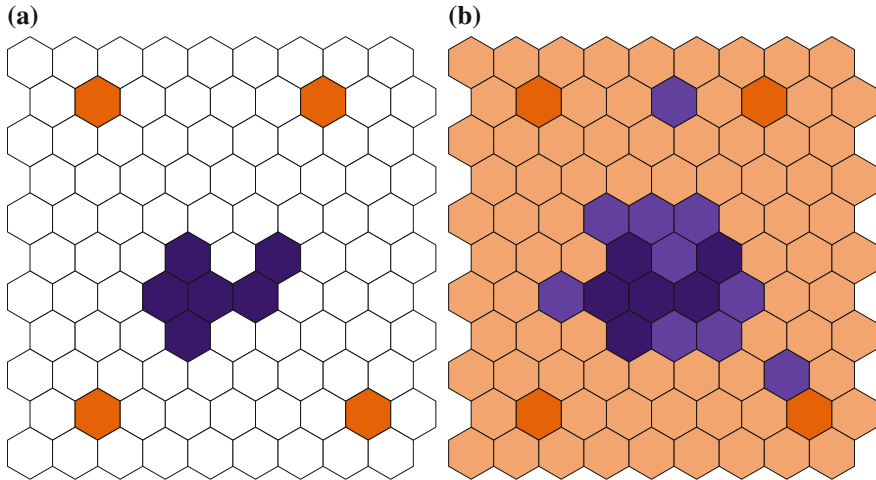
We propose the procedure reported in Algorithm 2 that works as follows. For each noun  $n \in \mathcal{N}$  we retrieve (line 4) from the geographic database a set  $\mathcal{P}$  of features whose names pattern-match (i.e. via regex expression) against  $n$ . In defining the regex expressions, particular attention must be given to encode case-insensitivity and special characters (e.g. vowel mutations) to deal with spelling issues occurring when people write place names in a non-native language (e.g. the German word Straße vs. Strasse). Of all the retrieved features  $\mathcal{P}$  we are only interested in selecting (lines 5–9) one whose name best matches against the set  $\mathcal{C}_n$  of candidate geonames associated to  $n$ . For each candidate geoname  $c \in \mathcal{C}_n$  and for each feature  $p \in \mathcal{P}$  we compute the Weighted Levenshtein distance<sup>3</sup> between  $c$  and the name of  $p$  (line 7). The Weighted Levenshtein distance is a reasonable choice in case of UGC as it allows to cope with incompleteness and irregularities typical of UGC. To find possible matches (line 9) we enforce (line 8) the distance not to be bigger than a given threshold  $\epsilon$ . Since a candidate geoname might be a compound name, we multiply  $\epsilon$  by the number of words making up the candidate geoname. The best matching, then, is the one with the smallest Levenshtein distance. At the end of the loop the variable  $\underline{p}$  is either empty or it contains a geographic feature. In the first case no match has been found. Otherwise we must make sure that the feature is unique in the geographic database (line 10). This might not be the case for features like e.g. shops or restaurants that have several branches in the same city.

Let us resume the example sentence introduced in Sect. 3.1 and whose dependency tree is shown in Fig. 2. Assume that for the noun  $n = \textit{Karntner}$  and for the case-insensitive regex expression “k(alaelä)rntner” the function *patternMatch* (line 4) returns only one feature named ‘Kärntner Straße’. The following table then shows the resulting Levenshtein distance for each candidate geoname in  $\mathcal{C}_n$  and the threshold (assumed to be  $\epsilon = 3$ ) multiplied by the number of words in each noun:

$c \in \mathcal{C}_n$	Levenshtein distance	$\epsilon \cdot \textit{WordsIn}(c)$
$\emptyset$	15	0
‘Karntner’	8	3
‘Strasse’	11	3
<b>‘Kärntner Strasse’</b>	<b>3</b>	<b>6</b>
‘Strasse Karntner’	12	6

It is easy to see that there is only one entry in this table whose distance is admissible and is minimum: the entry ‘Kärntner Strasse’.

<sup>3</sup>The Levenshtein distance is a string metric that measures similarity by the minimal number of required editing steps to transform one string into the other string.



**Fig. 3** A schematic representation of the classification process—given training vectors for the cognitive region of interest (*purple cells* in (a)) and for the counter-examples (*orange cells* in (a)), the machine learning classifier associates each other cell to one of the two classes (*light purple* and *light orange* in (b)). **a** Initial configuration. **b** Classified area

### 3.3 Machine Learning

This stage relies on a machine learning model called Multinomial Naïve Bayes: a probabilistic approach mainly used for text classification that learns from a given set of pre-classified samples (called training vectors) how to classify other, unclassified feature vectors according to their similarity with the given training vectors.

We adapt Multinomial Naïve Bayes to classify geographic areas as either being part of the cognitive region of interest (class 1) or not being part of it (class 2). The training vectors are obtained by tessellating the map with a regular grid (step 4 in Fig. 1) and retrieving the cells  $\mathcal{S}_1$  containing at least one of the geographic features  $\mathcal{G}$  derived in the previous stage. Such cells are the training vectors for the first class. The training vectors  $\mathcal{S}_2$  for the second class consist of the same number of randomly selected cells that do not contain any of the geographic features in  $\mathcal{G}$ .

We adopt a bag-of-words model<sup>4</sup> to obtain a simplified ‘semantical’ representation of the training cells by extracting certain categorical attributes from all the geographic features contained in each such cell (step 5 in Fig. 1). Let  $\mathcal{T} := \{t_i; i = 1, \dots, n\}$  be a vocabulary containing all categorical attributes of interest from the whole map. Then, each cell is represented by a vector  $(x_1, \dots, x_n)$ , where  $x_i$  is the

<sup>4</sup>The bag-of-words model is typically used for text classification. A text is represented as the bag (multiset) of its words and the frequency of occurrence of each word is used as a feature vector for training a classifier.

frequency of the categorical attribute  $t_i$  in this cell. Since our focus is on cognitive regions conceptualized as homogeneous areas in terms of the activities they allow to be performed, we only (so far manually based on an educated guess of the most typical activities for a cognitive region such as historic city centers<sup>5</sup>) select categorical attributes proper of geographic features that offer a service (e.g. bars, shops, restaurants, banks, ...).

Given the two training sets  $\mathcal{S}_1$  and  $\mathcal{S}_2$  as described above, the machine learning procedure is capable of classifying all the remaining cells (step 6 in Fig. 1) as graphically exemplified in Fig. 3.

## 4 Evaluation

This section describes an implementation of the processing workflow described in Sect. 3 and the results we obtained for the case study of the cognitive region *historic center of Vienna*.

As data sources (see Fig. 1) we selected two well-known UGC and Volunteered Geographic Information (VGI) projects: TripAdvisor<sup>6</sup> and OSM.<sup>7</sup> By means of a customized crawler we retrieved English textual descriptions of the *historic center of Vienna* from a dedicated comment page on TripAdvisor. For the geographic database we used the OSM extract of Vienna as provided by Mapzen Metro Extracts.<sup>8</sup> OSM provides spatial data in the form of *points* (e.g. a park bench), *ways* (e.g. streets and buildings), and *relations* (e.g. spatial entities consisting of several parts). Semantic information such as name and categorical attributes are defined as ‘tags’, which are key-value pairs. For example, OSM contains an entry for the “Hofburg Imperial Palace” that includes the name of the feature in several languages and is described by the following tags (among others): (*building, yes*), (*historic, castle*), (*castle\_type, palace*), (*tourism, attraction*). The spatial dataset (see Fig. 4) was stored in a dedicated database where the geometry of ways and relations was simplified by their centroid.

Finally, for the implementation of the machine learning stage (see Sect. 3.3) we resorted to a hexagonal grid with uniform cells with an edge-length of 0.0025°,<sup>9</sup> and we used the MatLab implementation of the Multinomial Naïve Bayes<sup>10</sup> classifier.

---

<sup>5</sup>We are working on an extension to select activities from textual descriptions.

<sup>6</sup><http://www.tripadvisor.com/>.

<sup>7</sup><https://www.openstreetmap.org/>.

<sup>8</sup><https://mapzen.com/>.

<sup>9</sup>The cell size can be shrunk or enlarged to obtain finer-grained or coarser results, respectively.

<sup>10</sup><http://de.mathworks.com/help/stats/naive-bayes-classification.html>.



**Fig. 4** Visualization of the OpenStreetMap (OSM) dataset of Vienna as used in our experiments—the whole dataset consists of 290,586 nodes, 368,112 ways, and 810,145 relations, for a total of 1,468,843 features

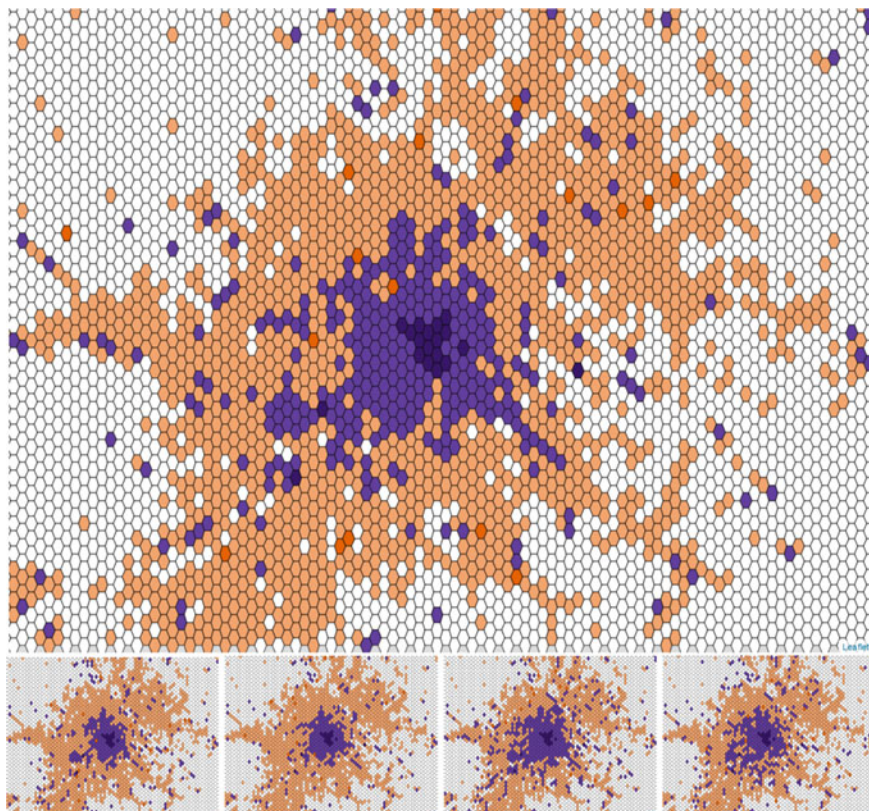
### ***4.1 Experimental Results***

We ran our workflow implementation on two experimental scenarios. Both scenarios use the same data sources with the following difference: in the first scenario (see Fig. 5), the training vectors have been kept in their integrity. In the second scenario (see Fig. 6), we manually removed outlier cells from the training vector associated to the cognitive region—i.e., those cells that fall far away from the actual city center (compare the distribution of dark purple cells in Figs. 5 and 6).

For the pictorial representations of the results we adopted the following color scheme: dark purple cells represent training vectors for the cognitive region *historic center of Vienna* as extracted from the textual descriptions; dark orange cells represent training vectors for the counter-example. Light purple and light orange cells show the areas classified as *historic center of Vienna* and counter-example, respectively. White cells denote areas that have not been classified because of insufficient semantic information.

Since counter-examples are selected randomly from the grid we decided to perform several runs for each scenario. Figure 5 shows the results obtained for five runs on the first scenario. Figure 6 shows similar results for the second scenario, where outlier cells were removed from the training vector of the cognitive region. The results for the two scenarios mostly coincide, and the cells classified as similar to the cognitive region *historic center of Vienna* form a region approximately corresponding to the central district of the city and its immediate surroundings. Interestingly,





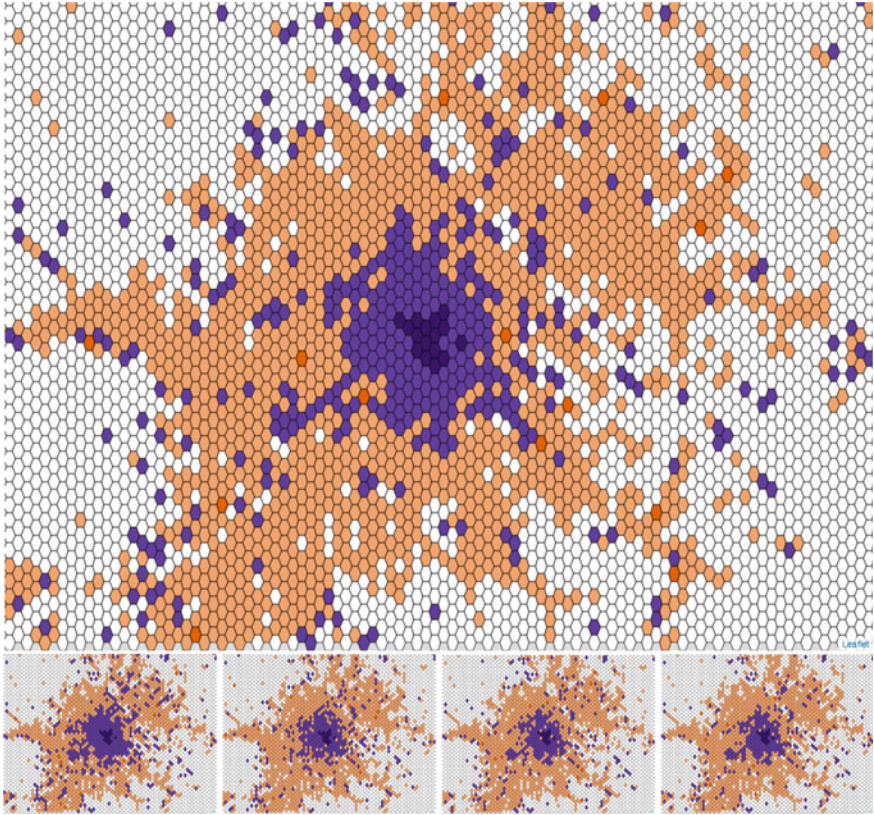
**Fig. 5** Visualization of classification results for the first scenario (several runs)—dark purple cells represent training vectors for the cognitive region *historic center of Vienna*; light purple cells are classified as *historic center of Vienna*; dark orange cells represent training vectors for the counter-example; light orange cells are classified as counter-example; white cells are unclassified

the cells that were manually removed in the second scenario are associated to the class corresponding to the cognitive region anyway.

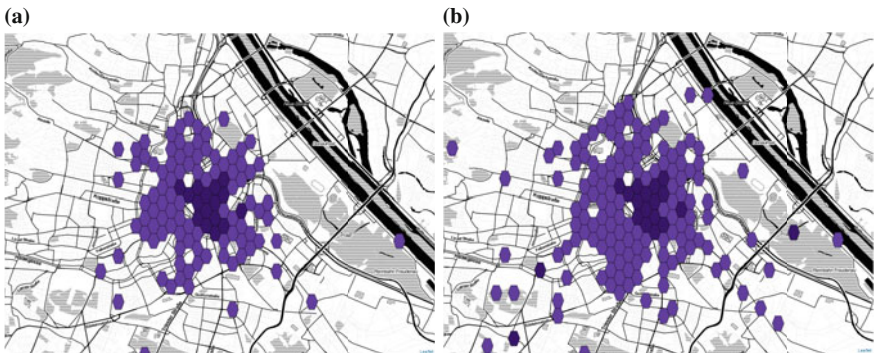
To mitigate the effects of using randomly selected counter-examples, we performed ten runs for each scenario and intersected the results to obtain ‘robust’ results: only cells classified as *historic center of Vienna* that occur in the result of each run form the robust results, as shown in Fig. 7.

## 4.2 Preliminary Evaluation

A sound evaluation of the results would require investigating how the derived cognitive regions fit to human conceptualization. To that end we are currently in the process of designing a questionnaire similar to that used in Montello et al. (2003).



**Fig. 6** Visualization of classification results for the second scenario (several runs)—dark purple cells represent training vectors for the cognitive region *historic center of Vienna*; light purple cells are classified as *historic center of Vienna*; dark orange cells represent training vectors for the counter-example; light orange cells are classified as counter-example; white cells are unclassified



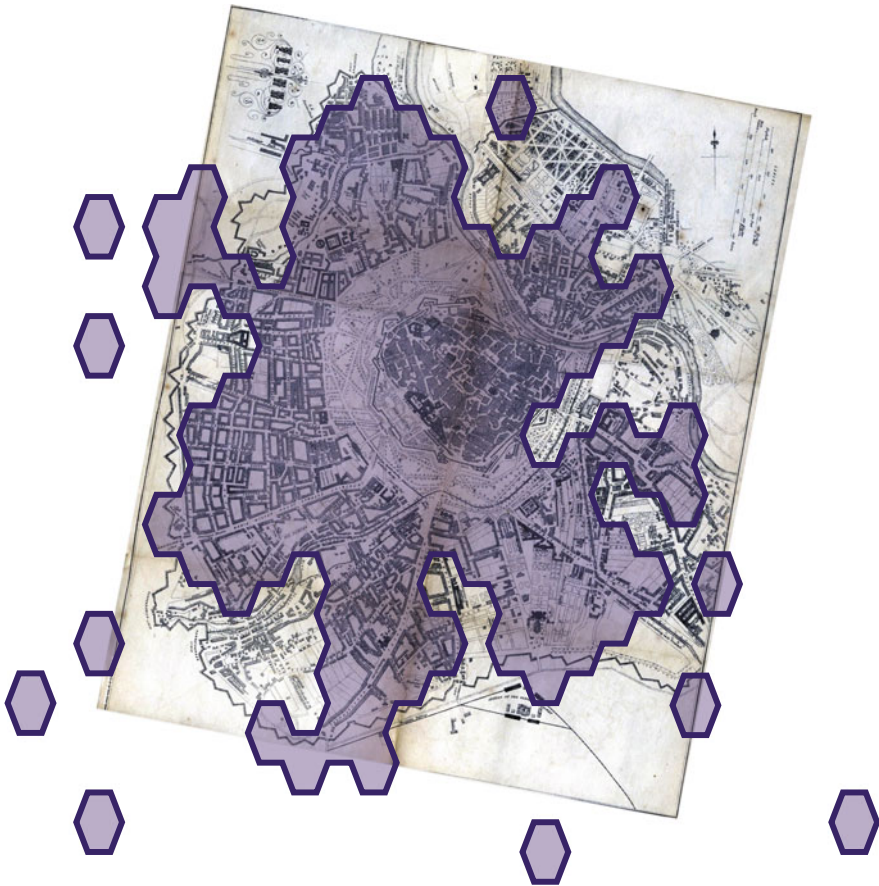
**Fig. 7** Visualization of robust results. **a** Scenario 1. **b** Scenario 2



A pilot study with a small user group already showed that the footprint derived for the cognitive region *historic center of Vienna* fits well to human conceptualization. We plan to publish the final results of this study in a future publication.

Meanwhile, we present here a preliminary qualitative evaluation of the outcomes by comparing the obtained robust results with a historical map of the city of Vienna that dates back to 1850. For that, we geographically overlaid the derived regions with the map, as shown in Fig. 8 for the first scenario. It is easy to see how the shape and extent of the derived region nicely fit with the city boundaries of 1850: The outer boundary of the main part of the classified area coincides with a physical separation which is now a major street of the city, while the few outlier cells correspond to historical sites that are not reported in the historical map (e.g. the Schönbrunn Palace).

In summary, the approach, which relied solely on a knowledge base derived from VGI and crowdsourced information sources, shows promising results.



**Fig. 8** Approximate overlay of the robust result (Scenario 1) over a historic representation of Vienna (map retrieved from <http://www.valentina.net>)



## 5 Conclusion and Future Work

We presented a novel automated approach to derive the geometric extent of “cognitive regions” by utilizing solely crowd-sourced geographic information as the fundamental knowledge bases. Based on Natural Language Preprocessing and a combinatorial place matching procedure tailored to identify unique geonames, the conceptualization of regions perceived as a whole based on the activities they allow is translated into a machine-processable form.

The proposed approach builds upon a representation of semantic attributes of geographic features, and allows for the automated clustering of cities into “cognitive regions”. We pointed out that the classification problem can be efficiently solved by utilizing the Multinomial Naïve Bayes model as classifier. For that, a bi-classification approach was discussed that operates on initial seeding cells identified by the combinatorial place matching procedure. Counter-examples are derived using a Monte Carlo approach.

While the method presented in this paper reveals promising results, it presents a number of limitations that we plan to overcome in future work.

First, our approach relies on uniquely identifiable places to derive the initial cells for the machine learning model, while non-unique features (like shops with several branches) are completely discarded and not used to create the training vectors. Different approaches can be devised to also exploit such non-unique features, according to whether they are mentioned in a comment together with uniquely identifiable features or not. In the first case, one approach would be to use the dependency tree to locate the syntactically closest unique feature mentioned in the text. This could be used as a reference point to locate on the map the spatially closest feature matching the non-unique reference. In the second case, a solution would be to run a two-step geomatching. In the first step only uniquely identifiable features are used (as done in the current approach) to generate a starting set of training cells. In the second step this initial set is recursively extended by disambiguating non-unique features according to their vicinity to the training cells.

It could be argued that the random selection of counter-examples for the machine learning model can be improved by applying sophisticated methods. For example, one could derive an ontology of cognitive regions and select counter-examples from those that are semantically furthest away from the region of interest.

We adopted a bag-of-words model as a semantical approximation of the training cells. This is a rather coarse semantical representation, as it only accounts for the frequency of categorical attributes in a given cell. An improvement would be to resort to a model that also takes into consideration the ontological relations among the attributes as well as their spatial distribution and configuration.

Finally, we are working on a further extension of this approach that also exploits verbs and other syntactical classes to derive the activities that can be carried out at a given place. This extension may allow for a variety of more advanced applications such as the enrichment and/or validation of semantic attributes in geographic

databases, as well as enabling natural language interfaces for Geographic Information Retrieval (GIR) systems.

**Acknowledgments** We acknowledge the work of © OpenStreetMap contributors (<http://www.openstreetmap.org/copyright>), and Leaflet (<http://leafletjs.com>). This research was partially funded by the Vienna University of Technology through the Doctoral College Environmental Informatics.

## References

- Adams B, McKenzie G (2012) Frankenplace: an application for similarity-based place search. ICWSM
- Adams B, McKenzie G, Gahegan M (2015) Frankenplace: interactive thematic mapping for ad hoc exploratory search. In: Proceedings of the 24th international conference on world wide web. International World Wide Web Conferences Steering Committee, pp 12–22
- Adams B, Raubal M (2009) A metric conceptual space algebra. In: Spatial information theory. Springer, pp 51–68
- Alazzawi AN, Abdelmoty AI, Jones CB (2012) What can i do there? Towards the automatic discovery of place-related services and activities. *Int J Geogr Inf Sci* 26(2):345–364
- Ballatore A (2014) The search for places as emergent aggregates
- Ballatore A, Bertolotto M, Wilson DC (2015) A structural-lexical measure of semantic similarity for geo-knowledge graphs. *ISPRS Int J GeoInf* 4(2):471–492
- Ballatore A, Wilson DC, Bertolotto M (2013) Computing the semantic similarity of geographic terms using volunteered lexical definitions. *Int J Geogr Inf Sci* 27(10):2099–2118
- Ballatore A, Bertolotto M, Wilson DC (2014) An evaluative baseline for geo-semantic relatedness and similarity. *Geoinformatica* 18(4):747–767
- Chang AX, Savva M, Manning CD (2014) Interactive learning of spatial knowledge for text to 3d scene generation. Sponsor: Idibon, p 14
- Chang AX, Savva M, Manning CD (2014) Learning spatial knowledge for text to 3d scene generation. EMNLP
- Chang A, Monroe W, Savva M, Potts C, Manning CD (2015) Text to 3d scene generation with rich lexical grounding. [arXiv:1505.06289](https://arxiv.org/abs/1505.06289)
- Couclelis H, Gale N (1986) Space and spaces. *Geografiska annaler. Series B. Hum Geogr* 68(1):1–12
- Coyne B, Sproat R (2001) Wordseye: an automatic text-to-scene conversion system. In: Proceedings of the 28th annual conference on computer graphics and interactive techniques. ACM, pp 487–496
- Cunha E, Martins B (2014) Using one-class classifiers and multiple kernel learning for defining imprecise geographic regions. *Int J Geogr Inf Sci* 28(11):2220–2241
- Freundschuh SM, Egenhofer MJ (1997) Human conceptions of spaces: implications for gis. *Trans GIS* 2(4):361–375
- Gao S, Janowicz K, McKenzie G, Li L (2013) Towards platial joins and buffers in place-based gis. In: Proceedings of the 1st ACM SIGSPATIAL international workshop on computational models of place (COMP’2013), pp 1–8
- Goodchild MF (2011) Formalizing place in geographic information systems. In: Communities, neighborhoods, and health. Springer, pp 21–33
- Grothe C, Schaab J (2009) Automated footprint generation from geotags with kernel density estimation and support vector machines. *Spat Cogn Comput* 9(3):195–211
- Hobel H, Abdalla A, Fogliarini P, Frank AU (2015) A semantic region growing algorithm: extraction of urban settings. In: AGILE 2015. Springer, pp 19–33

- Jones CB, Purves RS, Clough PD, Joho H (2008) Modelling vague places with knowledge from the web. *Int J Geogr Inf Sci* 22(10):1045–1065
- Jordan T, Raubal M, Gartrell B, Egenhofer M (1998) An affordance-based model of place in gis. In: 8th international symposium on spatial data handling, SDH, vol 98, pp 98–109
- Kuhn W (2001) Ontologies in support of activities in geographical space. *Int J Geogr Inf Sci* 15(7):613–631
- LÄscher P, Weibel R (2013) Exploiting empirical knowledge for automatic delineation of city centres from large-scale topographic databases. *Comput Environ Urban Syst* 37:18–34
- Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D (2014) The stanford corenlp natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pp 55–60
- McKenzie G, Adams B, Janowicz K (2013) A thematic approach to user similarity built on geosocial check-ins. In: *Geographic information science at the heart of Europe*. Springer, pp 39–53
- Montello DR (1993) Scale and multiple psychologies of space. In: *Spatial information theory a theoretical basis for GIS*. Springer, pp 312–321
- Montello DR (2003) Regions in geography: process and content. In: *Foundations of geographic information science*, pp 173–189
- Montello DR, Goodchild MF, Gottsegen J, Fohl P (2003) Where's downtown? Behavioral methods for determining referents of vague spatial queries. *Spat Cogn Comput* 3(2–3):185–204
- Montello DR, Friedman A, Phillips DW (2014) Vague cognitive regions in geography and geographic information science. *Int J Geogr Inf Sci* 28(9):1802–1820
- Popescu A, Grefenstette G, Mo ëllic PA (2008) Gazetiki: automatic creation of a geographical gazetteer. In: Proceedings of the 8th ACM/IEEE-CS joint conference on digital libraries. ACM, pp 85–93
- Richter D, Vasardani M, Stirling L, Richter K-F, Winter S (2013) Zooming in-zooming out hierarchies in place descriptions. In: Krisp JM (ed) *Progress in location-based services*. Lecture notes in geoinformation and cartography. Springer, Berlin, pp 339–355
- Rösler R, Liebig T (2013) Using data from location based social networks for urban activity clustering. In: Vandenbroucke D, Bucher B, Crompvoets J (eds) *Geographic information science at the heart of Europe*, Lecture notes in geoinformation and cartography. Springer, pp 55–72
- Santorini B (1990) Part-of-speech tagging guidelines for the penn treebank project (3rd revision)
- Schatzki TR (1991) Spatial ontology and explanation. *Ann Assoc Am Geogr* 81(4):650–670
- Scheider S, Janowicz K (2014) Place reference systems: a constructive activity model of reference to places. *Appl Ontol* 9(2):97–127
- Smith B, Mark DM (2001) Geographical categories: an ontological investigation. *Int J Geogr Inf Sci* 15(7):591–612
- Tversky B, Hemenway K (1983) Categories of environmental scenes. *Cogn Psychol* 15(1):121–149
- Winter S, Truelove M (2013) Talking about place where it matters. In: Raubal M, Mark DM, Frank AU (eds) *Cognitive and linguistic aspects of geographic space*. Lecture notes in geoinformation and cartography. Springer, Berlin, pp 121–139
- Winter S, Kuhn W, Krüger A (2009) Guest editorial: does place have a place in geographic information science? *Spat Cogn Comput* 9(3):171–173

**Part II**  
**Crowdsourcing and Social Networks**

# Android-Based Multi-Criteria Evaluation Approach for Enhancing Public Participation for a Wind Farm Site Selection

Pece V. Gorsevski and Alberto Manzano Torregrosa

**Abstract** This project presents a hypothetical case study of an interactive mobile-based Public Participation Geographical Information Systems (PPGIS) prototype for selection of best alternative for new offshore wind farm development in Lake Erie, northern Ohio. The prototype implements a client-server architecture where Android operating system is used for the client side, and Google Cloud Platform services and GeoServer/PostgreSQL for the server side. The potential benefits from this prototype are demonstrated through an interactive Android interface where the importance of three decision alternatives is evaluated by multiple participants using different evaluation criteria. The individual evaluation scores are aggregated by using a mathematical Pairwise comparison voting method while the sum of all individual Pairwise comparison scores yields the group solution. The results from the group solution are interactively returned and used for building consensus and to aid understanding of potential solutions coalesced from multiple participants' perspectives.

**Keywords** Spatial decision support system • Wind farm siting • Multi-criteria evaluation • PPGIS • Mobile GIS • Android GIS

## 1 Introduction

Traditional Geographical Information Systems (GIS) have been adopted in many wind farm suitability applications, such as wind farm siting (Al-Yahyai et al. 2012; Baban and Parry 2001; Tegou et al. 2010; Voivontas et al. 1998), evaluation of

---

P.V. Gorsevski (✉)

School of Earth, Environment & Society, Bowling Green State University,  
Bowling Green, OH 43403, USA  
e-mail: peterg@bgsu.edu

A.M. Torregrosa

Computer Science, Bowling Green State University, Bowling Green, OH 43403, USA  
e-mail: amanzan@bgsu.edu

visual impacts (Aydin et al. 2010; Berry et al. 2011; Hurtado et al. 2004; Rodrigues et al. 2010; Yeo et al. 2013) and other ecological impacts on birds and their habitats (Aydin et al. 2010; Baban and Parry 2001; Farfán et al. 2009; Gorsevski et al. 2013). However, one of the shortcomings is that traditional GIS alone has limited functionality because it lacks the analytical modeling capabilities and does not support complex spatial planning activities such as the involvement of multiple participants (Bishop and Stock 2010; Gorsevski et al. 2012, 2013; Simão et al. 2009).

In the last decade, the evolution of spatial decision support systems (SDSS) added new capabilities intended for solving complex problems that often involve a large number of decision alternatives including economic, social, or environmental and for enhancement of variety of management solutions (Malczewski 1999a, 2006a, b). The primary goal of those SDSS is to integrate GIS capabilities with multiple criteria evaluation (MCE) methods used to support decision making and complex spatial planning problems (Borouhaki and Malczewski 2008, 2010a, b; Donevska et al. 2012; Gorsevski et al. 2012). However, current SDSS include complex methodological computer applications and are mostly designed for advanced GIS uses but they lack support of a multi-user interface and do not support access for general public participation.

Public participation in wind farm planning problems is a very important process that helps to facilitate consensus building and to ensure future legitimacy and acceptance of such projects (Gorsevski et al. 2013; Jankowski and Nyerges 2001; Mekonnen and Gorsevski 2015; Nyerges and Jankowski 2009; Simão et al. 2009). However, some of the difficulties with the inclusion of public participants are conflicting views, diverse interests, values, and objectives that are inherent to different community members involved in the process. Such problems are referred as “ill-defined” or “wicked” decision problems, which result in an infinite number of solutions that are driven by differences in values, motives, and/or locational perspectives (Malczewski 1999b). Because of this complexity, alternative methods that involve public participation are needed.

A recent GIS development, coined the Public Participation Geographical Information Systems (PPGIS) concept, aims to provide broad public accessibility by using a web-based environment that can support an unlimited number of users who are free to participate at their own convenient time and location (i.e., asynchronous distributed interaction model) (Berry et al. 2011; Borouhaki and Malczewski 2010b; Simão et al. 2009). While several current web-based systems exist for visual and computational decision support that can support an ample number of participants, the exponential growth of cell phone usage and ubiquitous infrastructure creates even more opportunities for wider public involvement.

The latest development of smartphones such as iPhone, Windows 10 Mobile, Ubuntu Touch OS, and Android-based phones, offer a large amount of storage capacities, high processing and memory capabilities, advanced connectivity through Wi-Fi or other 3G/4G networks, and are easy to use, deploy and scale (Saeed et al. 2013; Weng et al. 2012). However, smartphone applications typically require backend components and services to feed applications with relevant and interactive

user data. Google Cloud Platform<sup>1</sup> is one of the popular solutions that can easily integrate and support Android and other iOS devices as a backend platform for different mobile solution tasks such as storage, retrieval, and processing data externally of the mobile devices (Anon n.d.).

Given the capabilities of mobile applications to integrate to current PPGIS tools and the surge in mobile phone growth, in this research project we demonstrate a custom-built, Android based prototype that is intended for aiding public participation for wind farm site selection. Although this research was implemented on an Android platform, other hardware platforms, operating systems, or cross-platform software could be considered. In this research, the PPGIS prototype tool integrates mobile, cloud and local server technologies for enabling visualization of decision criteria and alternatives, and efficient timely data collection/communication, which can play a vital role in participatory decision making. The potential of the tool is illustrated by evaluating three predefined decision alternatives using various evaluation criteria in the southwestern part of Lake Erie, Ohio. The four main components that are emphasized in the illustrated tool include: a discussion forum, mapping, decision, and result visualization tool. The discussion forum is used to facilitate communication and debate among participants regarding different criteria before they use the decision tool. The map tool is used in conjunction with the discussion forum for exploration and visualization of the decision alternatives associated with different criteria while the decision tool allows participants to make their personal decisions by ranking the decision alternatives using different sets of criteria and casting their votes. Finally, the result visualization tool is used to display charts with real-time results of the voting process. The methodology, the proposed conceptual framework, and the system architecture are discussed in the sections below.

## 2 System Architecture

This project implements a client-server architecture that uses Android Operating System (OS) with JavaScript Object Notation (JSON) and Remote Procedure Calls (RPC) communication. The Android OS is specifically designed for mobile devices, and it is developed and trademarked by Google's Android Developers.<sup>2</sup> JSON, which is an alternative for Extensible Markup Language (XML), is a lightweight data-interchange format that uses readable text to transmit data objects. JSON is widely used in Java or Android applications for communication and exchange of data over the internet (Nurseitov et al. 2009). On the other hand, RPC is a powerful technique for constructing distributed, client-server based applications that extends local procedure calling and provides transfer of data across the communication network. Such architecture is an ideal environment for mobile devices, especially

---

<sup>1</sup><https://cloud.google.com>, accessed 5 April 2014.

<sup>2</sup><http://developer.android.com>, accessed 5 April 2014.

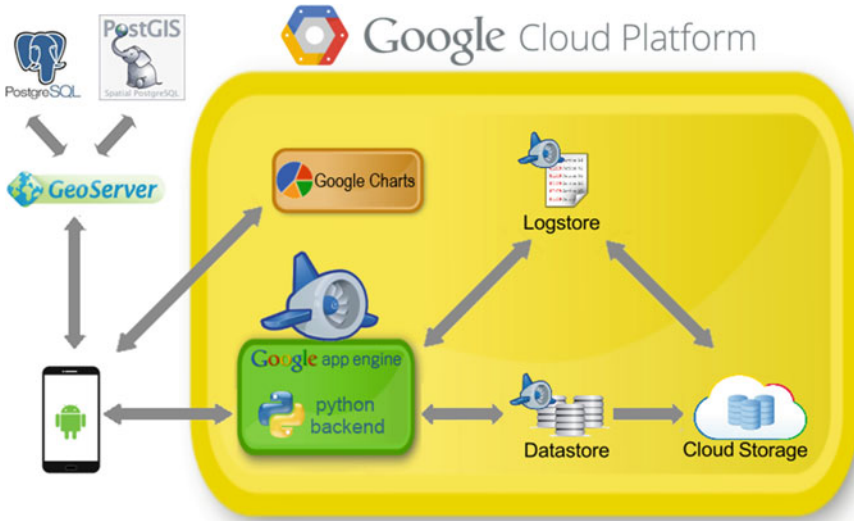


Fig. 1 System architecture

because it offers powerful built-in services. Such services include a large storage capacity, computational capabilities, and connectivity used for continuous data manipulation and interaction, which are deployed by minimal customization and coding.

Figure 1 shows the conceptual framework of the proposed client-server architecture and the connectivity between the components that comprise the framework. The Android OS is used for the client side, while Google Cloud Platform Services are used for the server side. The Google Cloud Platform Services integrate the Google App Engine that provides a cloud computing platform for developing and hosting web or mobile applications in Google-managed data centers. In addition, the Google Charts, which are used for creating charts for the Android interface, are also services provided by Google Cloud Platform. The Google Chart API's are tools for creating charts from the user input data which are consequently embedded in the web application. Finally, GeoServer<sup>3</sup> completes the set of tools in this architecture that is used to add geospatial capabilities for serving custom maps to the client (Youngblood 2013). In our application, the spatial database extender PostGIS<sup>4</sup> was used to connect GeoServer with PostgreSQL,<sup>5</sup> which is an open source object-relational database management system that stored the decision criteria layers associated with the proposed alternatives (Krosing 2013; Llarío 2013). The following subsections will describe each part of the implemented architecture.

<sup>3</sup><http://geoserver.org>, accessed 8 April 2014.  
<sup>4</sup><http://postgis.net>, accessed 8 April 2014.  
<sup>5</sup><http://www.postgresql.org>, accessed 8 April 2014.



## 2.1 *Client Design*

The client side for this project aims to develop an interactive mobile application where participants can request a variety of integrated services including: using Google and custom-built maps; accessing a forum to interact with other participants; implementing a multi criteria decision analysis and data processing tool; and visualizing of real-time graphical chart data. The proposed approach uses an Android OS development framework for the client side, which is an open source mobile operating system created by Google with a large community of active developers. Google provides developers with an Android Software Development Kit (Android SDK), for building Application (App) for Android-powered devices. The Android App development used Java<sup>6</sup> programming language with Eclipse<sup>7</sup> as the integrated development environment (IDE). The Android Development Tools (ADT) Plugin and Android SDK were integrated for the development of the PPGIS interface. Android OS provides developers with extensive tutorials and a great variety of examples completely free in order to make the most out of its operating system. In addition, Google also provides design recommendations with an objective of creating intuitive, easy to use, and visually attractive mobile applications. Some of the design recommendations relate to icons design, animations, performance enhancements, layouts design, dashboards design, and Google Maps integration.

## 2.2 *Server Design*

Google Cloud Platform was used for the server side in conjunction with the Google App Engine, which is a cloud computing platform. The Google App Engine is a Platform as a Service (PaaS) where developers can host and execute their applications in Google managed data centers. Google App Engine is designed for real-time dynamic application and offers multiple services such as social networking sites, mobile applications, survey applications, project management, collaboration, publishing, and other traditional website content (i.e., documents and images) (Sanderson 2012). The App Engine involves three parts: the runtime environment, the data management tool (i.e., datastore), and the scalable service which is the management console.

The runtime environment initiates into existence when the request handler begins (client contacts the application with an HTTP request) and disappears when it ends. The App Engine receives the request from the domain/subdomain name that is used for registering and setting up Google Apps and distributes traffic among multiple servers by giving every request the same treatment. Applications run in a secure environment, where requests are distributed across multiple servers and applications

---

<sup>6</sup><http://www.java.com>, accessed 31 March 2014.

<sup>7</sup><http://www.eclipse.org>, accessed 31 March 2014.

run within its own secure, reliable environment that is independent of the hardware, operating system, or physical location of the server. Python<sup>8</sup> is one of the programming languages supported by Google App Engine and used with Python Software Development Kit (SDK) in this project (Fig. 1). The Python datastore is the tool used for storing and retrieving data generated by applications running on this platform. The datastore resembles an object database design which differs from traditional relational databases models. A typical arrangement for a small project involves a single database server, and one or more web clients that connect to the database to store or retrieve data. As with the runtime environment, the datastore allows App Engine to handle the details of distributing and scaling the application for better code performance. The advantage of the datastore is to optimize computational performance for applications experiencing high traffic. The Google Query Language (GQL), which is simplified version of the Structured Query Language (SQL), is used to provide an alternative method for accessing the data.

Finally, the Google App Engine contains a management console that gives administrators full control of their application. For instance, some of the main features in the Administration Console are used to manage the application and view its access, resource usage, statistics, and message logs. In addition, the console gives access to real-time performance about data application usage as well as access to log data emitted by the use of the application.

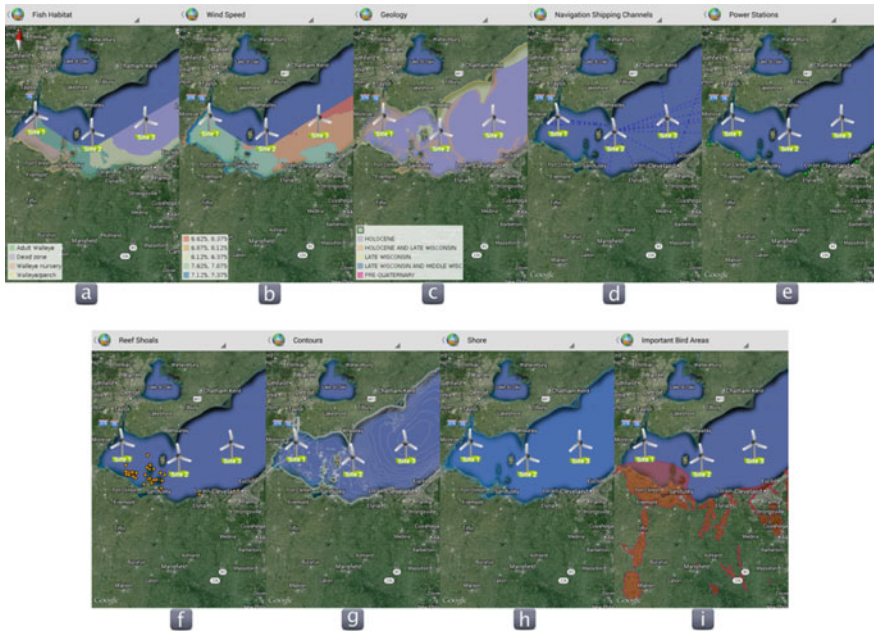
### ***2.3 Geospatial Integration and Study Area***

The geospatial data was acquired for a hypothetical study area located in northern Ohio along the western Lake Erie shore. The study area for this project encompasses mostly the western basin of Lake Erie. The Western Erie Basin represents just a small subset of the entire Great Lakes system, which is the largest fresh surface water system in the world. The long, narrow orientation of Lake Erie parallels the direction of the prevailing southwest winds while the western basin is relatively shallow having an average depth of 7.4 m and a maximum depth of only 19 m. The study area is characterized by strong winds with an annual average speed of 7–7.5 m/s, which is highest in November and lowest in July, and is favorable for offshore wind farm development. The characteristics surrounding the lake are unique where hydrology, soils, vegetation, land uses and land cover of the study area are highly variable in space and time that contribute to the unique wetlands that create dwelling and migratory habitats for a variety of avian species. The Ottawa National Wildlife Refuge (ONWR),<sup>9</sup> which is known for its rich biodiversity, is part of the study area and includes different species such as raptors including bald eagle and osprey, bats, waterfowl, wading birds, shorebirds, gulls and terns, and

---

<sup>8</sup><https://www.python.org>, accessed 13 July 2014.

<sup>9</sup><http://www.fws.gov/refuge/ottawa>, accessed 19 March 2014.



**Fig. 2** Study area and decision criteria **a** fish habitat, **b** wind speed, **c** geology, **d** navigable waterways, **e** power stations, **f** reef shoals, **g** bathymetry, **h** distance from shore, and **i** important bird habitat

perching birds which use the surrounding wetlands for migration, stopover, nesting, and feeding (Mekonnen 2014; Mekonnen and Gorsevski 2015; Mirzaei et al. 2011, 2012a, b, c). The large concentration of population is another factor associated with the lake and which needs to be considered in wind energy production.

Figure 2 also shows multiple criteria data layers of the study area and the three site locations (decision alternatives) considered for the selection of a new wind farm development. The background map information in the Android client application is powered by the Google Maps interactive map interface (API). The Google Maps were used because of general public’s familiarity with the interface and its potential to allow higher participation in such complex decision making planning settings. In this demonstration, the criteria that were used are limited to currently available data which address different wind farm development planning issues that impede the decision making process. Table 1 shows a brief description of the criteria used in the client. However, (ODNR 2012) the methodology is flexible and other decision criteria could be considered for different site-specific purposes. In this project the Android client interface used a Web Map Service (WMS) protocol that handles HTTP requests in order to retrieve geo-registered map layers generated by the GeoServer mapping server and the PostgreSQL spatial database.

**Table 1** Evaluation criteria for wind farms

Criteria	Description
Fish habitat	These areas are habitat for larval and young-of-year fishes that support different classes of fish including walleye nursery, adult walleye, walleye/perch, and dead zone (absence of habitat). The walleye nursery class is considered as the least suitable, adult walleye and walleye/perch are moderately suitable, while the dead zone class is the most suitable habitat for wind farms
Wind speed	Energy output of wind turbines increases as wind speeds increase until nominal wind speed is reached, which is the speed that maximizes the energy production. Therefore, areas classified with higher wind speeds are more suitable than areas classified with lower speeds. The data is partitioned in four categories of annual average wind speeds classified by the NREL ranging from poor (1) to good (4).
Geology	Different types of geology can affect the installation costs especially foundation design of a wind farm when a potential location does not contain adequate lakebed conditions that can support large structures like wind turbines
Navigable waterways	Offshore wind sites that are located further away from navigable waterways are more suitable and they will not affect any transportation routes. The data is organized as buffers at 1, 2, and 3 miles from existing navigable waterways
Power stations	The proximity to power stations and transmission lines is an important consideration for wind farm development for minimizing the cost of delivered electricity to the consumer. Point data was used to represent the spatial locations of existing power stations.
Reef Shoals	The reef effects are important to local fish habitats because create opportunities for larvae and other foods in a sheltered-like habitat. Point data was used to represent the spatial locations of existing reef shoals
Bathymetry	The bathymetry is the measurement of the depth of water in the lake and important criteria that affects the installation and transmission costs. Also, it is a limiting factor for the size of vessels that are used for transporting the wind turbine components during the installation phase
Distance from shore	As the distance from shore increases the visibility of the farm diminishes and potentially public approval is assumed to increase. On the other hand, increased distances affect transmission and efficiency costs as well as making operation and maintenance more challenging and restrictive
Important bird habitat (IBA)	The significance of designated important bird habitat is intended to minimize collisions and mortality of birds and bats by operating wind turbines. A major concern is avian collisions near bird habitats and migratory routes and the change of air pressure around a wind turbine that is fatal especially for bats. A digital map published by the Ohio Audubon Society (OAS) (2009) that depicts all the IBAs located in the state was used to represent the locations of those areas

The three decision alternatives for participant's consideration are shown in Fig. 2, where "Site 1" is located near Maumee Bay, "Site 2" is located east of Kelley's Island, and "Site 3" is located off the far northwestern corner of Lorain County. These decision alternatives were identified since they fulfill the wind resource required for offshore wind farm development, which is at least 7 m/s at a turbine height of 90 m above the lake surface. To demonstrate the functionalities of this prototype tool and to illustrate the potential application for evaluation of alternatives for wind farm selection, the simple hypothetical scenario is detailed through the description of the Android client interface.

### 3 Wind Farm Siting: Implementation of an Android Spatial Decision Support Tool

#### 3.1 User Interface

The intention of the prototype was to provide a simple interface for mobile applications and non-experienced GIS users. The main components that comprise the prototype were organized under different themes, including a main page, a criterion selection and mapping tool, discussion forum, and a decision tool for voting and communicating results.

Figure 3a shows the default parent application page that is used for smooth, slider-like transitions and navigation between different subset levels and to acquaint the participants with the project and the process of selection of alternatives for suitable offshore wind farms. The dashboard layout of patterns, which is a Google I/O (2013) design, is implemented to simplify the navigation process. The grid-like organization of the dashboard layout is used to automatically organize the elements in vertical and horizontal screen orientation and to allow optimized view for accessing the modules.

Figure 3b is the mapping interface that contains basic mapping functionalities. The mapping interface uses a combination of Google Maps and GeoServer capabilities to provide users with visual information of the study area and the available decision criteria. The contribution from the GeoServer, which is an open source server for sharing geospatial data, is that it allows for customization of different spatial raster or vector layers. For instance, this module allows for the visualization of the geospatial data layers in Fig. 2. In the figure the first layer (Fig. 2a) is the "Fish Habitat" layer showing the legend and the distribution of the four main species, including walleye nursery, adult walleye, walleye/perch, and dead zone, each of which is separated by different colors. In addition, Fig. 3b is the initial screen for the voting module that is used for ranking of the decision alternatives associated with the selection of a given wind farm location. The voting tool is the one of the most important components of the system that is used for collecting ballots from the participants. Here, the participants can vote on the importance of



**Fig. 3** Different android components **a** main page, **b** mapping interface, **c** forum module, **d** results from the votes associated with the best three alternatives, **e** results from the most preferred criteria from all the participants, and **f** the help module

criteria (i.e., select the top five) and rank the alternatives. The details and the steps for implementing the voting module are further discussed below.

Figure 3c represents the forum module that is used to facilitate discussion among participants for exchanging views and ideas such as impacts on the local community as influenced by the selection of different criteria and alternatives. The module allows for an ongoing, asynchronous discussion, so it can help each participant to better formulate his or her opinion. In addition, in a case of moderated

discussions that take place before the voting process occur, the module can increase the background knowledge of the participants, which may evolve their personal opinion related to the wind farm siting issue.

The basic implementation of this module uses ProBoards,<sup>10</sup> which is the largest host of free forums on the internet. The main themes used in this research include general discussion, announcements for facilitating the decision process, and a feedback section that enables generation of new ideas or concerns through each participant's responses.

Figure 3d, e show the results from the votes associated with the best three alternatives and the most preferred criteria from all the participants. The scores are generated when the Android application client connects to the Python backend server deployed on Google App Engine, using the RPC method that is embedded directly in the Uniform Resource Locator (URL) and executed with an HttpClient. The response from the Google App Engine is then represented by JSON data-interchange format, which is parsed and analyzed for subsequent integration of the Google Charts module that accommodates real-time visualization of the results. Finally Fig. 3f shows the help and tutorial module that contains a brief explanation of how to utilize the different components in this prototype. Additional information such as basics of wind energy, pros and cons associated with wind farms, the basis of the PPGIS concept and voting are also included in this section. For instance, although the voting process is intuitive and includes simple instructions that guide participants, the tutorial section integrates small thumbnails that can be executed by a simple click to animate the steps and the process for the participants.

### ***3.2 Voting and Decision Making Process***

The flow chart in Fig. 4 highlights the data flow process behind the voting tool. After the participant initiates the voting module, the decision module requires the participant to select the five most important criteria, which are used for the ranking of the three decision alternatives. Before the information is submitted to the database that is the Google App Engine Datastore, the system computes the scores and checks for errors such as required selection of criteria and ranking information. Figure 5 shows the sequential steps for the voting process. The individual process starts with a blank form Fig. 5a that requires a selection of a total of five criteria. For simplified use of the form, the current implementation limits the selection to the top five most important criteria, but this can be altered based on different needs and requirements. When a participant selects the five criteria, the "Next" button located at the bottom right of the screen in Fig. 5b is activated for initiating the ranking process. The individual steps for casting the votes for the five criteria are shown in Fig. 5c through d. For each criterion, the most important alternative from the

---

<sup>10</sup><http://www.proboards.com>, accessed 13 July 2014.

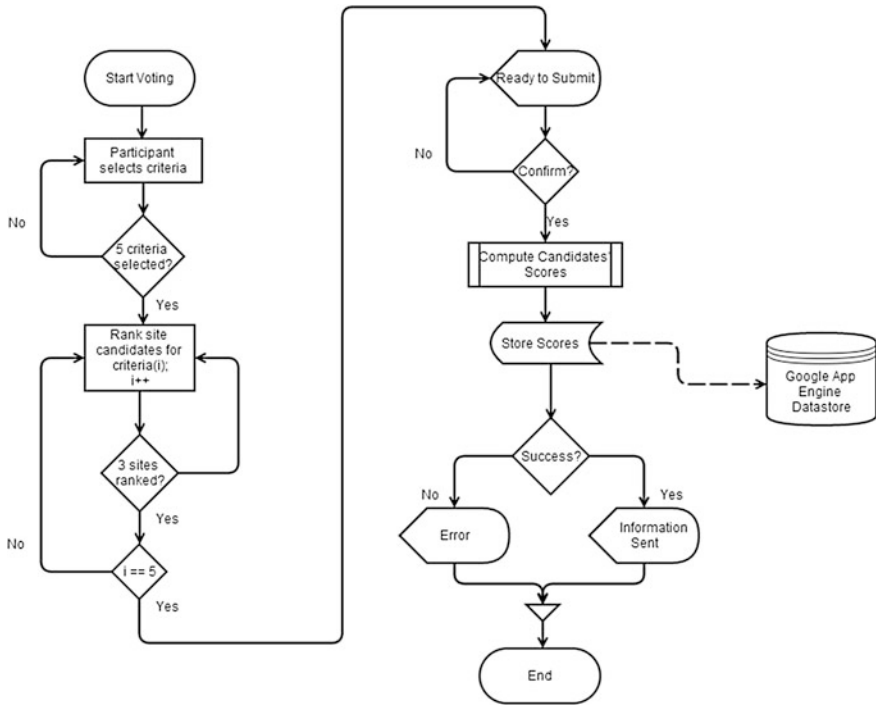


Fig. 4 A flow chart of the voting process

ranking process is placed first while the least important alternative is placed third by each participant. For example, Fig. 5d displays a ranking outcome where the participant ranked “Site 2” as most important for the “Wind Speed” criterion, “Site 1” is ranked on the second position, and “Site 3” is ranked on the third position or being the least important alternative.

### 3.3 Calculation of Suitability Scores

In this study, the decision alternatives were ranked from first to third position in terms of preferred importance for each of the selected five criteria. The Pairwise comparison voting method was used to select the best alternative by treating the comparison as a series of paired alternatives where the preferred alternative gets one point and in the case of a tie each alternative gets one half of a point. The participants’ ranking preferences for each criterion (i.e. Fig. 5g) are first computed by the Pairwise comparison before individual evaluation scores are aggregated into a group score that would yield the solution for the best three alternatives.



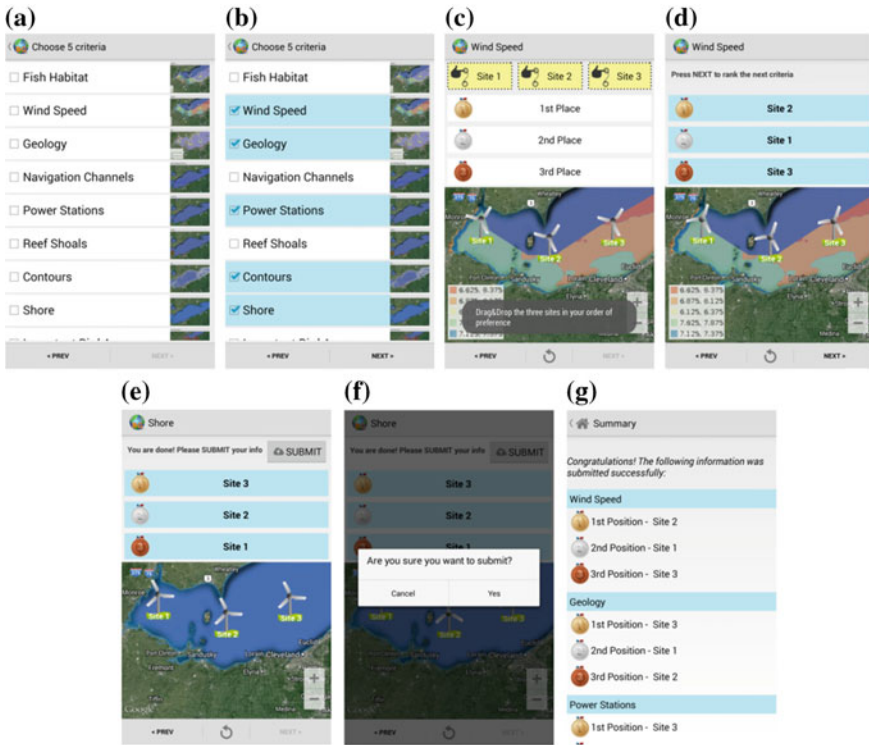


Fig. 5 Sequential steps of the GUI for the voting process

To get a clear understanding of this Pairwise comparison, the method is further explained using a numerical example. For a set of  $n$  decision criteria, the formula for the total comparison of each possible pair is  $n(n - 1)/2$  that determines the number of comparisons for selecting the best alternative from the Pairwise comparison voting method. For instance, 3 alternatives Pairwise comparison requires a total of  $3(3 - 1)/2 = 3$  comparisons. Thus in our example the comparison includes the following pairs: “Site 1” versus “Site 2”, “Site 1” versus “Site 3”, and “Site 2” versus “Site 3”.

Figure 6 illustrates an example from the calculation using the Pairwise comparisons method. As shown in the figure, the score for alternative ‘Site 1’ is the highest and 2 points are assigned based on the vote from a single participant. For instance, the comparison between “Site 1” versus “Site 2” shows that “Site 1” has received higher voting ranks for the Fish Habitat (FH), Power Stations (PS), and Distance from Shore (DS) while “Site 2” has received higher voting ranks for the Wind Speed (WS) and Reef Shoals (RS). Finally, the scores by personal preferences from the participants are summed to produce a group solution and determine the best alternative.

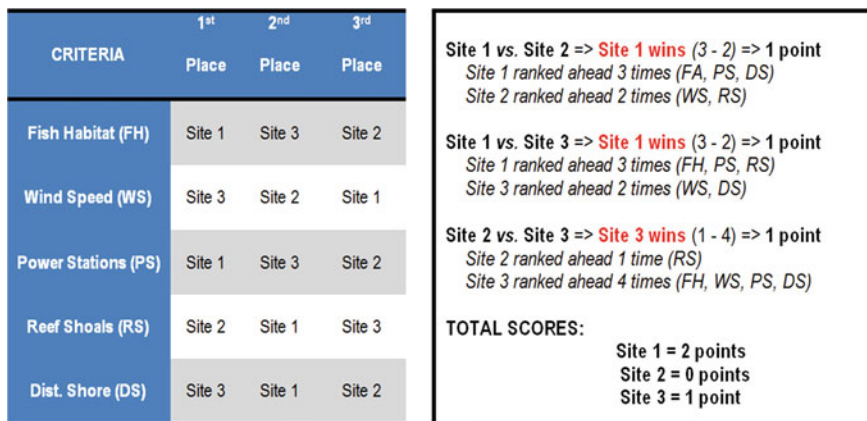


Fig. 6 Participant’s vote example and score calculation using Pairwise comparison method

Figure 3d shows an example of the group results associated with the three alternatives and generated by participant’s votes. For example ‘Site 3’ has received the highest score from the aggregated result from all participants’ votes. The chart in Fig. 3e shows the importance of the criteria valued by the participants. The figure shows that the most preferred criterion for all participants is the Power Stations, while Important Bird Areas and Fish Habitat are the most controversial criteria used for the selection of this hypothetical example. It is also important to note in this particular example that the summed solution used equal importance for each criterion but a weighted solution could be generated based on the selection of the most preferred criteria. Of course, a successful implementation of this prototype requires an adequate selection of participants and stakeholders, which is one of the critical considerations to maintain justice, equity and trust in the voting process (Devine-Wright 2005; Gorsevski et al. 2013; Wolsink 2000). For example, for implementation of local level policy-making, key participants and stakeholders can vary by location and policy regulations that are in place, but adequate selection should involve interest groups who are directly affected by a decision and its planning consequences.

## 4 Conclusions

This research presents an application of a smartphone-based Android approach to enhance public participation for Multi-Criteria Evaluation for assessing decision alternatives for a potential wind farm development. The potential implementation was illustrated by using a hypothetical case study to show the strengths and the benefits from this Android driven PPGIS in facilitating suitable offshore wind farm site selection in Lake Erie. The hypothetical case study demonstrated a standard decision-making scenario by ranking three predefined sites or decision alternatives using nine different spatial criteria.

The Android client application prototype provides multiple modules to guide the participants during the decision-making process by integrating simplified user-friendly graphical interface using familiar Google Maps and GIS capabilities that are simplified for non-expert users. Some of the featured modules integrated in the Android client include a mapping tool, a discussion forum, a voting and decision-making tool, and results modules. The mapping tool is used for visualization, exploration and comparison of decision alternatives and their corresponding properties, such as total score and rank. The discussion forum is used to facilitate communication and debate among participants while the voting and decision-making tool is used to perform ranking of the decision alternatives based on the evaluation criteria using the Pairwise comparison method. The calculated individual scores for each decision alternative from the Pairwise comparison method are used for subsequent group scores. In this hypothetical example, the relative importance of the criteria was not considered in the final solution, but weighed outcome along with sensitivity analysis ought to be the focus of future development. Finally, the participants can visualize the voting results at any time by accessing the results modules which shows charts that are generated in real-time.

The implementation of the server side framework used the Google App Engine application called Mobile Backend Starter that is a part of the Google Cloud Platform. The simplified server-side development provides a number of cloud services such as easy-to-use libraries for storing data in the cloud, sending device-to-device push notifications, event driven programming, user authentication, and infrastructure to accommodate scalability with a multi-user environment. The Google APIs libraries for Python were used to improve efficiency with data interchange and to store the voting results from the JSON-RPC communication.

The evaluation criteria used in the case study were limited to currently available data but the presented methodology is flexible, so different evaluation criteria could be added based on site specific problems and requirements. In addition, the proposed methodology could be used to determine and evaluate decision alternatives for other conflicting issues that affect decision-making and planning consequences. For instance, other conflicting issues that can be incorporated may include criteria that address problems related to landscape aesthetics, turbine noise, avian impact, shadowing and flickering and other environmental or socioeconomic issues. The main aim of this study was to show the potential of this Android-based PPGIS for offshore wind suitability analysis; another key objective was to present the potential of this integrated tool that can facilitate effective public involvement and it can be used for complex planning problems and building consensus. Moreover, the synergistic potential that integrates mobile technologies to facilitate decision-making through group collaboration and flexible problem-solving environments creates opportunities for robust public-private collaborations and formulation of public policies that consider socio-political influence. Thus the contribution of the proposed approach is to demonstrate a new decision making tool that can increase the potential of a participatory planning process, especially through empowerment of key players who are directly affected by a decision and its planning consequences at the local level.

**Acknowledgments** This investigation was supported by the US Department of Energy, Coastal Ohio Wind Award Number DE-FG36-06GO86096 and the Coastal Ohio Wind Project for Reduced Barriers to Deployment of Offshore Wind Energy Award Number: DE-EE0003871.

## References

- Al-Yahyai S, Charabi Y, Gastli A, Al-Badi A (2012) Wind farm land suitability indexing using multi-criteria analysis. *Renew Eng* 44:80–87
- Anon (n.d.) Google App Engine—Google Developers. <https://developers.google.com/appengine/>. Accessed 31 July 14
- Aydin NY, Kentel E, Duzgun S (2010) GIS-based environmental assessment of wind energy systems for spatial planning: a case study from Western Turkey. *Renew Sustain Eng Rev* 14 (1):364–373
- Baban SMJ, Parry T (2001) Developing and applying a GIS-assisted approach to locating wind farms in the UK. *Renew Eng* 24(1):59–71
- Berry R, Higgs G, Fry R, Langford M (2011) Web-based GIS approaches to enhance public participation in wind farm planning. *Trans GIS* 15(2):147–172
- Bishop ID, Stock C (2010) Using collaborative virtual environments to plan wind energy installations. *Renew Eng* 35(10):2348–2355
- Boroushaki S, Malczewski J (2008) Implementing an extension of the analytical hierarchy process using ordered weighted averaging operators with fuzzy quantifiers in ArcGIS. *Comput Geosci* 34(4):399–410
- Boroushaki S, Malczewski J (2010a) Using the fuzzy majority approach for GIS-based multicriteria group decision-making. *Comput Geosci* 36(3):302–312
- Boroushaki S, Malczewski J (2010b) Measuring consensus for collaborative decision-making: a GIS-based approach. *Comput Environ Urban Syst* 34(4):322–332
- Devine-Wright P (2005) Beyond NIMBYism: towards an integrated framework for understanding public perceptions of wind energy. *Wind Eng* 8(2):125–139
- Donevska KR, Gorsevski PV, Jovanovski M, Peševski I (2012) Regional non-hazardous landfill site selection by integrating fuzzy logic, AHP and geographic information systems. *Environ Earth Sci* 67(1):121–131
- Farfán MA, Vargas JM, Duarte J, Real R (2009) What is the impact of wind farms on birds? A case study in southern Spain. *Biodivers Conserv* 18(14):3743–3758
- Gorsevski PV, Cathcart SC, Mirzaei G, Jamali MM, Ye X, Gomezdelcampo E (2013) A group-based spatial decision support system for wind farm site selection in Northwest Ohio. *Eng Policy* 55:374–385
- Gorsevski PV, Donevska KR, Mitrovski CD, Frizado JP (2012) Integrating multi-criteria evaluation techniques with geographic information systems for landfill site selection: a case study using ordered weighted average. *Waste Manag* 32(2):287–296
- Hurtado JP, Fernández J, Parrondo JL, Blanco E (2004) Spanish method of visual impact evaluation in wind farms. *Renew Sustain Eng Rev* 8(5):483–491
- Jankowski P, Nyerges T (2001) GIS-supported collaborative decision making: results of an experiment, vol 91(1). pp 48–70
- Krosing H (2013) PostgreSQL server programming. Packt Publishing, Birmingham
- Llario JCM (2013) PostGIS 2 Análisis Espacial Avanzado, 1st edn. CreateSpace Independent Publishing Platform, Lexington, KY
- Malczewski J (1999a) GIS and multicriteria decision analysis. Wiley
- Malczewski J (1999b) GIS and multicriteria decision analysis. Wiley
- Malczewski J (2006a) GIS-based multicriteria decision analysis: a survey of the literature 20 (7):703–726

- Malczewski J (2006b) Ordered weighted averaging with fuzzy quantifiers: GIS-based multicriteria evaluation for land-use suitability analysis 8(4):270–277
- Mekonnen AD (2014) Wind farm site suitability analysis in lake erie using web-based participatory GIS (PGIS). Bowling Green State University. [https://etd.ohiolink.edu/ap/10?0::NO:10:P10\\_ACCESSION\\_NUM:bpsu1392975809](https://etd.ohiolink.edu/ap/10?0::NO:10:P10_ACCESSION_NUM:bpsu1392975809). Accessed 01 Aug 14
- Mekonnen AD, Gorsevski PV (2015) A web-based participatory GIS (PGIS) for offshore wind farm suitability within Lake Erie, Ohio. *Renew Sustain Eng Rev* 41:162–177
- Mirzaei G, Majid MW, Jamali MM, Ross J, Frizado J, Gorsevski PV and Bingman V (2011) The application of Evolutionary Neural Network for bat echolocation calls recognition. In: Paper presented at the 2011 international joint conference on neural networks (IJCNN), pp 1106–1111
- Mirzaei G, Majid MW, Ross J, Jamali MM, Gorsevski PV, Frizado JP and Bingman VP (2012a) Avian detection amp; tracking algorithm using infrared imaging. In: Paper presented at the 2012 IEEE international conference on electro/information technology (EIT), pp 1–4
- Mirzaei G, Wadood Majid M, Bastas S, Ross J, Jamali MM, Gorsevski PV, Frizado J and Bingman VP (2012b) Acoustic monitoring techniques for avian detection and classification. In: Paper presented at the 2012 conference record of the forty sixth asilomar conference on signals, systems and computers (ASILOMAR), pp 1835–1838
- Mirzaei G, Wadood Majid M, Ross J, Jamali MM, Gorsevski PV, Frizado J and Bingman VP (2012c) Implementation of ant clustering algorithm for IR imagery in wind turbine applications. In: Paper presented at the 2012 IEEE 55th international midwest symposium on circuits and systems (MWSCAS), pp 868–871
- Nurseitov N, Paulson M, Reynolds R and Izurieta C (2009) Comparison of JSON and XML data interchange formats: a case study. In: ISCA 22nd international conference on computer applications in industry and engineering. paper presented at the CAINE 2009. San Francisco, CA
- Nyerges TL, Jankowski P (2009) *Regional and Urban GIS: a decision support approach*, 1st edn. The Guilford Press, New York
- ODNR = Ohio department of natural resources (2012) Offshore wind energy, Resources, <http://www.ohiodnr.com/LakeErie/WindEnergyRules/tabid/21234/Default.aspx> (accessed 10/12/12)
- Rodrigues M, Montañés C, Fueyo N (2010) A method for the assessment of the visual impact caused by the large-scale deployment of renewable-energy facilities. *Environ Impact Assess Rev* 30(4):240–246
- Saeed A, Bhatti MS, Ajmal M, Waseem A, Akbar A and Mahmood A (2013) Android, GIS and web base project, emergency management system (EMS) which overcomes quick emergency response challenges. In: Rocha Á, Correia AM, Wilson T and Stroetmann KA (eds) *Advances in information systems and technologies*. Springer Heidelberg, pp 269–278. <http://link.springer.com/chapter/>. doi:10.1007/978-3-642-36981-0\_26. Accessed 31 July 14
- Sanderson D (2012) *Programming Google app engine*, 2nd edn. O'Reilly Media, Sebastopol, CA
- Simão A, Densham PJ and (Muki) Haklay M (2009) Web-based GIS for collaborative planning and public participation: an application to the strategic planning of wind farm sites. *J Environ Manage* 90(6): 2027–2040
- Tegou L-I, Polatidis H, Haralambopoulos DA (2010) Environmental management framework for wind farm siting: methodology and case study. *J Environ Manage* 91(11):2134–2147
- Voivontas D, Assimacopoulos D, Mourelatos A, Corominas J (1998) Evaluation of renewable energy potential using a GIS decision support system. *Renew Eng* 13(3):333–344
- Weng Y-H, Sun F-S, Grigsby JD (2012) Geotools: an android phone application in geology. *Comput Geosci* 44:24–30
- Wolsink M (2000) Wind power and the NIMBY-myth: institutional capacity and the limited significance of public support. *Renew Eng* 21(1):49–64
- Yeo I-A, Yoon S-H, Yee J-J (2013) Development of an environment and energy geographical information system (E-GIS) construction model to support environmentally friendly urban planning. *Appl Eng* 104:723–739
- Youngblood B (2013) *Geoserver beginner's guide*. Packt Publishing, Birmingham

# Presenting Citizen Engagement Opportunities Online: The Relevancy of Spatial Visualization

Thore Fechner and Christian Kray

**Abstract** Public administrations and cities increasingly use modern information and communication technologies to enhance their processes and services. As the fabric of cities becomes more complex, and collaboration and participation are emphasized, citizens thus need to be empowered to find available engagement opportunities and citizens need to identify those that they want to engage with. We report on the use of an online information platform that offered citizen engagement opportunities in a traditional textual form and via an interactive geo-visualization. The platform was deployed in a real-world study and integral component in a campaign to raise volunteer engagement in a medium-sized German city. We first introduce our approach to letting citizens explore engagement opportunities and follow up with an analysis of how people used the platform. Subsequently, evidence is presented and discussed that spatial visualization and interaction is relevant for informing citizens online. Since we released the information platform as open source, others can easily benefit from our insights.

**Keywords** Geo-visualizations · Maps · Citizen engagement

## 1 Introduction

Citizens are called upon to partake in political processes on different levels as public administrations face several challenges. Creating livable and sustainable environments or calls for greater transparency and participation are just some of them. Modern information and communication technologies (ICTs) increasingly govern the interaction between public administrations, cities, and citizens. ICTs can also

---

T. Fechner (✉) · C. Kray  
Institute for Geoinformatics, WWU Münster, Heisenbergstraße 2,  
48149 Münster, Germany  
e-mail: t.fechner@uni-muenster.de

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

© Springer International Publishing Switzerland 2016  
T. Sarjakoski et al. (eds.), *Geospatial Data in a Changing World*,  
Lecture Notes in Geoinformation and Cartography,  
DOI 10.1007/978-3-319-33783-8\_7

help to lay the foundation for smart cities (Ferro et al. 2013) and are an essential part of open government movements (Maier-Rabler and Huber 2011). Both movements play a crucial part in tackling the challenges cities, and public administrations are facing.

One key question in this context is how information should be disseminated and presented to citizens digitally. Before any citizen engagement can happen, citizens need to be aware of engagement opportunities.

In this article, we thus discuss the relevancy of spatial visualization and interaction for informing citizens about engagement opportunities. Based on a real-world deployment we examine the use of an online portal that informed citizens about engagement opportunities. The portal was an integral part of a campaign of 25 non-governmental organizations (NGOs) to raise volunteer engagement in a medium-sized German city. Two distinct and interactive ways of representing engagement opportunities were present: A classical grid-based mosaic-view with photographs and accompanying textual information, and an interactive geo-visualization that tightly coupled the textual information with the geo-visualization.

Our contributions in this article are threefold: (i) We propose an approach to tightly couple spatial interaction and visualization with textual information about engagement opportunities. (ii) We present evidence that spatial interaction and visualization is relevant for citizens while informing themselves about engagement opportunities. Our analysis is based on a two and a half month real-world study within a medium-sized German city. We also discuss potential reasons why citizens used the geo-visualization to explore available engagement opportunities, and why they did not use the available feedback and discussion component of the information portal. (iii) We provide our implementation as open-source application.<sup>1</sup>

In the remainder of this article, we first discuss the theoretical background and review related work in the fields of citizen engagement and Geographic Information Systems (GIS) in general. Section 3 present the use-case. Details on the study and the citizen information portal we implemented are outlined in Sect. 4. Results are reported in the subsequent Sect. 5 and discussed afterward, looking at limitations of our study and approach. We conclude with a summary and by outlining future work.

## 2 Related Work

The following section provides a short review of related work in the area of participation, citizen engagement and GIS. Furthermore, we provide the motivation to use interactive geo-visualizations to inform citizens about engagements opportunities by examining the intertwined nature of the spatial dimension and engagement.

---

<sup>1</sup><https://github.com/ubergesundheit/dialogmap>.

## 2.1 *Participation and Geographic Information Systems*

Research regarding GIS and participation is mostly being conducted in the area of Public Participation GIS or Participatory GIS (hereafter PPGIS). Early PPGIS research focused on GIS technology and ease of use to involve the public in the process of official decision-making (Sieber 2006; Obermeyer 1998). Most PPGIS use-cases revolve around aspects of urban planning or resource management. Due to the aim to involve the public in the decision-making process the concept of PPGIS has been investigated from several viewpoints: PPGIS role in governance (McCall and Dunn 2012), its potential for democratization of spatial decision-making (Dunn 2007), and forming consensus in a structured process (Bailey and Grossardt 2010) were discussed.

By relying on digital technology the risks to digitally divide and marginalizes parts of the population is present. Elwood (2006b) discusses barriers that hinder the use of PPGIS systems and how diverse PPGIS are used by experts and non-experts alike.

Public participation, engagement or collaboration are often mentioned together in the context of smart cities or in open government initiatives. While a vast body of knowledge exists that examines different forms of participation, deliberation, and engagement (see Carpini et al. 2004 for a survey) it is important to note that technological openness does not imply political openness that allows citizen engagement (Yu and Robinson 2012). Nonetheless, citizen participation tends to produce positive effects as a meta-case study shows (Gaventa and Barrett 2010).

An early and well-known topology for different levels of citizen participation is provided by Arnstein (1969). She uses the analogy of a ladder to describe eight levels of increasing citizen participation—ranging from categories like non-participation, tokenism to citizen power, increasing with each step on the ladder. Her initial concept and analogy was adopted and updated by several researchers (e.g. Connor 1988; Steinmann et al. 2005). Rowe and Frewer (2005) established a “working model” topology to overcome the imprecise usage of public participation and review different mechanism classes for public engagement.

## 2.2 *Spatial Dimension and Engagement*

In general, geo-visualizations can foster communication, facilitate rapid insights into what is known by whom, how it is understood, and they encourage reasoning (Hopfer and MacEachren 2007; Andrienko et al. 2007, 2010). Maps are ubiquitously present to provide information at our fingertips (Weiser 1991), e.g., on our smartphones to help us navigate, allow us to contextualize information on web pages or visualize facts. Elwood (2006a) notes that the spatial analysis function is not the most valuable function of PPGIS systems for activists and NGOs. Instead, they produce cartographic spatial narratives to support political projects and characterizations.



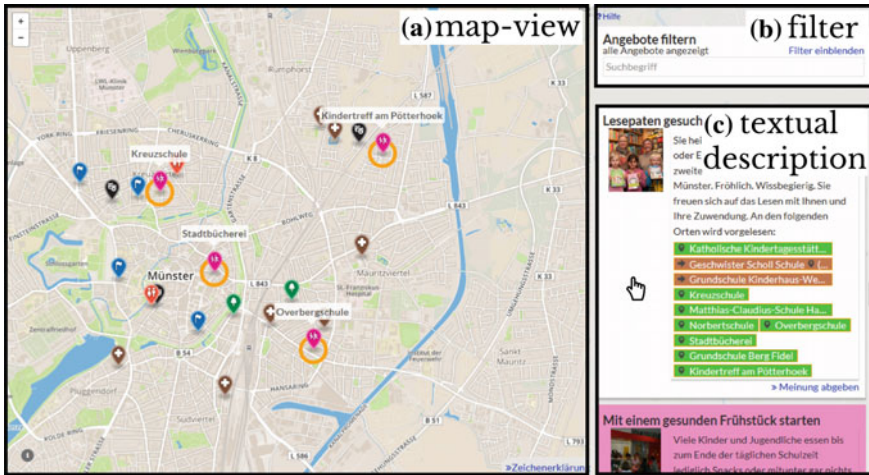
The role of the spatial dimension for citizen engagement is also documented for local citizen activism: Taylor et al. (2015) report on how data is tied to a place, both in terms of physical and social geography. They reported on engagement projects for civic activism in one specific neighborhood and observed tight knight relations between people, place and data and how data materializes differently to people and places. Similarly, other reports (Crivellaro et al. 2015) highlight how the spatially situated discovery of issues can connect city residents with the processes of their cities. Hecht and Gergle (2010) show that user generated content has a strong link to space and place. Due to this, we consider space and its depiction in the form of geo-visualizations as a well-suited medium to convey a message or narrative. It can form part of a canvas that relays or highlights information—similar to Sui and Goodchild (2011) who recognize that GIS technology can be used as a medium and that it converges with social media.

The idea to present engagement opportunities spatially to citizens is motivated by the observation that geo-visualizations can help to experience and explore content by providing structure to the experience (Elwood and Leszczynski 2013). We designed the geo-visualization that is presented in Sect. 3 around the notion that *spatial and textual interaction should equally help to structure, explore and filter the content*. While all functionalities and components that we use are well-known and much used individually, we argue that their combination and composition substantially increase their value and usefulness.

### 3 Presenting Engagement Opportunities Spatially

In previous work, we argued that geo-visualizations can serve as a catalyst on different levels of citizen engagement and as an integrator for open data (Fechner and Kray 2014). We also reported on techniques to achieve this and resulting citizen behavior in the context of real-time collaboration on maps (Fechner et al. 2015). In this article, we present a user interface (UI) to facilitate exploration and interactivity via a synchronized and integrated geo-visualization approach. The UI intertwines spatial representation of engagement opportunities with their textual descriptions in real-time. This results in the entire exploration process becoming highly responsive, and it allows for different strategies of exploration and to seamless switch between them. Figure 1 shows the interface while the mouse pointer is hovering over a textual description of an engagement opportunity.

The largest portion of the interface is a map-view displaying the locations or affected areas of citizen engagement opportunities. A vertical sidebar shows textual information and images about the available opportunities. Selecting entities in one of them automatically selects or highlights the corresponding entities in the other view. Spatial references within the textual descriptions are directly linked to the map-view and visually emphasized in the text. Non-spatial filtering capabilities such as instantaneous full-text search or filtering through preferences are also available.



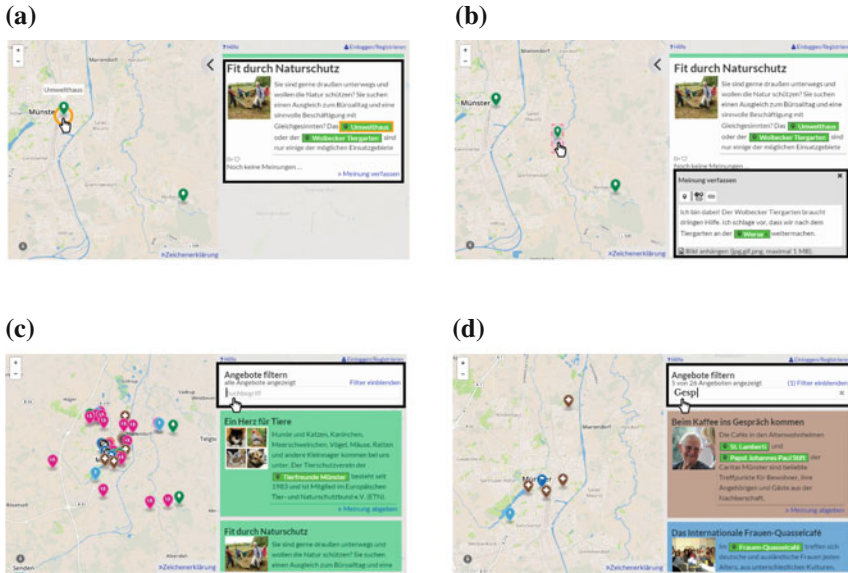
**Fig. 1** The interactive interface consists out of an interactive map-view (a), a sidebar featuring non-spatial filters like a full-text search (b) and textual descriptions with images of the engagement opportunities (c). User interactions in any of the parts automatically update the entire interface

A hovering cursor over a textual description of an engagement opportunity automatically results in highlighting the corresponding locations that are currently within the viewport of the map-view. The spatial references in the textual description are also emphasized. We do not update the spatial extent to encompass all spatial references that are linked to the textual description, though. As mouse hovering over textual description frequently occurs this would produce a lot of disturbance in the map-view. This is one example of the intertwining of textual components and the spatial dimension.

The map view supports standard operations like zooming, panning and hovering over spatial references. Interacting with the map view affects the textual description of the engagement opportunities in the sidebar: Hovering over a spatial reference in the map-view highlights that particular reference and shows its title over the marker. At the same time, the sidebar will display the corresponding textual content or images of the corresponding citizen engagement opportunity. Zooming and panning the map-view do not automatically update the sidebar, for example, to display only descriptions of engagement opportunities that are currently shown in the map-view. This behavior would result in rapid and unexpected changes and could thus irritate users. Refer to Fig. 2 for an illustration of some of the functions.

Clicking either on a spatial reference or a textual description updates the map-view and triggers the expanded, “detailed-view” of that particular engagement opportunity in the sidebar. Expanded content is then displayed on its own in the map-view and sidebar.

Users can now leave comments, ask a question or voice an opinion that relates to the selected engagement opportunity. Comments can be enriched by links to existing spatial reference or uploaded materials. This function is heavily inspired by the



**Fig. 2** Some functions of the geo-visualization: detailed-view, detailed-view with opened feedback function and the textual search that updates displayed content immediately. **a** Detailed-view: shows one particular engagement initiative. **b** Detailed-view: opened feedback function and a referenced location. **c** Textual search interface: displays all content without a search term. **d** Textual search interface: displays matches immediately in the geo-visualization

concept of argumentation maps that was originally introduced by Rinner (1999) and then developed further (e.g. Kessler et al. 2005; Cai and Yu 2009; Sidlar and Rinner 2009). The implementation also enables users to create new engagement opportunities or to create activities for a certain period. However, these features were disabled in this case. The focus of the campaign that we participated in to evaluate the interface was on recruiting volunteers for *existing initiatives* rather than creating new ones. Deployment and use case are described in Sect. 4.

The user interface provides two further functions: an instantaneous full-text search and filters for citizen engagement preferences. Preference filters are collapsed by default and can be expanded on demand. The full-text search enables users to search for an individual or combined search terms. Results are shown immediately after the user typed the first two characters and updated continuously as they type.

Search terms filter the textual descriptions, and sidebar and map-view update to display matches. This creates a highly interactive experience driven by user input. As users type search terms, the search space is reduced, and only matches are displayed both in the sidebar and the map-view. Hence, users can assess search results rapidly from a spatial and textual point of view. Filtering for citizen engagement preferences (e.g. children and youth, supporting the elderly or sustainability) is straightforward and works identically. All functions work in combination, granting users the power to easily and quickly explore the available engagement opportunities.

Colors and icons are consistently used across the entire user interface to provide additional clues and to tie the different UI elements together. Specific colors indicate different preferences of engagement opportunities. The background color of the textual descriptions in the sidebar corresponds to the color of their depictions on the map.

## 4 1000 h for Münster

We cooperated with the “Stiftung Bürger für Münster” (hereafter SBM) in a campaign to raise citizen engagement in the city of Münster, Germany. The SBM is a foundation aiming to cultivate and support citizen engagement backed up by 260 private and corporate donors. The campaign offered the chance to evaluate the geo-visualization in cooperation with established NGOs that address several topics (e.g. sustainability, elderly, children and youth). As Gaventa and Barrett (2010) note citizen engagement does not occur automatically, and active involvement of NGOs is needed.

The two and a half month campaign was called “1000 Stunden für Münster” which translates to “a 1000 h for Münster” and was a joint undertaking of non-governmental partners from the city. Citizens should be made aware of the various citizen engagement opportunities that are available across town and incentivized to partake. Total preparation time of the campaign was a year, it took place from mid-January to end of March in 2015. Two research institutes, a professional graphic artist, a journalist, and 25 NGOs participated in the campaign. It was undertaken as a SBM survey revealed that a portion of the population did not feel sufficiently informed about existing citizen engagement opportunities.

The objective of the campaign was to make citizen aware of the available opportunities and to incentivize them to volunteer time. Although the claim of the campaign was “a 1000 h for Münster”, the cooperation partners knew that reaching this bar would not be likely and hard to measure. Nonetheless, the claim was used as it was catchy. The city in which the study took place has roughly 300000 inhabitants, a high student density, and active civic community with various NGOs.

The use-case allowed gaining new insights, as the targeted user base is diverse and in an actual citizen engagement context. Therefore, the study is not controlled in the sense of lab-based research. It was performed out in the “wild” with actual citizens, granting insights into their preferences.

*Research interest:* The research interested was to *evaluate the relevancy of spatial visualization and interaction for presenting citizens engagement opportunities online*. We wanted to investigate whether citizens would rather rely on the spatial dimension to explore potential engagement cases or if they preferred a categorized and gridded mosaic-view.

*Citizen information portal:* To evaluate the research interest we embedded the geo-visualization in a citizen information portal that we developed for the campaign. The basis of the citizen information portal is a content management system featuring

general information about the campaign and partners. Its primary function was to allow citizens to inform themselves online about the offered engagement opportunities.

Two possibilities were present for that: The first option was the mosaic-view, a visualization form frequently encountered in citizen engagement portals. Such mosaic-views are quick to realize and mimic a content organization form that is usually found in non-digital media, e.g., booklets that present their content structured through categories or some other linear representation. The second option was the geo-visualization, intertwining the spatial dimension of citizen engagement opportunities with their textual descriptions. While both options could be used individually, they were also linked: Users could click on a link in the details page of the mosaic-view opening up the same detailed-view in the geo-visualization, or go back to the mosaic-view via the header of the website.

The mosaic-view displayed 25 engaging images of the offered activities in a grid, see in Fig. 3. Engagement opportunities were clustered linearly by preference in the grid and color coded. Each picture showed the title of the engagement activity below the image. A short description would be displayed on top of an image if a user hovered over the image with the mouse. Detailed-views could be reached via a mouse click, providing descriptions about the initiative, who was organizing it, where it would take place, and when. The geo-visualization was not placed prominently compared to the gridded mosaic-view. The header of the information portal displayed seven entries in this order: A click-able logo to return to the landing page, “1000 h” with general information, “25 Offers” displaying the engagement opportunities in the mosaic-view, “News,” “Map” offering the geo-visualization, “Partners”, and “SBM.”

The mosaic-view would open up if a user clicked on the second tab while the geo-visualization could be accessed via the fourth tab. We suspected that the mosaic-view would often be examined first, due to the ordering of the navigation bar. This assumption is confirmed by the logging system (see Sect. 5). The SBM insisted on this order to prominently display the pictures in the mosaic-view and slightly longer



**Fig. 3** Main view for the mosaic-view and a detailed-view for the mosaic-view. **a** Overview of the gridded mosaic-view displaying all engagement opportunities. **b** Detailed information (what, who, when, where) in the mosaic-view

textual descriptions in the mosaic view. We agreed to this ordering to avoid favoring the geo-visualization.

*Data Logging:* We used an automated logging system to capture user actions. Google Analytics provided a baseline and was complemented by an event-driven custom logging framework to capture user interactions in the browser. With this custom logging, we were able to capture user interactions such as zooming, panning, hovering or typing as Google Analytics does not support these actions.

*Recruiting Participants:* We can only report on the advertisement actions of the campaign and cannot report on age or gender, as we logged the use of the citizen information portal without additional questionnaires, as this study was “in the wild”. Ten thousand glossy a6 booklets with 48 pages each were printed and 9250 distributed in first two weeks. Booklets described engagement opportunities and included some additional material about the partners. Some the booklets (1400) were mailed to partner organizations of the SBM. Aside from the booklets citizen were informed via 25 posters in university buildings, theaters, and exhibition halls. The campaign was featured once in the local newspaper and during a broadcast of the local radio station. Online advertisement included Facebook posts, university-wide newsletters, blog posts and a 50 s YouTube video that was distributed via social networks. All materials included a link to the portal.

## 5 Results

In the following we present the findings of our study, starting with logged data from the citizen information portal (Sect. 5.1). Interactions in the mosaic-view and geo-visualization are looked at in Sect. 5.2, while we close with a user-flow analysis that identifies usage patterns.

### 5.1 Citizen Information Portal

Due to the present ordering of the navigation bar we suspected that the mosaic-view would receive a lot more hits through the navigation bar than the geo-visualization. This assumption holds true: The sub-page that displayed the engagement opportunities in the gridded mosaic-view entitled “25 Offers” was accessed 10.3 (51.3 %) times as much as the entry “Map” (5.0 %). See Table 1 for a compilation of the general data from the website.

We consider 468 out of 713 users as active users, as they did not leave the citizen information portal after glancing at it. The geo-visualization was accessed by 223 active users—roughly every second active user (47.76 %). Considering all users, 227 of the 713 users accessed the geo-visualization (31.83 %), every third user. These access numbers of the geo-visualization are comparatively high, considering the fact



**Table 1** Aggregated data about website usage from January, 15th to March, 31st 2015

Unique users	713 total users; 245 user left after displaying the landing page
User acquisition	47.90 % via direct URL; 43.19 % via referrals from other websites; 8.91 % via social networks or search engines
Visitors	70.9 % new visitors; 29.1 % returning visitors
Geo-ip	97.74 % of visits originated in Germany, 92 % of these visits originated from the federal state in which the city is located
Total pageviews	6627
Number of sessions	998 total; 654 these occurred in the first month
Dropped sessions	26.82 % were dropped after displaying the first page and 7.51 % after displaying two pages
Average session time	03:34 min
Click distribution on navigation entries	Landing page: 17.0 %, 1000h: 13.7 %, 25 Offers: 51.3 %, News: 6.5 %, Map: 5.0 %, Partners: 3.1 %, SBM: 2.9 %, Imprint: 0.6 %

that the entry in the navigation bar was accessed ten times less than the entry for the mosaic-view concerning pageviews. The fact that still 47.76 % of all active users accessed the geo-visualization can be attributed to the link “display on a map”, which was present on each detailed-page of the engagement opportunities in the mosaic-view. Active users who did not access the geo-visualization (245) viewed on average 5.6 sub-pages in the citizen information portal before leaving. The interactions with the mosaic-view were limited as well as those 245 users opened up 2.4 engagement opportunities on average during their stay on the site. Users that accessed the geo-visualization were more active: They viewed on average 11 sub-pages and 3.1 engagement opportunities within the mosaic-view.

## 5.2 Interactions Mosaic-View and Geo-Visualization

The mosaic-view was accessed in 45 % of all sessions with the first interaction, probably due to its placement the navigation bar. The geo-visualization was only accessed in 1.7 % of all sessions with the first interaction, still 47.76 % of all active users accessed it during their stay on the citizen information portal. Users accessed various citizen engagement opportunities in the mosaic-view and looked at the detailed-page, switching back and forth between the overview and the detailed pages. Most views in the mosaic-view were accumulated by engagement opportunities that are already known in the city in the area of children and youth work, international activities that organize meet-ups or environment and sustainability initiatives.

Interactions in the geo-visualization are clearly focused on the map-view. Zooming and panning account for 39.9 % of the actions in the map-view, 57.4 % interactions were mouse hovers, and 2.7 % were clicks on markers to access the detailed-view with the interface for giving feedback. A lot of mouse hovers are not

surprising as this is an interaction that only required users to remain with the mouse cursor over a spatial reference on the map for a few milliseconds. Zooming and panning require a mouse click or scrolling, a more determined interaction in comparison to a mouse hover. The amount of performed actions in each of the areas of the geo-visualizations can serve as indicator to gain a first insight into their importance: 4476 interactions (78.83 %) (zooming, panning, hovering and clicking) occurred in the map-view, while 1141 interactions (20.10 %) occurred in the textual descriptions of the sidebar (hovering over textual spatial references, clicks to access the detailed view, scrolling). The least amount of interactions occurred in the textual filter function that was located on top of the sidebar. In total 61 interactions (1.07 %) were performed with it.

The filter functions were accessed by 14 users; filters were toggled 20 times. Full-text search was used by four users. Each of the users searched for one search term individually, and all occurring searches aimed at one specific and well-known engagement opportunity that helps children to learn to read.

The feedback mechanism that could be accessed in the detailed-view in the geo-visualization was not used. No user wrote a comment, asked a question or referenced an additional area. Seven users clicked the button to open up the feedback form in the detailed-view but did not write or send anything. The feedback form was opened up for six different engagement opportunities.

In total 2527 interactions were recorded for the mosaic-view, while 5678 were recorded for the geo-visualization. While both presentation forms offered interaction possibilities, the geo-visualization offered a wider variety of functions for interaction. Both values are hard to compare directly, still the total interaction counts show that users engaged with the geo-visualization.

### **5.3 User Flow**

For the comparison of the geo-visualization and mosaic-view we look at the user flow—the sequence of accessed sub-pages and interactions on each sub-page per user. If a user accessed the site multiple times, we aggregate the behavior that occurred during multiple sessions and report on the predominant pattern for the particular user. We limit the user flow analysis to active users of the citizen information portal that accessed the geo-visualization. Active users that did not access the geo-visualization amount to 245 users. They only looked at the mosaic-view and did not access the geo-visualization. We do not include them in the user flow analysis as they accessed on average 5.6 sub-pages and on average 2.4 engagement opportunities directly in the mosaic-view. They are still counted as active users as they spent time and did access the mosaic-view, although they interacted overall in a fairly limited fashion with the citizen information portal. Active users that accessed the geo-visualization viewed on average 11 sub-pages and interacted with 3.1 engagement opportunities in the mosaic-view.



**Table 2** User flow categories that were identified in a semi-automatic classification process for users that accessed the geo-visualization

Pattern	Total users
1. Mosaic-view → geo-visualization	114
2. Mosaic-view ↔ geo-visualization	80
3. Only geo-visualization	21
4. Geo-visualization → mosaic-view	8
$\Sigma$	223

We were able to identify four user flow patterns for active users that used the geo-visualization. In a semi-automatic classification process, we sorted and pre-classified all user interactions for all active users that accessed the geo-visualization. Based on the sequence of the interactions with the citizen information portal we grouped and classified the users into four patterns and verified them manually in a subsequent step. Table 2 provides a breakdown for each pattern.

In the *first user flow pattern* with the largest amount of users, a user started to explore the engagement opportunities in the mosaic-view. At some point during the exploration the user decided to either click on the link “display on a map” in the detailed-view or clicked on the “Map” entry in the header and started to use the geo-visualization to explore the engagement opportunities. In this pattern, users did not access the mosaic-view afterward again. A total amount of 114 users employed this pattern.

Switching between mosaic-view and geo-visualization is the *second user flow pattern* we identified with the second largest amount of users (80). Users started to investigate engagement opportunities by accessing the mosaic-view and followed up to use the geo-visualization. Subsequently, they accessed the mosaic-view and geo-visualization again switching between them.

In the *third user flow pattern* users only used the geo-visualization, without accessing the mosaic-view at all. Those users started by clicking on the “Map” entry in the navigation bar and used the geo-visualization exclusively. We counted 21 users that used this pattern.

The *fourth user flow pattern* is the reversed first pattern, but it does not occur often. Eight users started by investigating the geo-visualization first. After interacting with the geo-visualization, they followed up by investigating the mosaic-view. Figure 4 displays patterns one and two for two selected users. Reoccurring interactions are clustered for the mosaic-view and the geo-visualization. Pattern one displays a user that switched from the mosaic-view to the geo-visualization and pattern two depicts a user that switched between mosaic-view and geo-visualization repeatedly.

We computed dwelling times for the total amount of time spent on the geo-visualization and the mosaic-view. Dwelling times for the mosaic-view are slightly higher compared to the geo-visualization. For 117 users (52.5 %) the mosaic-view was displayed longer in the browser compared to the 106 users (47.5 %) where the geo-visualization was active longer in the browser. The computed dwelling times

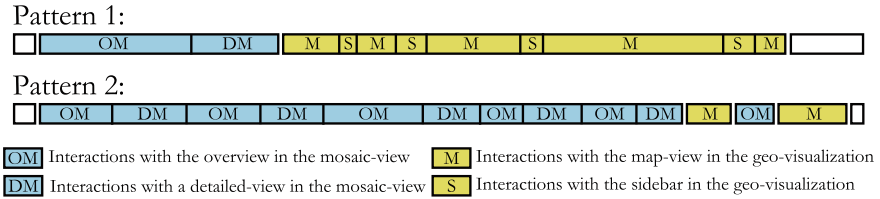


Fig. 4 Example of two user-flow patterns (one and two) that occurred most often. The figure is based on two actual users

for each presentation form account only for the presentation form, time on general information pages (landing page, news, etc.) were not counted and excluded.

### 6 Discussion

Spatial visualization and interaction seem to be relevant for citizens while exploring engagement opportunities online. The first indicator for this conclusion is that a not prominently placed geo-visualization (fourth entry in the navigation menu) was still accessed by roughly every second active user (47.76 %). In terms of raw clicks, the second entry in the menu that lead to the mosaic-view was accessed 10.3 times more than the entry for the geo-visualization. This is probably an effect due to the ordering of the navigation menu. That still a large portion of the user accessed the geo-visualization, although the textual descriptions had all the needed information including addresses, is one indicator that users were interested in spatial visualization of engagement opportunities.

The second indicator is that the total amount of recorded interactions within the geo-visualization is higher than the number of recorded interactions in the mosaic-view. While both presentation forms were interactive, direct comparisons are tricky as different types of interactions were possible. The geo-visualization allowed, for example, to zoom and pan within a map-view to find engagement opportunities. These are interactions that have no counterparts in the mosaic-view.

The third indicator is that a total of 114 users from 223 active users used the geo-visualization without accessing the mosaic-view again after they found it. These users from the first user-flow pattern started with the mosaic-view and did not return to it after discovering the geo-visualization. An additional 21 user interacted only with the geo-visualization, accessing the “Map” entry in the navigation bar directly, without accessing the mosaic-view at all. These users seem to value spatial visualization greatly.

Another indicator that is worth considering is that 80 users repeatedly switched between views. One potential reason for this behavior is that they needed information from both presentation forms. The geo-visualization offered a better spatial overview, allowing users to find engagement opportunities quickly that are close to

each other or located in a particular part of town. The mosaic-view offered slightly longer textually descriptions. This has likely affected and promoted this user behavior and is a limitation of our study. Nonetheless, switching between presentation forms also indicates engagement with both presentation forms. Only eight users started with the geo-visualization and switched to the mosaic-view afterward.

We reported on dwelling times for both presentation forms for active users that used the geo-visualization at least once. Slightly more users (52.5 %) had opened up the mosaic-view longer than the geo-visualizations in the browser. However, dwelling times can not indicate if users favored a particular presentation form, especially in light of tracking limitations and form of information presentation. Reading textual information might take more time than looking at a spatial representation, or a user might need to learn to interact with a spatial representation. The interpretation of the dwelling times for both presentations forms is therefore just another indicator. Despite these known issues, we find that the number of active users staying longer on the geo-visualization is quite high (47.5 %).

Our results show that users did not engage with the spatial discussion features in this particular citizen engagement case. One might argue that this is not surprising, as citizens were not called upon to voice opinions, and the entire campaign was not framed to provide feedback. Still, the total absence was unexpected and might indicate that the implementation of was too obscure or hidden away. Another potential reason is that it takes an active, organized civic community for this kind of engagement (Gaventa and Barrett 2010) and that volunteering time is an individual choice. Apparently citizens were satisfied, in this case, with informing themselves. They did not need to ask questions or exchange information. Krek (2005) found that citizen ignore PPGIS sometimes with rational ignorance—as the cost to learn a PPGIS is too high and the potential gain too small. This idea of rational ignorance might be employed here as well. Another feature that was largely unused were the filter functions. As the map was not crowded, users could easily distinguish between markers by zooming and panning. Supporting this assumption is the fact that users mostly used the map-view (78.83 %) in comparison to the filter functions and sidebar.

## ***6.1 Limitations and Implications***

A number of limitations are present in our evaluation: Our study is not comparative in the sense that we had a within subject design or exposed the geo-visualization or mosaic-view exclusively to specific users only and compared the results. To take part in the campaign and to avoid losing potential volunteers both presentations forms were accessible while the mosaic-view was favored in its placement in the navigation bar. Furthermore, pure data collection through logging actions in the browser is limiting as well. The most obvious limitation is that dwelling times can only be considered very carefully as there is no certainty of what the user did while the website was active. While we used two independent user logging systems to cross-validate our results the data can not be taken at face value. The unique user count of both

systems is identical and session times and dwelling times are similar, but the user count is based on the used device. Multiple users can share a device, distorting the identified patterns, or one user might use different devices at home or work. Ad- and tracking blockers that technology affine users use need consideration as well. Technology affine users that could have particular interaction preferences may have blocked the logging system entirely.

Controlling such effects can only be done in a lab-based environment. We are considering to run a lab-based study with a refined system that also accounts for the discrepancies in the displayed content and functions: The mosaic-view could feature a similar sidebar as the geo-visualization and offer filter functions. There is an amount of users that felt the need to switch between mosaic-views and geo-visualization, and the reasons for this behavior would certainly be interesting.

A couple of limitations arise from the concept itself: Online platforms have the risk to divide the population, as such they can always only be one part of the equation in citizen engagement cases. As Crampton (2009) notes aspects regarding the digital divide and net neutrality are important for the use of maps for any engagement. Unequal access to technological infrastructure and different knowledge levels within the population are likely to lead to an uneven use of such systems. Furthermore, it has to be ensured that websites used for citizen engagement are not treated differently based on the content. Online portals rely on the Internet to disseminate the data of such websites. Net neutrality is crucial to avoid favoring information that expresses views that are more in line with current political and societal agendas.

By using geo-visualizations to expose citizen engagement opportunities citizens have to have a solid understanding of maps. Map literacy is important, as every geo-visualization is an abstraction and simplification. The most popular projection in web applications is a simple spherical Web Mercator projection and unfolding a three dimension object to a two-dimensional plane can not be done without distortions. There are always trade-offs of some kind and citizens need to be aware that distances, shapes, and areas are distorted, and their perception is affected (Wright 2009).

Overall, the attending NGOs were satisfied with the campaign, press coverage, and citizen information portal. They formulated the wish to repeat the campaign in the next years. The SBM's view on the campaign was favorable as well and the 2016 campaign is underway with our continued support.

## 7 Conclusions

We presented and motivated an approach to tightly couple spatial interaction and visualization with textual information about engagement opportunities in this article. Our approach was evaluated based on an extensive real-world deployment with actual citizens. We participated in a campaign to raise citizen engagement in a medium-sized German city partnering with 25 NGOs that offered engagement opportunities. Citizens could inform themselves in an information portal we developed in corporation with the NGOs and a foundation. Two distinct ways to find

engagement opportunities were present: The first was a more traditional gridded mosaic-view that displayed representative images of the opportunities and the second was geo-visualization that exposed and visualized the engagement opportunities spatially.

Based on the presented indicators in Sect. 6 we conclude that spatial visualization and interaction are relevant for citizens who explore citizen engagement opportunities online. The map-view had by far the highest interaction count and seems to be the most relevant part of spatial interactions in the geo-visualization. A good portion of active citizens switched to the geo-visualization from the mosaic-view and continued to use it in the exploration process. Spatial feedback and discussion mechanisms seem unimportant in this particular case and setup. We do not claim that spatial interaction and visualization are more important than the traditional text-based presentation. However, this initial study—based on a large real-world deployment with actual citizens—indicates that citizen did engage with the geo-visualization consistently and in a structured fashion. Further controlled studies are needed to examine if there are users that prefer one presentation form over the other. Still, our results point out that designers of citizen information portals should consider including geo-visualizations to allow citizens to explore engagement opportunities spatially.

**Acknowledgments** We would like to thank the Stiftung Bürger für Münster, all participating NGOs and other partners that helped in the project 1000 Stunden für Münster. We give thanks to Gerald Pape, as the implementation of the information portal was based on his M.Sc. Thesis. The research reported here was partly funded by con terra GmbH.

## References

- Andrienko G, Andrienko N, Jankowski P, Keim D, Kraak MJ, MacEachren A, Wrobel S (2007) Geovisual analytics for spatial decision support: setting the research agenda. *Int J Geogr Inf Sci* 21(8):839–857
- Andrienko G, Andrienko N, Demsar U, Dransch D, Dykes J, Fabrikant SI, Jern M, Kraak MJ, Schumann H, Tominski C (2010) Space, time and visual analytics. *Int J Geogr Inf Sci* 24(10):1577–1600
- Arnstein SR (1969) A ladder of citizen participation. *J Am Inst planners* 35(4):216–224
- Bailey K, Grossardt T (2010) Toward structured public involvement: justice, geography and collaborative geospatial/geovisual decision support systems. *Ann Assoc Am Geogr* 100(1):57–86
- Cai G, Yu B (2009) Spatial annotation technology for public deliberation. *Trans GIS* 13(1):123–146
- Carpini MXD, Cook FL, Jacobs LR (2004) Public deliberation, discursive participation, and citizen engagement: a review of the empirical literature. *Ann Rev Polit Sci* 7(1):315–344
- Connor D (1988) A new ladder of citizen participation. *Nat Civic Rev* 77(3):249–257
- Crampton JW (2009) Cartography: maps 2.0. *Prog Hum Geogr* 33(1):91–100
- Crivellaro C, Comber R, Dade-Robertson M, Bowen SJ, Wright PC, Olivier P (2015) Proceedings of the 33rd annual ACM conference on human factors in computing systems—CHI '15. ACM Press, New York, pp 2853–2862
- Dunn CE (2007) Participatory GIS a people's GIS? *Prog Hum Geogr* 31(5):616–637
- Elwood S (2006a) Beyond cooptation or resistance: urban spatial politics, community organizations, and GIS-based spatial narratives. *Ann Assoc Am Geogr* 96(2):323–341

- Elwood S (2006b) Critical issues in participatory GIS: deconstructions, reconstructions, and new research directions. *Trans GIS* 10(5):693–708
- Elwood S, Leszczynski A (2013) New spatial media, new knowledge politics. *Trans Inst Br Geogr* 38(4):544–559
- Fechner T, Kray C (2014) Geo-referenced open data and augmented interactive geo-visualizations as catalysts for citizen engagement. *eJ eDemocracy Open Gov* 6(1):14–35
- Fechner T, Wilhelm D, Kray C (2015) Ethermap—real-time collaborative map editing. In: Proceedings of the 33rd annual ACM conference on human factors in computing systems—CHI '15. ACM Press, New York, pp 3583–3592
- Ferro E, Caroleo B, Leo M (2013) The role of ICT in smart cities governance. In: Parycek P, Edelmann N (eds) Proceedings of 13th international conference for E-democracy and open government. Donau-Universität Krems, pp 133–145
- Gaventa J, Barrett G (2010) So what difference does it make? Mapping the outcomes of citizen engagement. *IDS Working Pap* 2010(347):01–72
- Hecht BJ, Gergle D (2010) On the “localness” of user-generated content. In: Proceedings of the 2010 ACM conference on Computer supported cooperative work—CSCW '10. ACM Press, New York, pp 229–232
- Hopfer S, MacEachren AM (2007) Leveraging the potential of geospatial annotations for collaboration: a communication theory perspective. *Int J Geogr Inf Sci* 21(8):921–934
- Kessler C, Rinner C, Raubal M (2005) An argumentation map prototype to support decision-making in spatial planning. *Proc AGILE* 5:26–28
- Krek A (2005) Rational ignorance of the citizens in public participatory planning. 10th symposium on Information-and communication technologies (ICT) in urban planning and spatial development and impacts of ICT on physical space, CORP 5
- Maier-Rabler U, Huber S (2011) Open: the changing relation between citizens, public administration, and political authority. *eJ eDemocracy Open Gov* 3(2):182–191
- McCall MK, Dunn CE (2012) Geo-information tools for participatory spatial planning: fulfilling the criteria for good governance? *Geoforum* 43(1):81–94
- Obermeyer N (1998) The evolution of public participation GIS. *Cartography Geogr Inf Syst* 25(2):65–66
- Rinner C (1999) Argumaps for spatial planning. In: Laurini R (ed) Proceedings of teleGeo'99. First international workshop on telegeoprocessing, Lyon, France, pp 95–102
- Rowe G, Frewer LF (2005) A typology of public engagement mechanisms. *Sci Technol Hum values* 30(2):251–290
- Sidlar CL, Rinner C (2009) Utility assessment of a map-based online geo-collaboration tool. *J Environ Manage* 90(6):2020–2026
- Sieber R (2006) Public participation geographic information systems: a literature review and framework. *Ann Assoc Am Geogr* 96(3):491–507
- Steinmann R, Krek A, Blasche T (2005) Can online map-based applications improve citizen participation? In: Böhlen M, Gamper J, Polasek W, Wimmer MA (eds) E-government: towards electronic democracy, vol 3416., Lecture notes in computer science Springer, Heidelberg
- Sui DZ, Goodchild M (2011) The convergence of GIS and social media: challenges for GIScience. *Int J Geogr Inf Sci* 25(11):1737–1748
- Taylor AS, Lindley S, Regan T, Sweeney D, Vlachokyriakos V, Grainger L, Lingel J (2015) Data-in-place: thinking through the relations between data and community. In: Proceedings of the 33rd annual ACM conference on human factors in computing systems—CHI '15. ACM Press, New York, pp 2863–2872
- Weiser M (1991) The computer for the 21st century. *Scientific American*
- Wright DR (2009) Towards fair world maps? A journey in our unfair world. *Int Res Geogr Environ Educ* 18(1):1–4
- Yu H, Robinson DG (2012) The new ambiguity of “open government”. *SSRN Electron J* 59:178–208

# Spatial Data Relations as a Means to Enrich Species Observations from Crowdsourcing

Stefan Wiemann

**Abstract** The general fascination of nature has always been a major driver for studies on living animal and plant species. A large number of professionals and especially volunteers are organized in related initiatives and projects from the local to the global level, leading to the vast amount of species observations nowadays available on the Web. This article seeks to enhance this knowledge base by the determination, management and analysis of feature entity relations among the observations. Those relationships are considered important for comprehensive biological monitoring and, in general, facilitate the integrated use of existing data sources on the Web. Particular emphasis is put on crowdsourcing, which increasingly receives attention and support by citizen science initiatives. The Linked Data paradigm, representing the core of the Semantic Web, is applied to describe, handle and exploit relations in a standardized and thus interoperable manner. Methodologies to determine and validate relationships are developed and implemented. The implementation combines the analysis of spatio-temporal behavioral patterns of species with a crowdsourcing approach for the validation of determined relations. The vagueness of results is addressed by assessing the probability of a relation.

**Keywords** Crowdsourcing • Species observation • Linked data • Spatial data relations

## 1 Introduction

Many people hold fascination for the study of natural history and the observation of nature. From initially being very popular among ancient scholars, the observation of species reached out to a wide public and is meanwhile subject to numerous crowdsourcing campaigns. One of the first and very prominent examples is the

---

S. Wiemann (✉)

Chair of Geoinformatics, Technische Universität Dresden, Dresden, Germany  
e-mail: stefan.wiemann@tu-dresden.de

Christmas bird count in the US, which established a regular annual census of birds since 1900. On the Web, the *Global Biodiversity Information Facility*<sup>1</sup> (GBIF) represents a kind of central repository for species observations worldwide. As of October 2015, the platform provides access to around 580 mio occurrences of circa 1.6 mio different species obtained from 767 different data publishers. It contains archeological, historic and up-to-date observations collected by various research institutes, national data centers as well as crowdsourcing projects. Prominent examples for the latter are *Naturgucker*<sup>2</sup> (~4.3 mio observations), *iNaturalist*<sup>3</sup> (~730.000 observations) and *Artenfinder*<sup>4</sup> (~260.000 observations).

Whereas spatial data on the occurrence of species, such as provided by GBIF, already forms an important biodiversity knowledge base, relations among the observations as well as relations to external data sources on the Web do merely exist. However, such relations are considered a significant extension allowing for more sophisticated data analysis, data enrichment and the inference of spatial information for decision making. In this context, the Linked Data paradigm offers the opportunity to formalize, maintain and analyze spatial data relations on the Web in a flexible and interoperable manner (Wiemann and Bernard 2015). An added incentive is that the combination of geospatial and Semantic Web standards is currently advanced in a collaboration between the *Open Geospatial Consortium* (OGC) and the *World Wide Web Consortium* (W3C), the main standardization bodies in the respective areas (Taylor and Parsons 2015). With respect to existing crowdsourcing campaigns, those developments pave the way for the establishment of an interlinked observations knowledge base as the foundation for comprehensive species analysis.

In OGC (2013), an observation is defined as an “act of measuring or otherwise determining the value of a property”. Although an observation is also considered a feature in the sense of ISO, the term feature hereinafter refers to the observed organism, which is the property described by an observation. Hence, the feature views and feature relations described in this paper refer to the occurrence of a particular species, rather than the observation process.

In the following Sect. 2, different views on a feature and corresponding relations are described. A conceptual workflow for the determination of feature entity relations and possible application scenarios are outlined in Sect. 3. A prototypical implementation for the determination and validation of crowdsourced species observations is described in Sect. 4. The paper concludes with a summary and outlook in Sect. 5.

---

<sup>1</sup><http://www.gbif.org>.

<sup>2</sup><http://www.naturgucker.de>.

<sup>3</sup><https://www.inaturalist.org>.

<sup>4</sup><http://www.artenfinder.rlp.de>.

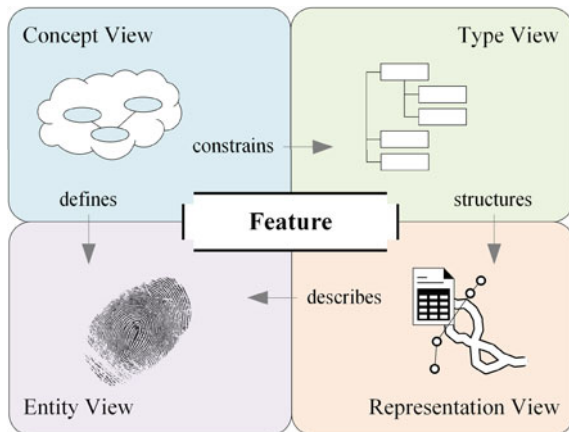


## 2 Differentiation of Relationships Between Species Observations

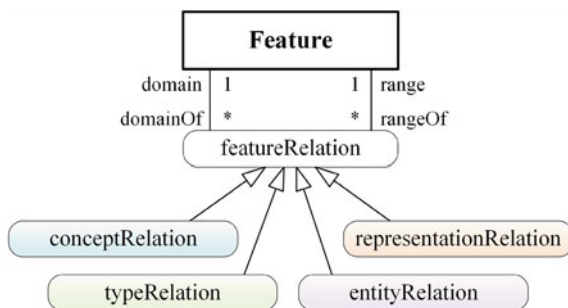
The modelling of geographic information is usually organized and structured on a number of different abstraction levels, from the physical world to the computational representation in software. With respect to the reference model for geographic information standardized by ISO (2014), and with particular focus on possible relations between species observations, four distinct views on a feature are hereinafter distinguished (Fig. 1).

- The concept view describes the general idea of the real-world phenomenon represented by a feature; the feature concept is often represented by an ontology and can be further divided into high-level, domain-specific and application-driven concepts (Guarino 1997). With respect to species observations, this could be the concept of an animal as a living thing.
- The feature type view represents the translation of the feature concept into an application schema and is accordingly constrained by the concept; the feature type defines a number of feature properties that describe a feature. This could be the intention to describe an animal by its name, size and color.
- The feature entity view is a realization of the feature concept and refers to a particular real-world phenomenon; the feature entity is based on the concept of feature identity as a “unique characteristic that distinguishes one object from another” (Hornsby and Egenhofer 2000). Accordingly, every individual animal is considered a separate entity.
- The feature representation view is the computational representation of a feature entity in a spatial dataset; the feature representation implements the feature properties that are defined by the corresponding feature type. With reference to the example, this refers to the description of an individual animal by its name, size and color value.

**Fig. 1** The four different views on a feature and their corresponding relations



**Fig. 2** Specialization of feature relations with reference to the different feature views



The described feature views are not bound to a single feature instance but can be related to an arbitrary number of features. However, while there are usually a lot of features sharing the same concept or type, there are rarely features sharing the same representation.

With respect to the different feature views, two kinds of relations are distinguished: internal and external. Internal relations are used to link the four different views on a single feature to each other and are thus ideally created during the feature modelling process. External relations, hereinafter referred to as feature relations, are used to link two distinct features on the same feature view level; they are accordingly classified into *conceptRelation*, *typeRelation*, *entityRelation* and *representationRelation* (Fig. 2).

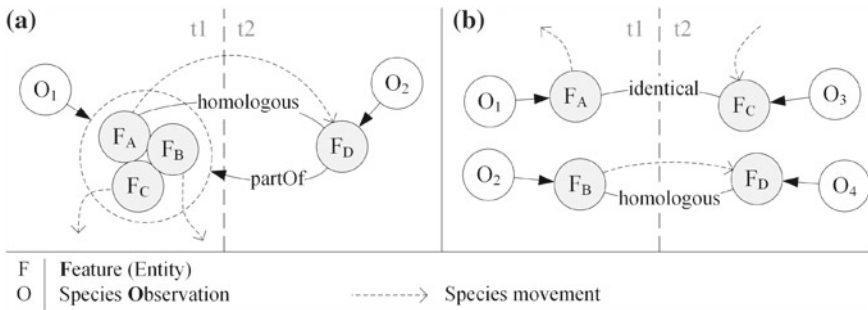
The observation of species by volunteers is touched by a multitude of concepts either in the narrower or the broader sense. Concerning a comprehensible definition, the *HumanObservation* defined by the *Darwin Core Vocabulary*<sup>5</sup> as an “output of a human observation process” can be used. However, if textual definitions are insufficient, e.g. for semantic reasoning, a more comprehensive conceptualization of the observation process, the *Semantic Sensor Network Ontology*, is introduced in Compton et al. (2012). A third example with explicit focus on interoperability is the *Species Distribution Vocabulary for INSPIRE*<sup>6</sup> developed within the *SmartOpenData* project. Relations between those different concepts can be used to bridge semantic dissimilarities and are therefore especially important for cross-domain applications. With regard to Web-based applications, they can best be expressed using the W3C SKOS vocabulary,<sup>7</sup> which recommends a set of Linked Data properties to express hierarchical and associative concept relations.

Type relations define hierarchical (e.g. inheritance) and associative (e.g. aggregate, spatial, temporal or thematic) relations between different feature types. In addition to the qualification of relationships, they may also include schema transformation rules for the mapping of feature properties. Concerning crowdsourced species observations, the majority of projects use their own feature type based on

<sup>5</sup><http://rs.tdwg.org/dwc>.

<sup>6</sup><http://www.w3.org/2015/03/inspire/sd>.

<sup>7</sup><http://www.w3.org/TR/skos-reference>.



**Fig. 3** Feature entity relations between species observations; **a** a parthood relationship and **b** identical and homologous feature entities

the particular data or application needs. GBIF provides global species observation data in a plain CSV structure following the *Darwin Core* vocabulary. A more generic observation schema for citizen science based on the OGC Observations and Measurements standard (OGC 2013) is currently under development, led by the OGC.<sup>8</sup> In addition, the INSPIRE directive defines a data type specification for the obligatory exchange of species information among the EU member states (INSPIRE 2013). However, crowdsourcing projects often introduce custom properties, such as observer profile information, user and expert validation or comments. In consequence, the mapping of crowdsourced observations towards the GBIF repository currently involves a respective loss of information.

Entity relations describe the relationships between real-world phenomena, which are not bound to a particular representation, and are thus hard to quantify. Examples include the mereological (*parthood*) relations described by Varzi (2007) and the temporal transition of objects modelled by Hornsby and Egenhofer (2000). An important distinction is hereinafter made between *identical* feature entities, which refer to real-world phenomena that are equal in value, i.e. belong to the same species, and *homologous* feature entities, which actually refer to the same real-world phenomenon, i.e. represent the same feature entity. Figure 3 exemplarily shows two observations with (a) a *parthood* relation and (b) an *identical* and *homologous* feature entity relation.

Relations between feature representations are primarily based on similarity measurements between location, spatial, thematic, temporal or metadata properties (cf. ISO 2005), which are usually derived from geographic, temporal, network, numeric, string or artificial distance measurements. The qualification of representation relations is accordingly based on the comparison of property values on an ordinal, interval or ratio scale. Examples include the topological relations in Egenhofer and Franzosa (1991), the distance and directional relations in Frank (1992) and the temporal relations in Allen (1983). With respect to species

<sup>8</sup><https://github.com/opengeospatial/swe4citizenscience>.

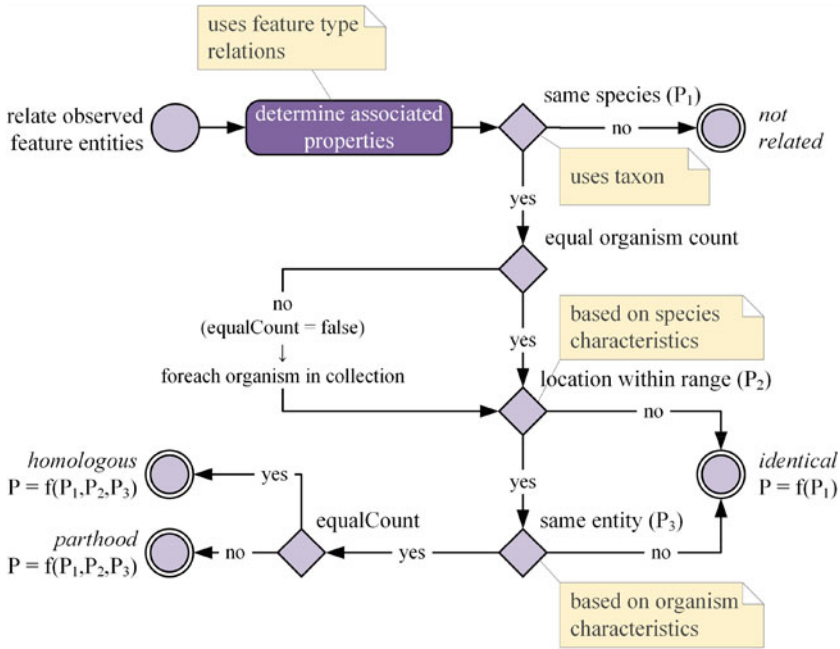
observations, common properties of an observation are: (1) the spatial localization using geometry primitives, usually points, (2) the timestamp or, less often, the time interval and (3) the taxon of the observed organism, which specifies the name and phylogenetic classification of the corresponding species. As mentioned earlier, crowdsourcing induces additional properties, such as information on the observer, the validation status or supplementary material, like photos or comments.

An important aspect for relation modelling, in particular for crowdsourced species observations, is the expression of uncertainty. The uncertainty of a relation is caused by the uncertainty of the observation measurement itself, e.g. determined by the spatial accuracy of the positioning system, and the inherently uncertain events that occurred between two observations. Both can be quantified using confidence measurements, e.g. on a probabilistic basis. A corresponding qualification of a relation can be realized using vague qualifiers, such as geographically ‘near’ (cf. Worboys 2001) or semantically ‘close’ (cf. SKOS *closeMatch*).

### 3 Determination and Application of Species Entity Relations

There are basically three ways to determine a feature relation: (1) the relation is inferred from existing relations, (2) the relation is determined by relation measurements or (3) a combination of the above. An exception to this is the entity relation, which can hardly be measured, and thus needs to be derived from other relations. In practice, an entity relation between two species observations is easier to exclude than to determine exactly. Accordingly, identified relationships are qualified from *impossible*, the negation of a relation, to *certain*, the absolute guarantee of a relation, with intermediate steps, such as *likely*, *even* and *unlikely*. The capability to reliably determine an entity relation significantly depends on the information content of the considered observations and the general characteristics of the observed species. Whereas the first comprises the captured properties of an observation including observer and observation quality metrics, the latter comprises contextual information on the species, such as typical movement patterns, the social and territorial behavior or the estimated population density in the area.

An entity relation between two species observations cannot unambiguously be inferred from concept or type relations. Nevertheless, both can assist the match-making process by imposing limitations on applicable entity relations or relation measurements. For example, features following the *Darwin Core* concept of a *LivingSpecimen* cannot be *homologous* to features following the disjoint concept of *FossilSpecimen*. Yet, they can be related based on the underlying taxon or even be identical in the case of a ‘living fossil’. In a similar manner, feature type relations define associated feature properties that can be used to compare and relate features on the representation level.

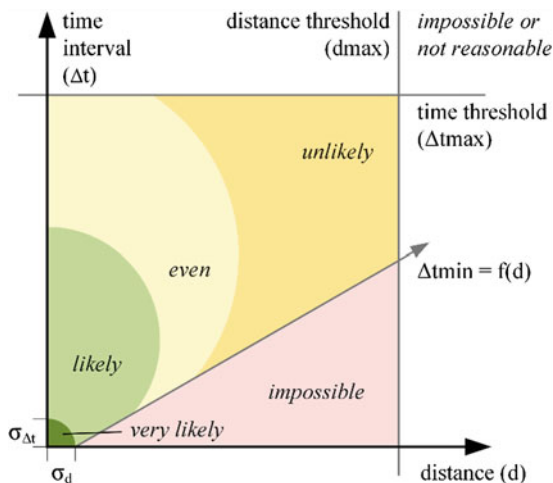


**Fig. 4** Basic workflow for the determination of feature entity relations, in particular parthood, identical and homologous

Figure 4 shows the basic workflow for the identification of an entity relationship between two species observations, revolving around four major decision steps. First of all, two observations need to be observations of the same species, which is evaluated based on the underlying taxon. The attached probability  $P_1$  expresses the probability of both species being equal, including the likeliness of a misclassification if the underlying taxon provides corresponding similarity information on species. In the second step, the count of observed feature entities is compared. If the numbers are neither exact nor approximately equal, the possible feature relation is limited to *identical* or *parthood*. The third and fourth step of the workflow deal with the identification of possibly *homologous* feature entities. In the third step, the two observations are analyzed with respect to known species characteristics in order to assign the probability  $P_2$ , which expresses the estimated probability of the two observations representing the same entity. Since autocorrelation in species movement data is reported to be omnipresent (Cushman 2010), valuable information can be derived from the spatio-temporal distance between observations. Figure 5 shows a corresponding evaluation scheme based on the geographic distance ( $d$ ) and the time interval ( $\Delta t$ ) between two observations. It presumes the following:

- $\Delta t_{min} = f(d)$ : a minimum time interval for each species exists that is required to move a certain distance. Features observed below this threshold cannot be *homologous*.

**Fig. 5** Determination of the likelihood of two observed species observations being homologous



- $d_{\max}$ ,  $\Delta t_{\max}$ : an upper threshold for both spatial and temporal distance exists, above which no relation can be reasonably identified.
- $P = f(d, \Delta t)$ : for each point in the quadrant, there is a function that determines the probability of a relation depending on the spatial and temporal distance.
- $\sigma_d$ ,  $\sigma_{\Delta t}$ : within a defined spatial and temporal uncertainty, two observed entities have a high probability of being homologous.

In consequence, the probability  $P_2$  is significantly influenced by the movement pattern of a species. However, the probability assignment significantly varies with different species and habitat types. As an example, spatial movement can be largely excluded for stationary species, i.e. plants. Moreover, there is an inverse proportional relationship between the probability of two observed features being homologous and the estimated population size of the particular species in the area. Finally, the time of an observation needs to be taken into account with respect to the typical daily routines of the observed species.

In the fourth step of the workflow (cf. Fig. 4), the probability of two observed features being *homologous*, needs to be determined. This decision relies on unique characteristics of the observed organism, such as a tracking marker, an individual coat pattern, visible injuries or a known non-overlapping territory. Although this information is usually very hard to capture and formalize in an observation, crowdsourcing has a great potential, especially because of the widespread use of mobile devices that facilitate the application of multimedia content. By means of this data, the potentially related observations that were identified before can be selected for further substantiation by experts or the crowd, e.g. using a gamification approach similar to the one described and successfully implemented in the Geo-Wiki project (Fritz et al. 2009).

To demonstrate the feasibility of the estimation of feature entity relations, the spatio-temporal distance patterns are exemplarily computed for existing animal tracking data, in particular:

- Four red kites (*Milvus milvus*) tracked between 2012 and 2013 by the red kite project *Rettet die Roten*.<sup>9</sup> In total 5279 observations are analyzed with in average 11 observations per day.
- Seven zebras (*Equus burchellii*) tracked in northern Botswana obtained from the *Movebank*<sup>10</sup> database (Bartlam-Brooks and Harris 2013). It contains about 54.000 GPS fixes for the years 2007–2009.
- 15 grey seals tracked at Sable Island in the years 2011 and 2012 with about 125.000 observations from the *Movebank* database (Lidgard et al. 2015)
- 25 wild baboons (*Papio anubis*) tracked in Kenya for a period of 14 days in 2012, with tracking data obtained from the *Movebank* database (Crofoot et al. 2015)

Figure 6 shows the calculated patterns, which represent the typical daily movement of the observed species entities. As can be seen from the plots, each species has very specific spatio-temporal movement characteristics. Although the variance between different entities of the same species should not be underestimated, the determination of a basic probability of two observed feature entities being homologous seems reasonable.

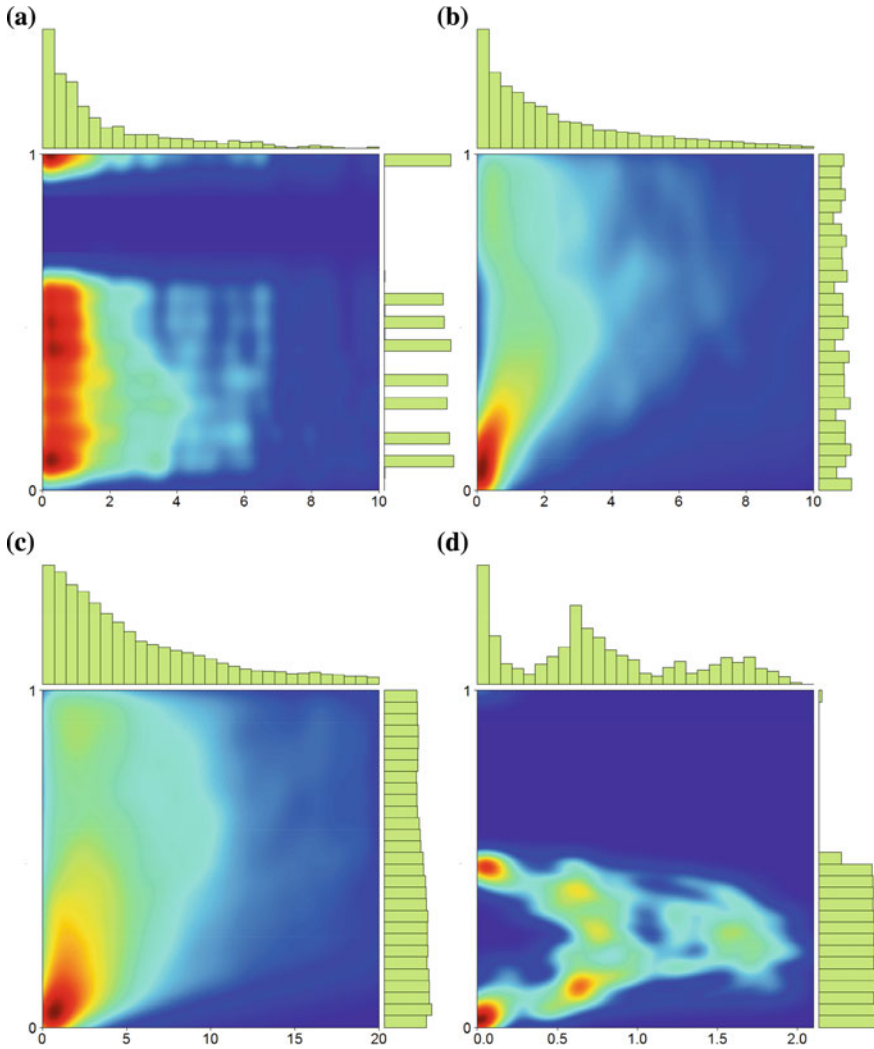
Once identified, the relations can be used to derive a value-added information. This is true not only for the explicit relations, but also for implicit relations, which are induced by transitive, symmetric, equivalent, disjoint and inverse relation characteristics. They allow for the inference of implicit information by the utilization of pre-defined composition tables, e.g. for temporal relations (Allen 1983), distance and directional relations (Frank 1992), topological relations (Egenhofer 1994) or feature class relations (Mäs 2008). The transitive inference of a *homologous* feature relation is exemplarily shown In Fig. 7, whereby the probability of the conclusion  $P_{AC}$  is a function of the probabilities of the premises  $P_{AB}$  and  $P_{BC}$ . However, the logical inference of a relation does not, per se, imply an explicit propagation rule for the attached probabilities (Pfeifer and Kleiter 2009).

As a general rule, *identical* entity relations add information on the observed species, while *homologous* and *parthood* entity relations add information on the particular organism or group of organisms. Since crowdsourcing inherently follows the open-world assumption, which assumes that everything that is not observed is simply not known, the absence of a relation does not entail disjoint feature entities. Thus, if it cannot be deducted from the applied inference rules, the explicit determination of disjoint entities supports subsequent analysis and reasoning. In the following, two applications are outlined that benefit from the exploitation of feature entity relations between species observations: (1) the estimation of a species

---

<sup>9</sup><http://rotmilane.eu>.

<sup>10</sup><https://www.movebank.org>.

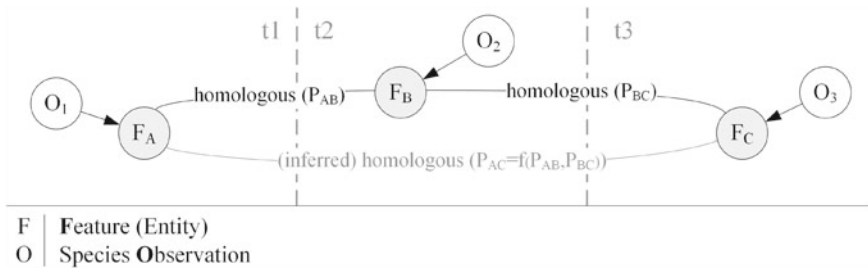


**Fig. 6** Spatio-temporal movement pattern for selected species; x-axis = spatial distance to starting point in kilometer; y-axis = temporal distance to starting point in days; histograms show the spatial (x-axis) and temporal (y-axis) density distributions of the observations; plot colors indicate the density of observations from *red* (many observations) to *dark blue* (no observation). **a** 4 Red Kites, 1 day. **b** 7 Zebras, 1 day. **c** 15 Grey Seals, 1 day. **d** 25 Wild Baboons, 1 day

population distribution and (2) the determination of the behavioral pattern of an observed organism.

Species population counts and density estimations are important biodiversity indicators with significant influence on nature conservation, resource management and policy making (INSPIRE 2013). The choice on an optimal counting procedure depends on the application scope and prior knowledge (Krebs 1999), which



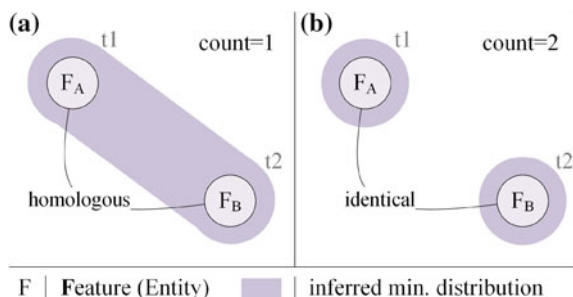


**Fig. 7** Transitive inference of a homologous feature relation with probability propagation

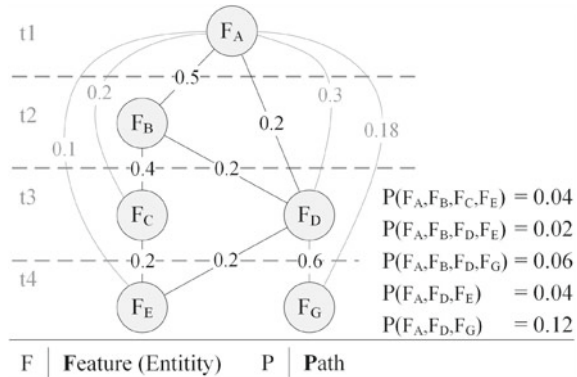
demands a consistent and coherent effort by expert ecologists. Crowdsourcing is different from this in terms of the organizational structure, the average observer expertise, the applied recording methods and the spatio-temporal coverage. Although expert field studies can hardly be replaced, their costs can be lowered by integrating crowdsourced observations, e.g. for the densification, update or enrichment of existing datasets (Wiemann and Bernard 2014). In theory, the total count of organisms belonging to a certain species is the sum of all observations reduced by the count of *homologous* observations. Hence, both the distribution of a species and the actual density of a population are based on the number of *identical* and *homologous* entity relations (Fig. 8). However, as mentioned before, crowdsourced data must be considered as incomplete. Whereas the single occurrence of a species observed by the crowd can be determined relatively well, the density, i.e. count, remains rather uncertain. This also affects information on the absence of a species, which can hardly be verified using solely crowdsourced data.

The identification and study of the behavioral pattern of an observed organism relies on *homologous* feature relations and primarily deals with observed spatial and temporal changes. Whereas a change in location gives an indication of possible movement paths (Fig. 9), a change in the appearance can provide information on different phenological stages of the organism. However, as described in the previous section, the known typical behavioral pattern of a species may support the identification of feature relations in the first place. Thus, to prevent circular reasoning, caution must be exercised when interpreting feature relations with respect to the determination of behavioral patterns.

**Fig. 8** Influence of *homologous* and *identical* feature relations on inferred species count and minimum distribution



**Fig. 9** Exemplary estimation of the movement path of an observed homologous feature entity; probabilities are propagated by multiplication



## 4 A Crowdsourcing Approach to Validate Species Entity Relations

The prototypical implementation developed in the course of this study demonstrates the technical feasibility of the described approach, and provides an example for the validation of feature entity relations using the crowd. For this purpose, observations with referenced multimedia files, i.e. photos, are obtained from GBIF (2015). With regards to the feature views described in Sect. 2, all of the observations follow the concept of a *HumanObservation* as defined by the *Darwin Core* vocabulary. The feature type is a plain attribute-value structure defined by GBIF, which in large parts uses *Darwin Core* terms for the definition of observation properties. The representation of the observed species occurrences include the spatial location given as geographic coordinates, the timestamp, a reference to the *GBIF Backbone Taxonomy* and the associated photos.

The software environment for the identification, maintenance and analysis of relations between the species observations is adapted from previous work described in Wiemann and Bernard (2015) and builds on the following components (Fig. 10):

- A Web-client is used to access existing observations and relations and allows for adding user ratings and comments to both of them. The application is mainly written in *Java* and *JavaScript*.
- The pre-processing analyzes species movement patterns to support the initial matchmaking between observations. Currently, it takes existing animal tracking data and generates spatio-temporal movement patterns (see Sect. 3) using the statistics software *R*.
- The evaluation component processes the user ratings and adds them to an existing relation. The reasoning allows conclusions on the probability of a match, the estimation of species distributions and the inference of possible movement paths (see Chap. 4). The processes are implemented in *Java* and can be invoked by the client.

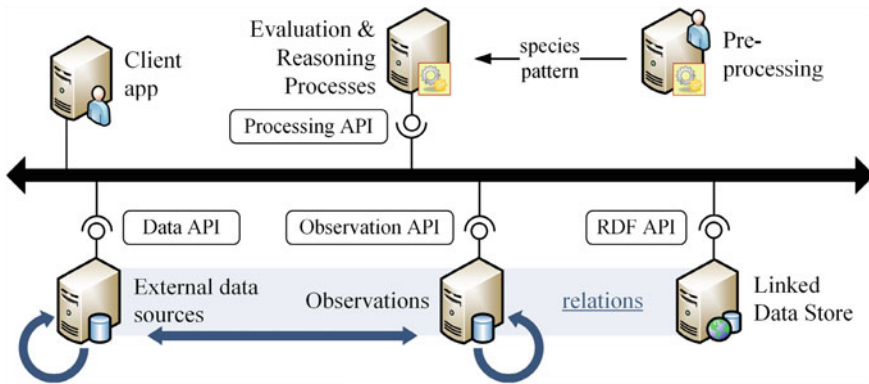


Fig. 10 Component infrastructure and possible relations between the considered data sources



- A Linked Data triple store is used to store and query identified relations using SPARQL (*SPARQL Protocol and RDF Query Language*) and is based on *Apache Jena*.<sup>11</sup> To uniquely identify observations in RDF (*Resource Description Framework*), a reference identifier is created by the combination of GBIF provider and observation id.
- The architecture in principle allows the use of open interface standards for the access to observations. However, since GBIF does not provide any services, the observations are directly read from the downloaded GBIF dataset for testing purposes.
- External data sources can be used to derive observation statistics, e.g. for distribution mapping, or enhance existing data sources with crowdsourced observations (Wiemann and Bernard 2014). The implementation currently supports access to external data via the OGC Web Feature Service (WFS) and the OGC Web Coverage Service (WCS).

Observations of the red kite are chosen as an application showcase, because of the good data availability and the prominent status as an indicator species for biodiversity analysis in Europe. Whereas the observations are assumed to be *likely identical* based on the underlying GBIF taxon, the probability of being *homologous* is first assessed using the identified spatio-temporal pattern in Sect. 3 (Fig. 6a). The corresponding probabilities are set to *very likely* if  $d \leq 100$  m and  $\Delta t \leq 5$  min, to *likely* if  $d \leq 1$  km and  $\Delta t =$  same year and to *even* if  $d \leq 3$  km and  $\Delta t =$  same year. As can be seen, the temporal relation does only play a minor role. This is due to the fixed breeding grounds of the red kite. The movement threshold is set to 40 km/h; below this threshold a relation is set to *impossible*. All remaining relations are classified as *unlikely*. The latter two relations, *impossible* and *unlikely*, are not

<sup>11</sup><https://jena.apache.org>.

explicitly stored in the triple store. The same is true for non-entity relations, i.e. between the concept, type and representation of a feature. Whereas concept and type relations are not required, because of the same origin of data, the representation relations can easily be inferred from the referenced observations

In total 186 red kite observations with referenced media files are available from GBIF (2015). Based on the previously described movement pattern for the determination feature entity relations, no observation pair is classified as *very likely* being *homologous*, 46 pairs are classified as *likely* and 365 pairs are classified as *even*; the rest is classified as *unlikely* or *impossible*. The identified relations are queried by the client for further validation. For this purpose, two possibly *homologous* features are shown to the user with occurrence descriptions, referenced photos, previously identified relations and, if applicable, existing user ratings (Fig. 11). Users can rate and comment on both the observations and the

1st Observation	Comparison	2nd Observation
Source: <a href="#">iNaturalist</a> Location: lat = 50.477; lon = 6.493 Timestamp: 10.05.2015 Taxon: Red Kite ( <i>Milvus milvus</i> )	distance: 1,5km time difference: 1 day	Source: <a href="#">iNaturalist</a> Location: lat = 50.465; lon = 6.504 Timestamp: 11.05.2015 Taxon: Red Kite ( <i>Milvus milvus</i> )
		
<b>Current relations</b> <a href="#">sameSpecies (likely); homologous (even)</a>		
<b>Your Rating</b>		
same species?	<input type="button" value="very likely"/> <input type="button" value="likely"/> <input type="button" value="even"/> <input type="button" value="unlikely"/> <input type="button" value="impossible"/> <input type="button" value="skip decision"/>	
same organism?	<input type="button" value="very likely"/> <input type="button" value="likely"/> <input type="button" value="even"/> <input type="button" value="unlikely"/> <input type="button" value="impossible"/> <input type="button" value="skip decision"/>	
Comments <input type="text"/>		
<input type="button" value="submit"/>		

**Fig. 11** Screenshot of the Web-client application for rating and commenting on the relation between two crowdsourced species observations; the original observations from this example are <https://www.inaturalist.org/observations/1473756> (left) and <https://www.inaturalist.org/observations/1477623> (right)

corresponding relation, or skip the decision. If included in the GBIF dataset, a link to the original observation is added to allow direct responses to the provider, e.g. in the case of a potential misclassification. Upon submitting the form, the user rating and comment are added to the relation via SPARQL Update query. The corresponding RDF encoding of a relation between two observations is depicted in Listing 1.

**Listing 1** RDF Turtle encoding scheme for a *homologous* relation between two species observations with user rating

```

@[base and prefix definitions]
_:relation_[uid]
  a featureRelation ;
  hasReference [1st Observation URI] ;
  hasTarget [2nd Observation URI] ;
  hasRelationType [
    a relationType ;
    type homologous ;
    probability [value]^^[value type]
  ] ;

-----
for each user rating
hasUserRating [
  a userRating ;
  user [user URI] ;
  xsd:dateTime [timestamp] ;
  probability [value]^^[value type] ;
  comment [comment]^^xsd:string
] .
-----

```

Currently, there are too little observations and relations available to perform reliable movement or species distribution analysis. However, queries on the relations and basic validity checks can be performed using SPARQL on top of the provided triple store. As an example, the request shown in Listing 2 returns all potentially *homologous* observations for a defined reference observation with associated estimated and user probability ratings. Moreover, all observations can be accessed by the unique identifier and thus be linked to external data sources following the data fusion approach described in Wiemann and Bernard (2015).

**Listing 2** SPARQL request for all relations that have [*input Observation URI*] as reference, with associated system and user probability ratings

```
[prefix definitions]
SELECT ?observation ?probability ?userRatingValue WHERE {
  ?relation a pre:featureRelation .
  ?relation pre:hasReference [input Observation URI] .
  ?relation pre:hasTarget ?observation .
  ?relation pre:hasRelationType ?relationType .
  ?relationType pre:type pre:homologous .
  ?relationType pre:probability ?probability .
  ?relation pre:hasUserRating ?userRating .
  ?userRating pre:probability ?userRatingValue .
  ?userRating pre:comment ?comment .
}
```

## 5 Conclusion and Outlook

Feature relations between species observations are considered an important means to enhance crowdsourced species observations for environmental analysis and decision making. Whereas the determination and validation of *identical* features is already addressed by most crowdsourcing projects, e.g. by expert ratings, the detection of *homologous* feature entities among the observations remains challenging. The approach described in this paper represents a first attempt to address this issue by the combination of a semi-automatic estimation procedure that relies on characteristic movement patterns of a species and a crowdsourcing approach for the validation of the estimated relations. It enables a standards-based way to determine, formalize and analyze relations between species observations and thus demonstrates the conceptual and technical feasibility. However, there is still a great potential for further developments:

- The estimation procedure for feature relations needs to be improved in order to include more species, in particular species with no trajectory movement data available. Moreover, a review of additional parameters should be carried out to extend the current spatio-temporal approach, e.g. with prevailing environmental or meteorological conditions in the area of the observation.
- The components of the prototypical implementation can be enhanced from both the functional and the esthetical perspective. For example, image enhancement and analysis tools from computer vision could support the determination of relations in the Web-Client. Furthermore, a mobile implementation would facilitate the immediate application in the field.

- A direct connection to existing crowdsourcing platforms on a technical and participatory basis is required to raise awareness and interest in the determination and validation of feature relations. Once linked to an existing application, relations offer additional user feedback mechanisms in the field. For example, the observer can be asked to pay attention to certain characteristics that were observed earlier for a potentially *homologous* feature to facilitate an ad hoc determination and validation of a relationship.
- A general engagement strategy is required to gather a reasonable amount of data for substantial analysis of and reasoning on identified relations. In this context, gamification is often reported to have positive effects on citizen engagement and participation (Morschheuser et al. 2016). In addition, further user studies need to evaluate the general capabilities of laymen and experts to distinguish species and organisms based on the provided observation information, including multimedia files

The approach described in this paper stands and falls with the capability to uniquely identify particular feature entities and distinguish them from another. This depends on many factors and will certainly not work for a large number of species, e.g. a population of small insects. Nevertheless, the determination and validation rate of relations is expected to increase with the continuous enhancement of camera systems, e.g. concerning the geometric resolution or anti-blur filters, the integration of specialized image matching strategies and the general awareness for unique characteristics of organisms belonging to a certain species. Thus, the ‘wisdom of the crowd’ is expected to significantly contribute to existing knowledge and methodologies in the field of biological monitoring.

**Acknowledgments** The work presented in this paper has received funding from the European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 308513, COBWEB.

## References

- Allen JF (1983) Maintaining knowledge about temporal intervals. *Commun ACM* 26:832–843. doi:[10.1145/182.358434](https://doi.org/10.1145/182.358434)
- Bartlam-Brooks HLA, Harris S (2013) Data from: in search of greener pastures: using satellite images to predict the effects of environmental change on zebra migration
- Compton M, Barnaghi P, Bermudez L et al (2012) The SSN ontology of the W3C semantic sensor network incubator group. *J Web Semant* 17:25–32. doi:[10.1016/j.websem.2012.05.003](https://doi.org/10.1016/j.websem.2012.05.003)
- Crofoot MC, Kays RW, Wikelski M (2015) Data from: shared decision-making drives collective movement in wild baboons
- Cushman SA (2010) Animal movement data: GPS telemetry, autocorrelation and the need for path-level analysis. In: Cushman SA, Huettmann F (eds) *Spatial complexity, informatics, and wildlife conservation*. Springer Japan, Tokyo, pp 131–149
- Egenhofer M (1994) Deriving the composition of binary topological relations. *J Vis Lang Comput* 5:133–149. doi:[10.1006/jvlc.1994.1007](https://doi.org/10.1006/jvlc.1994.1007)

- Egenhofer MJ, Franzosa RD (1991) Point-set topological spatial relations. *Int J Geogr Inf Syst* 5:161–174. doi:[10.1080/02693799108927841](https://doi.org/10.1080/02693799108927841)
- Frank AU (1992) Qualitative spatial reasoning about distances and directions in geographic space. *J Vis Lang Comput* 3:343–371. doi:[10.1016/1045-926X\(92\)90007-9](https://doi.org/10.1016/1045-926X(92)90007-9)
- Fritz S, McCallum I, Schill C et al (2009) Geo-Wiki.Org: the use of crowdsourcing to improve global land cover. *Remote Sens* 1:345–354. doi:[10.3390/rs1030345](https://doi.org/10.3390/rs1030345)
- GBIF (2015) GBIF occurrence download. doi:[10.15468/dl.uuel8m](https://doi.org/10.15468/dl.uuel8m)
- Guarino N (1997) Semantic matching: formal ontological distinctions for information organization, extraction, and integration. In: Paziienza MT (ed) *Information extraction: a multidisciplinary approach to an emerging information technology*, lecture notes in computer science, vol 1299. Springer, pp 139–170
- Hornsby K, Egenhofer MJ (2000) Identity-based change: a foundation for spatio-temporal knowledge representation. *Int J Geogr Inf Sci* 14:207–224. doi:[10.1080/136588100240813](https://doi.org/10.1080/136588100240813)
- INSPIRE (2013) INSPIRE data specification for the spatial data theme species distribution. INSPIRE drafting team “Data Specifications”
- ISO (2014) *Geographic information—reference model—Part 1: fundamentals* (ISO 19101-1:2014). International organization for standardization, ISO/TC 211
- ISO (2005) *Geographic information—rules for application schema* (ISO 19109:2005). International organization for standardization, ISO/TC 211
- Krebs CJ (1999) *Ecological methodology*, 2nd edn. Addison Wesley Longman, Menlo Park, California
- Lidgard DC, Bowen WD, Iverson SJ (2015) Data from: a novel approach to quantifying the spatiotemporal behavior of instrumented grey seals used to sample the environment
- Mäs S (2008) Reasoning on spatial relations between entity classes. In: Cova TJ, Miller HJ, Beard K et al (eds) *Geographic information science*. Springer, Berlin, pp 234–248
- Morschheuser B, Hamari J, Koivisto J (2016) Gamification in crowdsourcing: a review. In: *Proceedings of the 49th annual hawaii international conference on system sciences (HICSS)*, Hawaii, USA
- OGC (2013) *OGC Abstract specification: geographic information—observations and measurements*. Open geospatial consortium
- Pfeifer N, Kleiter GD (2009) Framing human inference by coherence based probability logic. *J Appl Log* 7:206–217. doi:[10.1016/j.jal.2007.11.005](https://doi.org/10.1016/j.jal.2007.11.005)
- Taylor K, Parsons E (2015) Where is everywhere: bringing location to the web. *IEEE Internet Comput* 19:83–87. doi:[10.1109/MIC.2015.50](https://doi.org/10.1109/MIC.2015.50)
- Varzi AC (2007) Spatial reasoning and ontology: parts, wholes, and locations. *Handb Spat Logics SE—* 15:945–1038. doi:[10.1007/978-1-4020-5587-4\\_15](https://doi.org/10.1007/978-1-4020-5587-4_15)
- Wiemann S, Bernard L (2014) Linking crowdsourced observations with INSPIRE. In: Huerta J, Schade S, Granell C (eds) *Proceedings of the AGILE’2014 international conference on geographic information science*
- Wiemann S, Bernard L (2016) Spatial data fusion in spatial data infrastructures using linked data. *Int J Geogr Inf Sci* 30(4):613–636. doi:[10.1080/13658816.2015.1084420](https://doi.org/10.1080/13658816.2015.1084420)
- Worboys MF (2001) Nearness relations in environmental space. *Int J Geogr Inf Sci* 15:633–651. doi:[10.1080/13658810110061162](https://doi.org/10.1080/13658810110061162)



# Cross-Linkage Between Mapillary Street Level Photos and OSM Edits

Levente Juhász and Hartwig H. Hochmair

**Abstract** Mapillary is a VGI platform which allows users to contribute crowd-sourced street level photographs from all over the world. Due to unique information that can be extracted from street level photographs but not from aerial or satellite imagery, such as the content of road signs, users of other VGI Web 2.0 applications start to utilize Mapillary for collecting and editing data. This study assesses to which extent OpenStreetMap (OSM) feature edits use Mapillary data, based on tag information of added or edited features and changesets. It analyzes how spatial contribution patterns of individual users vary between OSM and Mapillary. A better understanding of cross-linkage patterns between different VGI platforms is important for data quality assessment, since cross-linkage can lead to better quality control of involved data sources.

**Keywords** Volunteered geographic information • Mapillary • Openstreetmap • User contributions • Cross-linkage

## 1 Introduction

In recent years, technological developments in computer, sensor, and communication technology together with an increase in citizen's interest in sharing spatial information led to a significant growth of crowdsourced geographic information, often referred to as Volunteered Geographic Information (VGI) (Goodchild 2007), which became accessible on Web 2.0 platforms and social media. Contribution patterns for individual VGI applications, such as OpenStreetMap (OSM), photo sharing services, or drone imagery portals, have already been extensively analyzed in the Geoscience literature (Neis and Zielstra 2014; Hochmair and Zielstra 2015;

---

L. Juhász (✉) · H.H. Hochmair  
Geomatics Program, University of Florida, Fort Lauderdale, FL, USA  
e-mail: levente.juhasz@ufl.edu

H.H. Hochmair  
e-mail: hhhochmair@ufl.edu

Hollenstein and Purves 2010). However, it is less understood if and how users participate in several crowdsourcing platforms, whether an individual contributor's activity spaces in different VGI platforms are spatially co-located or spatially distinct, and how data are cross-linked.

Mapillary provides a crowdsourced alternative of street-level photographs to Google Street View. Since its public launch in February 2014 members of this project have so far provided more than 45 million street level photographs along a total of 1,250,000 km of roads and off-road paths. Besides being a crowdsourced and therefore free alternative to Google Street View, Mapillary has also the advantage that its users can take photographs with a smartphone and upload them with an app, without the need of professional camera equipment. This makes Mapillary particularly suitable for image collection on off-road paths, such as hiking or cycling trails. Since street level photographs provide supplemental information to other free alternative data sources, such as aerial photographs, satellite imagery, or census data, they are beginning to be used in other VGI platforms. For example, source tags of OSM edits indicate that Mapillary imagery is already used to edit OSM features. Mapillary image content can be used to identify features or feature attributes that cannot be seen on the aerial imagery but are visible on ground level photos only, such as names of bus stops or buildings. Therefore, Mapillary provides "local" knowledge for OSM remote mappers who do not map in the field. Mapillary imagery is included as a layer option in two of the common OSM editors (iD and JOSM), making it easier for mappers to use street level photos for data editing in OSM.

The overall objective of this study is to determine to which extent Mapillary imagery is currently used for OSM feature editing. More specifically, it aims to determine how the use of street level photos for this purpose varies between different parts of the world, which OSM features are primarily mapped and edited based on Mapillary imagery, how spatially distinct OSM and Mapillary contributions for an individual mapper are, and how cross-linkage to Mapillary is provided in OSM, i.e. which tags are used to reference this connection. This topic is relevant for VGI research since we assume that cross-linkage between different VGI data sources can improve data quality, both by increasing completeness of linked data sources, but also through quality control and review of the original data source, such as the location of an image in the linked platform.

The remainder of this paper is structured as follows: The next section provides a review of related literature, which is followed a description of the study setup. After this, analysis results are presented, which is followed by conclusions and directions for future work.

## 2 Literature Review

Understanding user contribution patterns is important in the context of spatial data quality of VGI (Budhathoki and Haythornthwaite 2013; Coleman et al. 2009). Therefore numerous studies examined data growth patterns for various

crowdsourced geographic data platforms and identified mapper types and mapping communities. Due to the topic of the presented paper the literature review section will focus on community and cross-linkage analysis in previous studies, where OSM is one of the most frequently analyzed VGI platforms. One study analyzed community development in OSM between 12 selected urban areas of the world and found that for cities with lower OSM community member numbers a significant percentage of OSM data contributions (up to 50 %) came from mappers whose main activity area was more than 1000 km away from these particular urban areas (Neis et al. 2013). The interaction between users in OSM was measured for seven selected cities in Europe, the United States, and Australia, by analysis of co-editing patterns in OSM (Mooney and Corcoran 2014). Results showed that high frequency contributors, so called senior mappers, perform large amounts of mapping work on their own but do interact, i.e. edit and update contributions from lower frequency contributors as well. A related study analyzing the OSM editing history for London revealed that there was limited collaboration amongst contributors with a large percentage of objects (35 %) being edited only once or twice (Mooney 2012). An earlier AGILE workshop focused on the activities and interactions which occur during VGI collection, management and dissemination on various VGI platforms (Mooney et al. 2013), including the semantic aspect of the integration of VGI datasets with authoritative spatial datasets. Community analysis among contributors was also conducted for other crowdsourcing and social media platforms. For example, a field experiment on the online encyclopedia Wikipedia showed that informal rewards (e.g. a thumbs-up) increase the incentive to continue contributing only among already highly-productive editors, but lower the retention of less-active contributors (Restivo and van de Rijt 2014). Another study on Wikipedia identified collaboration patterns that are preferable or detrimental for article quality, respectively (Liu and Ram 2011). For example, articles with contribution patterns where all-round editors played a dominant role were often of high quality. Analysis of the network of Twitter users and their followers shows that, although users can connect to people all over the world, the majority of ties from US based users are domestic (Stephens 2015). That is, the Twitter network in the US is spatially constrained and bound by national borders and population density. The connection between the Twitter online social network and the underlying real world geography is also discernable by the fact that 39 % of Twitter ties are shorter than 100 km, i.e. roughly the size of a metropolitan area, and that the number of airline flights emerges as a better predictor of non-local twitter ties than spatial proximity (Takhteyev et al. 2012).

Although contribution patterns and specifically contributor communities within various VGI and social media platforms have been recently discussed, as described above, comparison of contribution patterns of individual users across platforms and data-linkage across platforms has so far not been analyzed in great detail in the literature. Some studies do compare the density and spatial footprints of data contributions between different VGI and social media platforms, such as between Flickr and Twitter (Li et al. 2013), or between several photo sharing services (Antoniou et al. 2010). However, these studies do not analyze individual

contributor behavior or discuss how contributions of one data source are related to contributions from another. Recent trends show that linkage of geographic data across different VGI and social media platforms is a real phenomenon. For example, FourSquare/Swarm users use an OSM background layer to add new check-in places. Therefore OSM positional accuracy directly affects the positional accuracy of FourSquare/Swarm venues. Flickr, a prominent photo-sharing service, has about 30,000 photos tagged with OSM objects. These so-called machine tags potentially allow machine algorithms to automatically extract descriptive information from OSM for Flickr photos. Mapillary uses OSM for reverse geocoding as well. That is, for each photo and photo sequence, the name of the corresponding road is determined by the OSM Nominatim geocoder tool which provides descriptive information of the image locations. In turn, Mapillary photos can be used as a source to derive information for OSM mapping purposes.

It has been shown that OSM positional accuracy is better where high-resolution imagery is available (Haklay 2010). Also, data imports to OSM have clear benefits for areas with a smaller contributor base (Zielstra et al. 2013). Since the Mapillary licensing policy allows OSM contributors to derive information from its imagery, it is a valuable source of geographic information for OSM and other VGI platforms. Within its first year, Mapillary reached significant coverage in some selected cities, and even outperformed Google Street View in terms of completeness in some cases (i.e., in some rural areas and on some off-road segments) as of early 2015 (Juhász and Hochmair 2016). This explains why a growing number of OSM users utilize Mapillary data for OSM data editing, which will be more closely analyzed in the remainder of the paper.

### 3 Study Setup

The study is split into two parts. The first part uses worldwide data and is conducted at the aggregated level to get insight into how OSM contributors cross-tag feature edits and changesets to express the Mapillary source. It analyzes also which OSM primary features (e.g. highways or amenities) are mostly associated with Mapillary, and whether Mapillary information is used to create new features or to edit attributes of existing features. The second part of the study reviews in more detail the mapping behavior of individual users. More specifically it analyses for users that contribute to both platforms to which extent the areas they map in OSM and Mapillary overlap, and whether one of the two VGI platforms is preferred over the other as a mapping platform. This second part of the study is conducted for Europe.

Following the study design, the data retrieval process is also separated into two parts. The first part extracts from an 11 week period all OSM feature editing events and changesets worldwide that are associated with Mapillary according to their tags. The second part identifies, based on user names, users who contributed to both platforms in Europe. For these users, geometries of OSM changesets and edited or added OSM features, as well as Mapillary photo locations are extracted for subsequent comparison.

### ***3.1 Extraction of Mapillary Related OSM Events***

For the extraction of Mapillary related OSM events we used OSM diff files. These files contain all changes made to the OSM database over some time period and can be downloaded at different time granularities from the OSM Website in the compressed OsmChange format. In addition to the daily summary of changes in OSM diffs, we considered also OSM changesets. All these data were extracted between August 31, 2015 and November 15, 2015, covering an 11 week period of OSM edits.

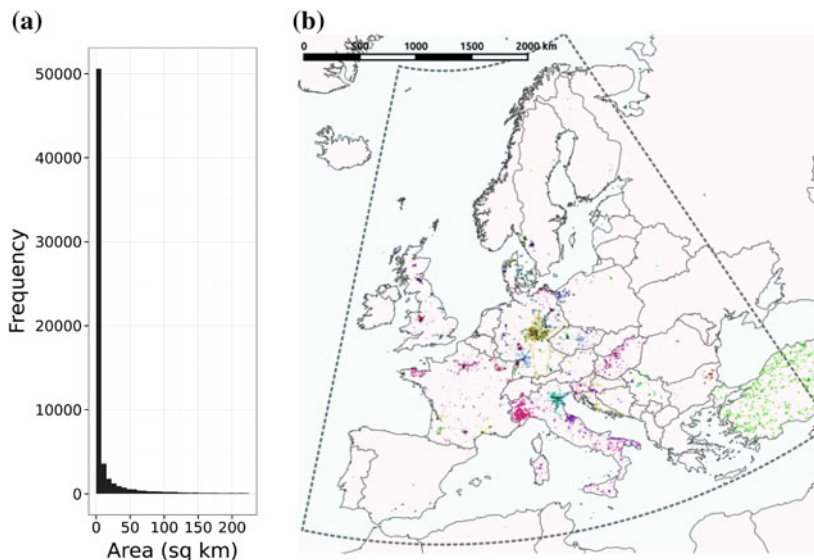
We relate to an OSM event as an insertion or modification of an OSM feature that has an explicit reference to Mapillary. Such references are usually source tags, descriptions, comments or URLs. We consider an OSM event Mapillary related if the expression “Mapillary” (or “mapillary”) can be found in a reference. A software tool, which we developed in Python and Bash, starts with downloading a diff file or changeset. After decompressing the file, the tool converts it into the OPL<sup>1</sup> format with the Osmosis tool. The resulting text files contain OSM edits in rows. Therefore it is possible to search and filter edits with UNIX grep commands. If a line (event) is associated with Mapillary (i.e. “Mapillary” or “mapillary” keywords can be found in tag names or values), it is inserted into a spatially enabled PostgreSQL table. Node and changeset geometries can be reconstructed from the object itself. However, way geometries were extracted from the OverpassAPI. As a result of this process, separate tables for OSM nodes, OSM ways and OSM changesets were available for analysis. Each table contains the unique OSM ID of the object (node, way, or changeset), username and ID of the OSM user that made the edit, tags in hstore format, and timestamps of the event.

### ***3.2 Extraction of Mapillary and OSM Features from Across-Platform Users***

For comparing the spatial editing and contribution behavior between Mapillary and OSM users, users from both platforms were extracted using string matching of usernames. We used a Mapillary database dump of photo sequences that is a suitable representation of the spatial coverage of photo mapping to extract usernames (Juhász and Hochmair 2016). To reduce the chance of extracting two different users who by coincidence share the same user name, only usernames that are longer than 7 characters were considered for this task. Next, it was checked whether the username from the Mapillary database dump exists also in the OSM database using the main API. Since this is not the intended use of the API, we limited our search to 100 matches. Then we reconstructed the OSM editing history of these

---

<sup>1</sup>OPL file manual—<http://docs.osmcode.org/opl-file-format-manual/>.



**Fig. 1** Histogram of filtered changeset areas (a) and selected changesets in Europe colored by user (b)

OSM users using their changesets. A changeset contains the map edits and their bounding area that are submitted by a user to the OSM database, which is typically done on a regular basis to avoid losing completed edits. We limited OSM contributions to the time period after a user signed up to Mapillary, ensuring that both data sources cover the same time range. Since changesets occasionally cover large areas, concealing details about a user's primary regions of edits we excluded changesets larger than 225 km<sup>2</sup>. Using an exploratory approach we found that eliminating the upper tail of the area distribution (Fig. 1a) results in a fairly accurate spatial representation of a user's OSM editing history, for which retained changesets are shown in (Fig. 1b).

To spatially match OSM and Mapillary contributions, a 10 by 10 km grid was created for Europe, limited to the region within the dashed boundaries shown in Fig. 1b. For each cell, OSM and Mapillary edits were extracted for all users that were active in that cell. Results were stored in a PostgreSQL table with unique cell IDs, allowing to spatially match user contributions from both sources. Based upon examination of OSM and Mapillary contributions (areas, descriptions and timestamps) we identified one username which clearly did not refer to the same individual (e.g. editing OSM based on local survey while uploading Mapillary photos from a distant country at the same time). This user was removed from the dataset. The final dataset, after limiting OSM contributions to after the Mapillary signup date and the geographic area to Europe, contained 83 individual users who uploaded photos to Mapillary, edited OSM data, and were most likely the same person.

## 4 Results

### 4.1 Contribution Patterns for Cross-Tagged OSM Features

#### 4.1.1 Cross-Linkage Between OSM Event Types and Mapillary

In a first step it was analyzed how and to which extent the OSM community uses Mapillary as a source of information. The analysis was conducted for tags in Mapillary related OSM events, i.e. node and way edits, and changesets (see Sect. 3.1), that explicitly mentioned “Mapillary” or “mapillary”. For OSM nodes, 1930 events were identified, consisting of new insertions or edits. These events occurred in connection with 1660 unique OSM nodes and were carried out by 68 unique users. For OSM ways, we found 1694 events relating to 1330 unique features that were edited by 96 individuals. Furthermore, the “Mapillary” or “mapillary” keywords appear in 5110 changesets submitted by 209 mappers. The weekly aggregated number of events is shown in Fig. 2. The number of users editing nodes or ways, or submitting changesets (smaller than 225 km<sup>2</sup>) with reference to Mapillary, together with the number unique users per week are summarized in Table 1. The table shows for the different weeks also the total number of OSM users who submitted any changes. Among this group, the percentage of OSM users who submitted changes based on Mapillary images is shown in the last column. Values between approximately 0.5 and 0.6 % indicate that the sub community that uses Mapillary images for OSM data contribution is still a small fraction.

To avoid storing redundant information, OSM users oftentimes attach source information to the changeset rather than to each individual feature. This approach is also recommended when editing multiple features in a mapping campaign. This explains the higher number of committed changesets and the higher number of changeset users with a reference to Mapillary compared to users associated with feature edits. However, it should be noted that not all edits in such tagged changesets are necessarily based on Mapillary alone, although the “Mapillary” or “mapillary” terms appear in the tag. For example, one changeset had a source tag value “bing”, referring to the available Bing imagery, accompanied by several

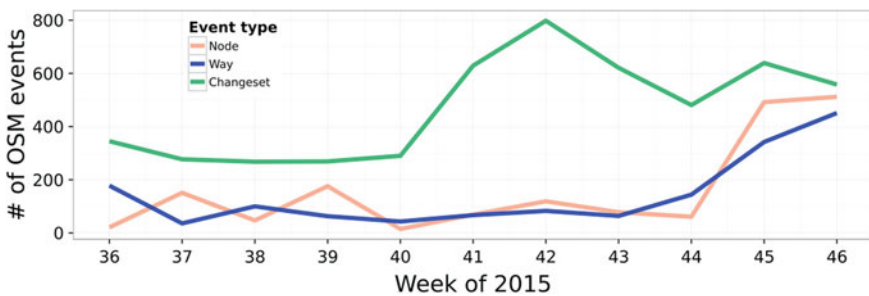


Fig. 2 Number of weekly OSM events cross-tagged to Mapillary



**Table 1** Weekly aggregated number of OSM users

Week	Users associated with Mapillary				All OSM users	% OSM users (Mapillary)
	Node	Way	Changeset	Unique		
Aug 31–Sept 6 (36)	10	12	55	67	11,701	0.63
Sept 7–Sept 13 (37)	12	14	52	63	10,918	0.58
Sept 14–Sept 20 (38)	13	11	47	56	10,476	0.53
Sept 21–Sept 27 (39)	10	14	40	55	10,298	0.53
Sept 28–Oct 4 (40)	9	9	38	49	10,108	0.48
Oct 5–Oct 11 (41)	8	17	56	66	10,606	0.62
Oct 12–Oct 18 (42)	9	13	48	59	10,270	0.57
Oct 19–Oct 25 (43)	18	19	51	68	10,607	0.64
Oct 26–Nov 1 (44)	17	20	52	69	10,872	0.63
Nov 2–Nov 8 (45)	14	13	47	58	11,185	0.52
Nov 9–Nov 15 (46)	7	15	41	54	11,305	0.48

comments, including “*Added crossing from Mapillary and Bing*” or “*Sidewalk + surfaces etc. from Bing, mapillary and local knowledge*”. Analysis of changeset source tags revealed that 29 % of identified changesets” rely solely on Mapillary, local knowledge and surveys, without indicating any other available sources, such as Bing or Mapbox imagery in OSM source tags. We checked also whether Mapillary images overlapped with cross-tagged OSM changesets and found that only 5 % of these changesets were more than 50 m away from the nearest available Mapillary imagery. 84 % of these changesets not located in the proximity of Mapillary imagery were created by the JOSM editor which does not reset the source tag when submitting a new changeset. Therefore these occurrences may be the result of this editor feature, and not of deliberately provided source information by the user. At least one changeset discussion confirms this.<sup>2</sup>

The spatial distribution of events (individual nodes, ways and changesets combined) is shown in the world map in Fig. 3. Table 2 summarizes relative frequencies of event counts by continent together with user numbers. The map shows that in all regions where Mapillary is mostly contributed, i.e. in Europe and the United States, Mapillary is frequently used as a data source for OSM edits as well. This is also confirmed by user numbers in Table 2, which are higher for Europe and the United States (as part of North and Central America) than for other continents. Table 2 reveals that over 61 % of all node edits and over 44 % of all way edits in OSM during the analyzed 11 week period occurred in Asia, which is surprisingly high given that the share of mapped tracks in Mapillary in Asia from all world contributions is only 4 % as of the beginning of 2015 (Juhász and Hochmair 2016). However, the user numbers for node and way edits in Asia are still much lower than those for Europe, which means that this pattern stems from a relative small group of OSM mappers that apply a source tag to edited individual OSM features rather than to changesets.

<sup>2</sup>OSM changeset—[www.openstreetmap.org/changeset/35291204](http://www.openstreetmap.org/changeset/35291204).





**Fig. 3** Spatial distribution of identified OSM events with reference to Mapillary

**Table 2** Identified OSM events with reference to Mapillary by continent

Continent	Nodes		Ways		Changesets	
	Event (%)	Users	Event (%)	Users	Event (%)	Users
Africa	0.05	1	0.00	0	0.08	1
Asia	61.63	11	44.29	16	8.59	19
Europe	37.28	48	50.43	66	45.08	139
North and Central America	0.73	7	4.91	11	43.51	38
Australia and Oceania	0.16	2	0.25	1	0.50	5
South America	0.16	1	0.12	2	2.24	15
Total	100	70	100	96	100	217

Contributions to North and Central America show that only very few mappers tag individually edited features (nodes, ways), but primarily tag changesets. OSM events cross-linked to Mapillary occur in all five continents. It should also be noted that the sum of users over aggregated continent data is greater than the number of users extracted from all nodes or changesets, which implies cross-continent mapping activities.

An analysis of tag distribution for OSM nodes and ways referencing Mapillary shows that the 10 most common tags, including the “source” tag, represent 60.4 % of all tag occurrences. These can be described as power tags (Peters and Stock 2010; Vandecasteele and Devillers 2015), i.e. tags used frequently by many users. The most common tag was “source”, which was attached to 1507 nodes and 1285 ways.

**Table 3** Distribution of primary features cross-linked to Mapillary

	OSM features referencing Mapillary		All OSM features
	#	(%)	(%)
Nodes			
Amenity	736	44.34	5.06
Natural	199	11.99	6.34
Highway	174	10.48	6.20
Tourism	102	6.14	0.81
Barrier	83	5.00	1.45
Leisure	74	4.46	0.31
Public transport	47	2.83	0.74
Ways			
Highway	620	46.62	28.07
Leisure	239	17.97	0.82
Barrier	194	14.59	1.27
Landuse	76	5.71	4.42
Amenity	50	3.76	1.06
Emergency	49	3.68	0.02

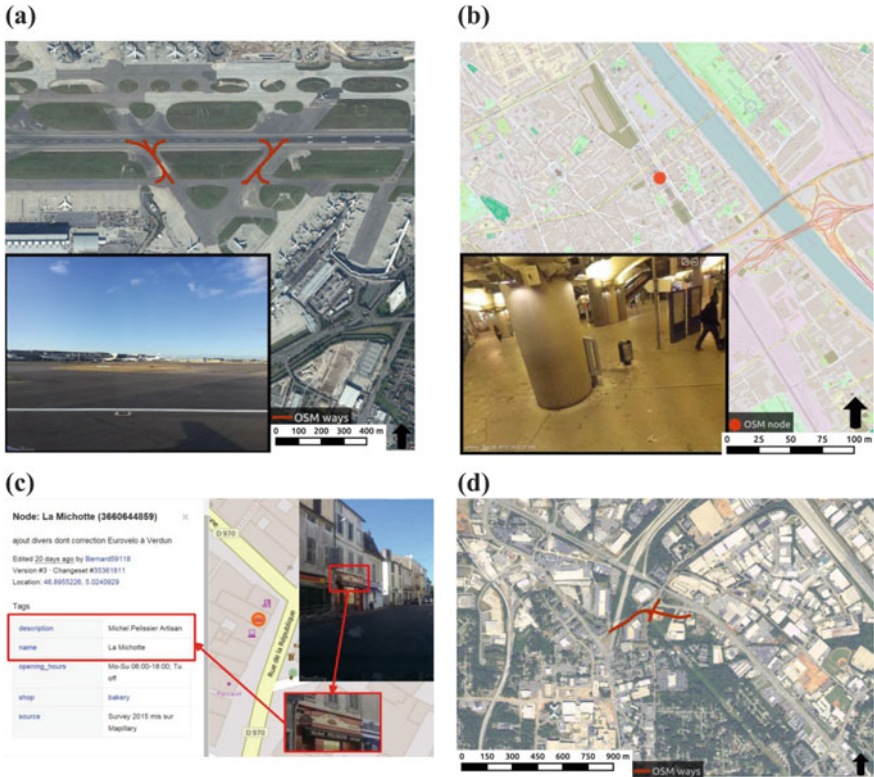
#### 4.1.2 Cross-Linkage for OSM Primary Features

The next step of the analysis examined the distribution of cross-linkages to Mapillary for OSM primary feature categories. For ways and nodes, features from 21 out of the 26 primary feature categories showed a reference to Mapillary in our dataset. Missing primary features are aerialway, boundary, craft, military, and office. Table 3 shows the most frequently used OSM primary features that were cross-linked to Mapillary. The tags of these OSM features show a clearly different frequency distribution than that of the complete set of OSM features, which was extracted from OSM Taginfo.<sup>3</sup> As an example, for node events OSM amenity features are frequently derived from Mapillary (44 %) as opposed to only around 5 % of amenity features that are present in the entire OSM dataset. For way events highway, leisure and barrier OSM features referenced to Mapillary occur at a higher relative frequency than this is the case for the corresponding primary features in the entire OSM dataset.

In addition to primary features, 64 OSM features with a key “traffic\_sign” that are cross-tagged with Mapillary (3.86 % of nodes) were also found. This de facto tag is also related to transportation and often used outside the “highway =\*” tagging scheme. With Mapillary extracting and displaying traffic signs on their website, it is convenient to map traffic signs in OSM.

Surprisingly, some aeroway features, which fall into the category to map air travel related features, appeared in the OSM event list. Although this is outside the

<sup>3</sup>OSM Taginfo—<http://taginfo.openstreetmap.org>.



**Fig. 4** Using street level imagery in OSM: Mapping runway features (a), indoor mapping (b), deriving descriptive information (c), and deriving new road pattern (d)

focus of typical street level imagery, the flexibility of Mapillary allows users to take and upload photos from virtually anywhere. As a result of this, some airport taxiways have been mapped on the London Heathrow airport based on the imagery (Fig. 4a). Another innovative use of Mapillary that can be seen in the analyzed dataset is indoor mapping. Since it is not possible to obtain GPS coordinates inside a building, postprocessing of images allows users to geolocate their imagery and upload it to Mapillary. The presence of an additional “indoor” OSM tag and negative “layer” and “level” values indicate object positioning through Mapillary indoor-imagery (Fig. 4b). In fact, 191 nodes and 161 ways were tagged as indoor or below surface features. For better integration of Mapillary images into the OSM tagging scheme, a new key called “mapillary” has also been introduced to the OSM community, which allows mappers to reference the corresponding Mapillary image in the OSM feature key. A new initiative, OneLevelUp, already renders this information to a web map. Users also tend to use namespaces, indicating from which direction a Mapillary photo shows the object in question (e.g. “mapillary: NE”). In addition, street level imagery provides the ability to capture descriptive

information of features, such as the name of a business (Fig. 4c), the surface type of a road or the material of street furniture. Furthermore, the crowdsourced nature of Mapillary and the ability to capture the rapidly changing world is sometimes a helpful source to obtain an update on geometry information, such as on a modified road layout (Fig. 4d). Interestingly, OSM features highlighted in Fig. 4d do not have a source tag indicating Mapillary, but the following note assigned to them: *“PLEASE DO NOT EDIT if you don’t live here. Roads have been completely reconfigured. High-zoom-level imagery is out-of-date (low zoom level imagery is correct). Consult Mapillary.com sequences for this area to see correct road configuration”*.

### 4.1.3 OSM Activity Types Associated with Mapillary

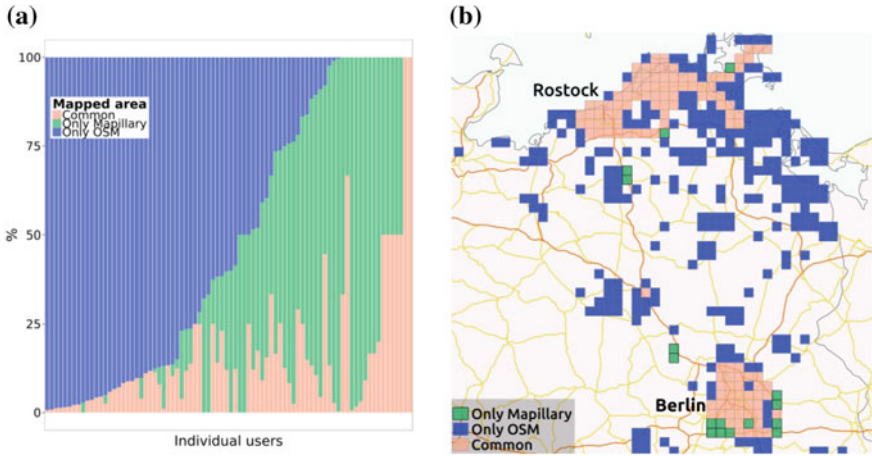
In another step the version numbers for edits of individually edited features in OSM that were cross-linked to Mapillary were extracted. This provides information about whether features were newly created (version number 1) or modified (version number >1). A summary of these activity types is provided in Table 4. The large number of edits with a Mapillary reference (last three columns) suggests that street level imagery is used not only to create new features but also to edit existing ones (e.g. to add descriptive information). The table distinguishes between edits applied to nodes that were created during the 11 week analysis period based on Mapillary (left part), and edits applied to nodes that were created before that period or without reference to Mapillary (right part).

## 4.2 Across-Platform User Contributions

For analysis of individual mapping behavior across the two VGI platforms we extracted Mapillary and OSM contributions of 83 individual users identified earlier as described in Sect. 3.2. This analysis was conducted for Europe (see Fig. 1b). To analyze whether mapped areas of edits are co-located or spatially distinct, for each user, the percentage of 10 by 10 km tiles mapped only in OSM, mapped only in Mapillary, or mapped in both platforms was computed. Results showed that 93 %

**Table 4** Number of OSM features based on activity type

	Created during data collection with Mapillary reference				Edited (created earlier or created without Mapillary reference)
	Total	Not edited further	Edited once	Edited more than once	
Nodes	692	596	65	31	968
Ways	681	593	74	14	649



**Fig. 5** Ratio of mapped areas in Mapillary and in OSM (a) and spatial distribution of a user’s contributions (b)

of users mapped at least some areas in both platforms, resulting in an overlap (Fig. 5a). Even though the sampling of users analyzed for this part of the study started with extracting users from Mapillary, the diagram shows that the majority of users focuses more on OSM (blue area) than on Mapillary (green area) in their data collection efforts. For five users, the exact same tiles are mapped both in OSM and Mapillary. Figure 5b highlights the spatial differences for a selected user in Northeastern Germany, showing that urban areas tend to be mapped both in OSM and Mapillary, whereas rural areas are predominantly mapped in OSM only. The latter may change once urban areas become more completely mapped and thus saturated in Mapillary, so that mappers need to divert more towards rural areas for additional Mapillary contributions. Areas of Mapillary-only contributions can be found along selected major roads (e.g. highway bypass of Berlin). This bypass was already mapped in OSM, but provided a novel contribution option to Mapillary. Mapillary requires users to be physically present at mapping locations, while editing OSM remotely is a common practice. This might be a reason behind OSM contributions being more spatially spread for this user.

Curves in Fig. 6 show which percentage of users mapped at least a given percentage of the total mapped area (constructed from OSM and Mapillary tiles combined) in OSM, Mapillary, or in both. For example, the leftmost values mean that 100 % of the users mapped (at least) 0 % of the area in OSM, Mapillary or both. Moving further to the right, one can see that 75 % of the users mapped OSM in at least 50 % of their combined areas, that 48 % of users mapped Mapillary in at least 50 % of their areas, but that only 10 % of users mapped at least 50 % of overlapping areas, i.e. in OSM and Mapillary. Furthermore the diagram shows that 2 % of users have a 100 % overlap in mapped areas.

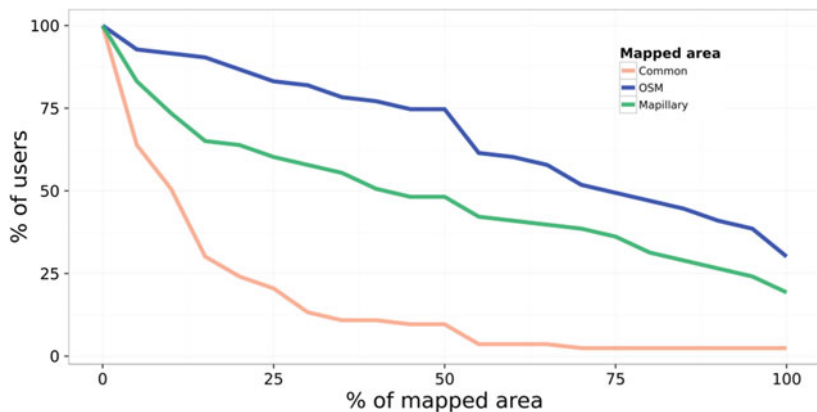


Fig. 6 Distribution of users by mapped area

## 5 Discussion and Conclusions

The first part of the study analyzed how Mapillary street level photographs are incorporated and cross-linked in OSM by matching the Mapillary keyword to tags in OSM edits and changesets. Results showed that even during a short period of time (August 31–November 15, 2015), Mapillary images have been used to edit OSM features. It was found that overall Mapillary is most frequently associated with changesets rather than with individually edited features, although the share of OSM events (nodes, ways, changesets) that are cross-tagged with Mapillary varies between the continents. The predominant tagging of changesets might be the result of batch changes in order to avoid the tagging of redundant information with individual features.

The geographic focus of Mapillary related OSM events corresponds to the core areas of Mapillary contributions, which are Europe and the United States. However, due to some local mapping activities, peaks in Japan and in Southeast Asia could also be identified.

The frequency distribution of cross-linked OSM primary features with a reference to Mapillary is significantly different from that of the entire OSM dataset. The percentage of cross-linked features compared to the entire OSM dataset is higher for transportation (highway, public transport, traffic sign) and leisure (natural, amenity, tourism). This finding is in line with common activities associated with Mapillary, which are recording photos while commuting, traveling, and outdoor and leisure activities, such as hiking. The crowdsourcing nature of Mapillary allows users to map OSM features in places where they are currently less frequently found, including airport taxiways or indoor objects. Cross-linking the two data sources can also help to improve data quality. An example was given where a changed road network pattern was reflected in Mapillary photographs, which were then used to update OSM road geometries. Furthermore the Mapillary images provide a potential data

source for adding OSM feature attribute information (e.g. surface type, name of business) without the need to conduct a field survey.

The second part of the study extracted areas of mapping activities from individual users who contributed both to OSM and Mapillary. The analysis revealed that an individual mapper is more likely to edit larger areas in OSM than in Mapillary. Despite this fact it could be observed that 93 % of users in our sample mapped at least some areas that overlapped between OSM and Mapillary. The overlapping areas tend to be located in locations where a user conducts frequent edits, for example in urban areas the user is familiar with.

For future analysis, we plan to extend our data collection methods to include the geographic areas of API calls from the iD and JOSM editors. These areas will reveal the spatial extent for which OSM users loaded Mapillary imagery into the editors. This will allow us to add a temporal component to the analysis, namely to check whether the viewing of the Mapillary street level photos coincides temporally with an OSM event, e.g. node edit, for an area of interest.

## References

- Antoniou V, Morley J, Haklay M (2010) Web 2.0 geotagged photos: assessing the spatial dimension of the phenomenon. *Geomatica* 64(1):99–110
- Budhathoki NR, Haythornthwaite C (2013) Motivation for open collaboration crowd and community models and the case of OpenStreetMap. *Am Behav Sci* 57(5):548–575
- Coleman DJ, Georgiadou Y, Labonte J (2009) Volunteered geographic information: the nature and motivation of producers. *Int J Spat Data Infrastruct Res* 4(1):332–358
- Goodchild MF (2007) Citizens as voluntary sensors: spatial data infrastructure in the world of web 2.0 (Editorial). *Int J Spat Data Infrastruct Res (IJSDIR)* 2:24–32
- Haklay M (2010) How good is Volunteered Geographical Information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environ Plan* 37(4):682–703
- Hochmair HH, Zielstra D (2015) Analysing user contribution patterns of drone pictures to the dronestagram photo sharing portal. *J Spat Sci* 60(1):79–98
- Hollenstein L, Purves R (2010) Exploring place through user-generated content: using Flickr tags to describe city cores. *J Spat Inf Sci* 1:21–48
- Juhász L, Hochmair HH (2016) User contribution patterns and completeness evaluation of Mapillary, a crowdsourced street level photo service. *Trans GIS*. <http://onlinelibrary.wiley.com/doi/10.1111/tgis.12190/full>
- Li L, Goodchild MF, Xu B (2013) Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geogr Inf Sci* 40(2):61–77
- Liu J, Ram S (2011) Who does what: collaboration patterns in the wikipedia and their impact on article quality. *ACM Manage Inf Syst (TMIS)* 2(2):11
- Mooney P, Corcoran P (2012). *How social is OpenStreetMap*. In: Proceedings of the 15th AGILE International Conference on Geographic Information Science. Avignon, France
- Mooney P, Corcoran P (2014) Analysis of interaction and co-editing patterns amongst OpenStreetMap contributors. *Trans GIS* 18(5):633–659
- Mooney P, Rehr K, Hochmair HH (2013) Action and interaction in volunteered geographic information: a workshop review. *J Location Based Serv* 7(4):291–311
- Neis P, Zielstra D (2014) Recent developments and future trends in volunteered geographic information research: the case of OpenStreetMap. *Future Internet* 6(1):76–106



- Neis P, Zielstra D, Zipf A (2013) Comparison of volunteered geographic information data contributions and community development for selected world regions. *Future Internet* 5 (2):282–300
- Peters I, Stock WG (2010) Power tags in information retrieval. *Libr Hi Tech* 28(1):81–93
- Restivo M, van de Rijdt A (2014) No praise without effort: experimental evidence on how rewards affect Wikipedia's contributor community. *Inf, Commun and Soc* 17(4):451–462
- Stephens M, Poorthuis A (2015). Follow thy neighbor: connecting the social and the spatial networks on Twitter. *Comput, Environ Urban Syst* 53:87–95
- Takhteyev Y, Gruzd A, Wellman B (2012) Geography of Twitter networks. *Soc Netw* 34(1):73–81
- Vandecasteele A, Devillers R (2015) Improving volunteered geographic information quality using a tag recommender system: the case of OpenStreetMap. In: Arjansani JJ, Zipf A, Mooney P, Helbich M (eds) *OpenStreetMap in GIScience (Lecture Notes in Geoinformation and Cartography)*. Springer, Berlin, pp 59–80
- Zielstra D, Hochmair HH, Neis P (2013) Assessing the effect of data imports on the completeness of OpenStreetMap—a United States case study. *Trans GIS* 17(3):315–334



# Geo-Privacy Beyond Coordinates

Grant McKenzie, Krzysztof Janowicz and Dara Seidl

**Abstract** The desire to share one's location with friends and family or to use location information for navigation and recommendations services is often overshadowed by the need to preserve privacy. As recent progress in big data analytics, ambient intelligence, and conflation techniques is met with the economy's growing hunger for data, even formerly negligible digital footprints become revealing of our activities. The majority of established geo-privacy research tries to protect an individual's location by different masking or perturbation techniques or by suppressing and generalizing an individual's characteristics to a degree where she cannot be singled out from a crowd. In this work we demonstrate that location privacy may already be compromised before these techniques take effect. More concretely, we discuss how everyday digital footprints such as timestamps, geosocial check-ins, and short social media messages, e.g., tweets, are indicative of the user's location. We focus particularly on places and highlight how protecting place-based information differs from a purely spatial perspective. The presented research is based on so-called semantic signatures that are mined from millions of geosocial check-ins and enable a probabilistic framework on the level of geographic feature types, here Points Of Interest (POI). While our work is compatible with leading privacy techniques, we take a user-centric perspective and illustrate how privacy-enabled services could guide the users by increasing information entropy.

**Keywords** Privacy · Place · Semantic signature · Location · Geosocial

---

G. McKenzie (✉) · K. Janowicz · D. Seidl  
STKO Lab, Department of Geography, University of California, Santa Barbara, USA  
e-mail: grant.mckenzie@geog.ucsb.edu

K. Janowicz  
e-mail: janowicz@ucsb.edu

D. Seidl  
Department of Geography, San Diego State University, San Diego, USA  
e-mail: dseidl@geog.sdsu.edu

© Springer International Publishing Switzerland 2016  
T. Sarjakoski et al. (eds.), *Geospatial Data in a Changing World*,  
Lecture Notes in Geoinformation and Cartography,  
DOI 10.1007/978-3-319-33783-8\_10

## 1 Introduction and Motivation

While data privacy continues to be an area of worry and confusion for many, recent concerns over the privacy of *location* information specifically have come to the societal forefront. With the increase in mobile devices, as well as technical advances in ambient intelligence powered by the Internet of Things (IoT), location information has become ubiquitous. It has been widely recognized that the resulting technological and social implications will change our understanding of privacy (Bohn et al. 2005; Weber 2010). In fact, personal location information is now arguably a commodity to be traded for services, e.g., for navigation applications, local search, and coupons. Social media have also had a role to play in the advancement of location information usage. An increasing number of social applications allow, and increasingly require, some aspect of location to be shared, be it through posts, messages, check-ins, or photos. While many of these services request location information to improve the user experience, e.g., to show nearby places recommended by friends, other services do not provide clear benefits to the user and collect a variety of personal data in the background (McKenzie and Janowicz 2014). A recent study, for instance, shows that smartphone users are still unaware of the extent and also the frequency at which their personal data are being collected and that they would benefit from more fine grained privacy settings and alerts (Almuhimedi et al. 2015). Even coarse location information can be revealing. In fact, 95 % of individuals can be uniquely identified by just 4 spatio-temporal fixes from cell antennas (de Montjoye et al. 2013).

Consequently, when discussing geo-privacy, people primarily think of geographic coordinates and positioning techniques such as Global Navigation Satellite Systems (GNSS), Wi-Fi-based positioning systems (WPS), Bluetooth Low Energy (BLE) beacons, or radio towers. There are, however, various other possibilities to infer somebody's location and, at least in terms of geo-privacy, some of them may be more revealing than geographic coordinates alone. Additionally, these approaches do not require access to the user's mobile device. This is particularly important as it dramatically increases the number of parties that may infer a user's location. In contrast to positioning techniques, these approaches rely on the notions of *place* and *place types* instead of merely focusing on geographic *space*. Intuitively, there are certain, often latent, place characteristics that emerge from human behavior towards these places and define them as being of a common type, e.g., *bar* or *office*. With respect to temporal characteristics, for instance, a place that is mostly visited during the evenings and weekends is more likely a bar than an office building. Similarly, a place where people predominantly talk about tacos, burritos, and tequila is more likely to be a Mexican restaurant than a Polish restaurant. In an analogy to remote sensing, a set of spatial, temporal, and thematic characteristics that jointly identify a type of place is referred to as the *semantic signature* of said type (Janowicz 2012).

In this work, we employ these signatures to demonstrate how apparently harmless digital footprints such as social media messages, check-in timestamps, and so forth can be used to compromise a user's geo-privacy before position masking techniques come into play. While our work is compatible with established methods for

location privacy, we focus on digital footprints here and how types of places impact geo-privacy. The concern in this case is that people should be aware that even if they don't explicitly share their geographic coordinates that their location can be probabilistically determined based on the words that they write, the timestamps that they make public, and a basic understanding of the spatial and platial<sup>1</sup> configuration of a city.

**The contributions of this work are as follows:**

1. We build on existing work in the area of geo-privacy to show how non-spatial content published by an individual can lead to the disclosure of information directly related to her location.
2. We demonstrate how semantic signatures, built from millions of geosocial footprints, can be used to infer the place type of the location someone is visiting. Moreover, we show that it is possible to quantify this inference and calculate the probability of determining one's location based on her content.
3. We offer a window into what is possible provided seemingly innocuous information. This work suggests ways that content publishers may adjust one or more pieces of published content in order to reduce the risk of revealing their location.

The remainder of the paper is organized as follows. Section 2 introduces related research relevant for the work at hand. Section 3 introduces the datasets used for our study and briefly reviews how the *semantic signatures* were constructed. Three different groups of semantic bands (spatial, temporal and thematic) are discussed in the section following this (Sect. 4). In Sect. 5, we implement our approach through a use case that demonstrates the importance of the semantic signatures in privacy preservation. Finally, we conclude with ideas for future work in Sect. 6.

## 2 Related Work

Geo-privacy research efforts in the GI science community have focused primarily on geomasking or obfuscation techniques, which introduce inaccuracy to geographic coordinates in an effort to balance the protection of location privacy and preservation of spatial information (Armstrong et al. 1999). Attention to the development and evaluation of geomasking procedures has given rise to a large body of work in recent years (Hampton et al. 2010; Zandbergen 2014; Keith C 2015; Kounadi and Leitner 2015; Seidl E. et al. 2015; Seidl 2015; Zhang et al. 2015). The foci of masking studies, which include the testing of distance thresholds and quantification of personal reidentification risk, remain unable to address the impact on location privacy of individuals generating location-bearing content outside a masked data set. A major missing component from these works is the consideration of other data disclosing personal locations even when geographic coordinates are omitted or masked to remain confidential.

---

<sup>1</sup>Following recent literature, we will use the term *platial* here for 'place-based' (Goodchild 2015).

Geo-privacy in masking studies is often defined as the right of the individual to determine how, when, and the extent to which his or her location data is shared with others (Duckham and Kulik 2006). This definition places an emphasis on human agency in privacy rights and is arguably unrealistic in a digital age characterized by frequent and rapid data exchange, where it is difficult to keep track of the parties to which personal data are transmitted. Setting a concrete definition of geo-privacy also opposes other frequently cited conceptual approaches that eschew specific definitions. The definition presented here, however, is in line with the purpose of this paper, which is to introduce unique means by which content publishers, e.g., social media users, may control the release of their location data, namely by considering what is possible with semantic signatures.

The measurement of privacy in a release of data is framed as the risk of identity disclosure. The principle of  $k$ -anonymity describes a release of data where each person in the data set is indistinguishable from  $k - 1$  other individuals in the same data set (Sweeney 2002). The  $k$ -anonymity property does not recognize the side information that an adversary might have about an individual in the database. Another development in information privacy studies is differential privacy, which addresses the problem auxiliary information outside a database poses to the notion of absolute disclosure prevention (Dwork 2011).

Compared to data collected and transferred to third parties in traditional data collection models, individuals do have some agency in the location information they share in user-generated content. The benefits of participation in location-sharing applications (LSAs) or other social networks tend to outweigh perceived privacy risks for users. Social influence is shown to have a strong impact on the adoption of a location sharing application (LSA) among university students (Beldad and Kusumadewi 2015), which extends from having friends or peers known to use the application. Users of the location check-in application Foursquare report that motivations for location sharing include coordination with friends, presentation of self, gaming aspects, and peace of mind or safety purposes (Lindqvist et al. 2011). Location reporting in other social media is not limited to GPS-assisted check-ins, and may be based on text content. Consider the message, “finally home,” which may be posted for peace of mind or coordination purposes. The site “Please Rob Me”<sup>2</sup> used a classifier predicting whether or not a Twitter user was home based on tweets to demonstrate how such information could be exploited by an adversary (Gambis et al. 2010).

Another consideration for this work is whether content publishers are likely to embrace new options for protecting their geo-privacy. A survey of location privacy preferences for personal GPS data finds that providing more complex privacy options, including setting temporal limits and specific locations that may not be shared, leads to more location sharing (Benisch et al. 2011). This provides support for developing an application that allows users to fine-tune privacy settings based on semantic signatures. It also debunks the idea that increased privacy support is at odds with information sharing.

---

<sup>2</sup><http://pleaserobme.com>.

### 3 Data and Semantic Signatures

For the analysis and examples used in this paper we accessed POI data from Foursquare's public facing application programming interface (API).<sup>3</sup> A total of 908,031 randomly selected Foursquare venues<sup>4</sup> were accessed, each categorized into one of 421 Foursquare-defined place types. These types are hierarchically organized into three levels, e.g., Arts and Entertainment > Movie Theater > Indie Movie Theater. Analyzing attributes of these POI and aggregating them to the type level allows us to derive *semantic signatures* (Janowicz 2012). Semantic signatures use digital footprints emitted from humans such as terms that are associated with certain place types, times at which places of a given type are typically frequented, and so forth.

To construct *temporal bands*, each POI in the dataset was accessed every hour for 4 months starting in October 2013. The number of *check-ins* was recorded and cleaned allowing for a *popularity* distribution to be calculated through aggregating data to the place type level. To further strengthen the temporal bands, the 4 months of check-ins were distilled down to hours of the day over the course of a single week. This produced an array of 168 temporal bands (24 h × 7 days). These bands can be further aggregated into coarser resolution bands which are discussed in Sect. 4.2.

*Thematic bands* are constructed from the unstructured textual content provided as *tips* by people that have visited POI. Tips are essentially reviews that a visitor uses to describe or comment on a place. All tips were accessed for each POI in the Foursquare venue dataset mentioned previously. The tips were combined based on place type, stemmed, and cleaned (punctuation and stop words were removed). To ensure robust data signatures, only those place types with 30 or more tips were included in this textual analysis. *Latent Dirichlet allocation (LDA)* (Blei et al. 2003) was used to mine topics from the text and assign probabilistic topic distributions to each of the place types. LDA analyzes documents (aggregate of tips by place types in this case) and extracts topics based on the co-occurrence of words. This allows place types to be described as a distribution of topics extracted from the textual content contributed by individuals to those place types. We call these topic distributions *thematic bands*. In this work, 200 topics (thematic bands) are used.

*Spatial bands* are developed by exploring the geospatial patterns within the POI data. A number of different approaches are used to create these bands. Spatial descriptive statistics such as *Ripley's K* function are used to estimate the deviation of POI place types from spatial homogeneity. In previous work these place type functions have been *binned* by distance and combined with other spatial dispersion techniques such as *Average Nearest Neighbors (ANN)* and *Voronoi place-type variance* to produce a range of spatial bands (McKenzie et al. 2014).

For the purposes of this research, further investigation into the role of semantic signatures in location privacy focuses specifically on examples in the greater Los Angeles region. The boundary of this region was determined through the 2014 U.S.

---

<sup>3</sup><https://developer.foursquare.com/docs/venues/search>.

<sup>4</sup>*Venue* in this case is the Foursquare-specific term for Point of Interest.

*census urban areas* dataset and the boundaries of 240 neighborhoods within this region were ascertained from the 2014 *census designated places* dataset.

## 4 Indicativeness of Digital Footprints

In this section, we present a number of ways that information shared by an individual could be used to expose her location. A multidimensional approach is outlined exploiting the spatial layout of POI, the unique temporal popularity distributions of place types, and the thematic structure that can be extracted from text. The impact of each group of semantic bands is discussed individually and implemented as a whole in Sect. 5.

### 4.1 Spatial Indicativeness

To start with an illustrative example, imagine a user publishing content via her favorite social networking application, stating that she is at a *Mexican restaurant* in neighborhood  $N$ . We assume for the purposes of this research that we have access to a complete POI gazetter for the greater Los Angeles region (e.g., Foursquare venue set).

If  $N$  is *East Los Angeles*, the probability of determining her location is quite low compared to other neighborhoods (Fig. 1a). East Los Angeles has one of the highest ratios of Mexican restaurants to all other POI types in the region, namely 50 out of 809 (0.062). In comparison, the probability of randomly selecting a Mexican restaurant in *Beverly Hills* (Fig. 1b) is merely 4 out of 900 (0.004).



**Fig. 1** Mexican restaurants compared to all POI in two greater Los Angeles neighborhoods. **a** East Los Angeles. **b** Beverly Hills

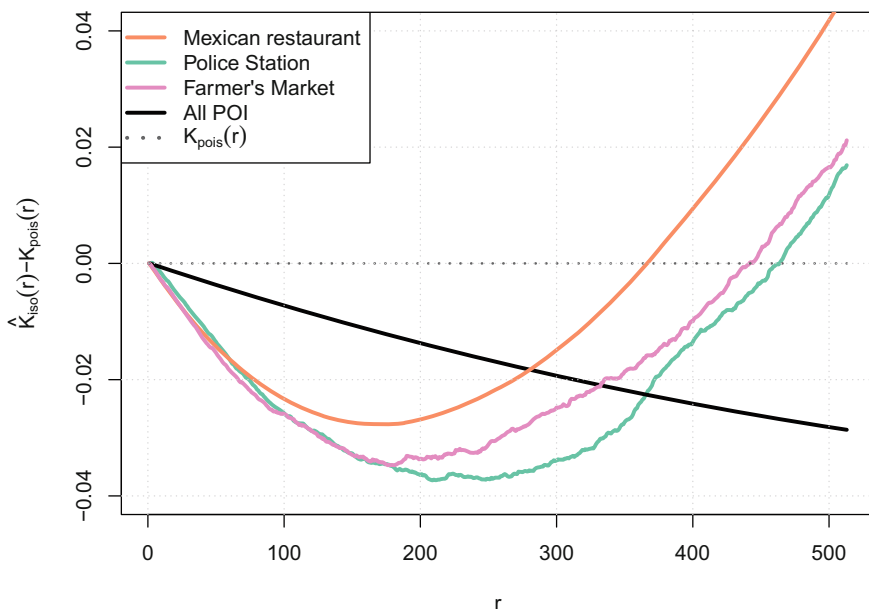
Consequently, knowing that a user is at a Mexican restaurant and in a specific neighborhood significantly impacts the ability to locate this individual. With access to a public POI dataset, the above example shows just how different two neighborhoods are with regards to platial privacy. In other words, the same place type can be revealing in one neighborhood, while it does not expose the user’s likely location in another neighborhood.

If an individual were to state the name of the establishment, e.g., indicate that she were at the chain restaurant *Chipotle Mexican Grill*, this would further increase the probability of determining her exact location within *Beverly Hills*. In this case, two of the four Mexican restaurants in Beverly Hills belong to the chain and therefore have the same name. In comparison, in *East Los Angeles*, no two Mexican restaurants have the same name. Thus, any indication of the place name on the part of the user immediately identifies her location to the place *instance* level.

Given the hierarchy of place types introduced in Sect. 3, we can increase location privacy by simply moving one level up in the place type hierarchy. For example, in the Foursquare place type vocabulary, *Food* is the category into which *Mexican Restaurant* is assigned (along with numerous other restaurant types, grocery stores, etc.). Comparing the number of POI categorized as *Food* to all POI in the dataset, the ability to locate someone in Beverly Hills based purely on place types drops considerably from 4 out of 900 POI (Mexican Restaurant) to 163 out of 900 (Food). Of the 240 neighborhoods in the greater Los Angeles region, Beverly Hills drops from 4th to 193rd with regards to its ability to locate someone based on place type. *East Los Angeles* on the other hand drops to a ratio of 0.234 (189 out of 809). This signifies a substantial decrease in identifiability, but not to the same extent as in Beverly Hills. Table 1 shows a sample of LA neighborhoods along with ratios for

**Table 1** A sample of neighborhoods in Los Angeles showing total POI within each neighborhood along with ratios for four different place types at two different levels in the place type hierarchy

Neighborhood	POI count	Mexican restaurant	Food	Museum	Arts and entertainment
Redondo Beach	948	0.014	0.217	0.000	0.023
Inglewood	998	0.025	0.200	0.000	0.024
Monterey Park	1,085	0.007	0.190	0.001	0.013
Torrance	2,731	0.011	0.168	0.001	0.017
Malibu	1,070	0.006	0.089	0.002	0.026
Santa Monica	1,443	0.016	0.243	0.001	0.038
Culver City	993	0.011	0.209	0.003	0.050
Stevenson Ranch	19	0.000	0.316	0.000	0.000
East Los Angeles	809	0.062	0.234	0.000	0.011
Beverly Hills	900	0.004	0.181	0.002	0.047
All POI	208,682	0.015	0.150	0.001	0.025



**Fig. 2** Plot of Ripley's K functions for three POI categories as well as all POIs in the greater Los Angeles region

*Mexican Restaurants* and *Museums* as well as their parent categories *Food* and *Arts and Entertainment* respectively.

The importance of spatial clustering within the POI dataset must also be considered. Simply knowing a place type and its prevalence within a region is valuable, but knowledge of the spatial distribution of the place type within the region may also lead to an increase in identifying a user's location. For example knowing that an individual is located at a place type that is highly clustered in a region minimizes the time necessary to find them (e.g., search and rescue operation).

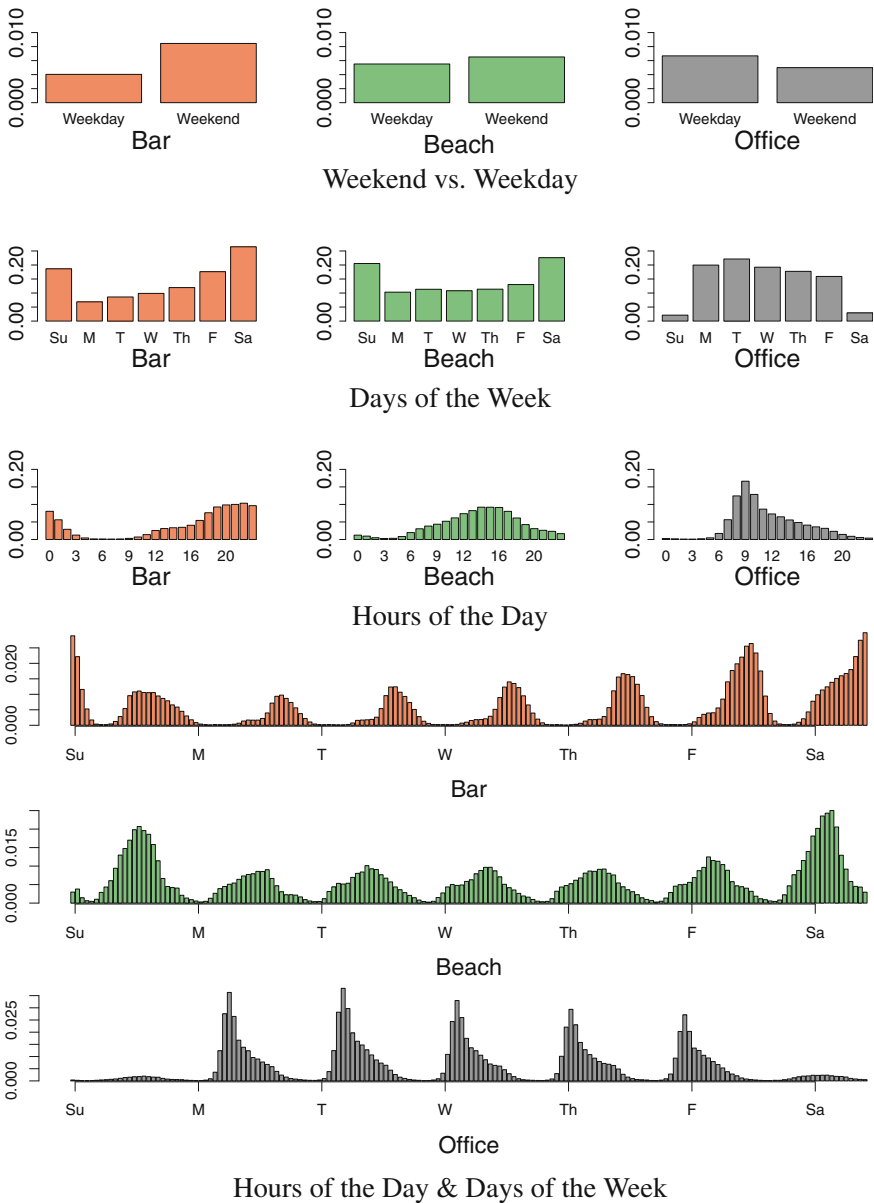
Figure 2 depicts Ripley's K statistics (Dixon 2002) for three place types as well as all places of interest in the Los Angeles. It shows the deviation from spatial homogeneity (shown as the dashed gray line in this figure). Naturally, place types such as Mexican restaurants show stronger clustering at a smaller distance than police stations or farmer's markets. Other methods for assessing the spatial indicativeness of a geospatial dataset have also proved valuable, including spatial entropy (Batty 1974).

## 4.2 Temporal Indicativeness

By way of another example, let us assume that an individual chooses not to publish the place type of the location but rather the *time* at which she is visiting a specific neighborhood  $N$ . Previous research has shown that time is highly indicative of



the types of places that people visit (McKenzie and Janowicz 2015). As one might expect, it is highly unlikely that someone posting from Los Angeles at 5 am on a Monday is at the *Department of Motor Vehicles*. Similarly, one is less likely to locate someone at a nightclub at 9 am on a Monday.



**Fig. 3** Temporal bands aggregated to different granularities and split by three example place types

Using the temporal bands we can probabilistically estimate an individual's location given a specific time. These probabilities can work at multiple levels of granularity. Figure 3 shows temporal signatures for three different place types with increasing levels of temporal granularity. Consulting the values in this Figure, an individual that is very precise in mentioning the time in an online post, e.g., 9 pm on a Friday night, would be more likely to be found at a *bar*, then at an *office building*. These bands can be aggregated based on the level of temporal granularity published. Say an individual solely mentioned the time of day, e.g., 9 am, and not the day of the week, then this method would return *office building* as the most probable place type.

Unsurprisingly, different temporal bands offer different amounts of information about the platial location of an individual. For instance, someone who only mentions 5 am on a Monday when publishing content is unlikely to be at *Department of Motor Vehicles*. Realistically, the probability of this person being anywhere except at home is rather small. On the other hand, if this person were to mention 6 pm on a Friday there is a much wider range of places this person could be given the activities that are possible at this time. To put it more formally, each temporal band can be defined by the unpredictability of the place types one might visit, which can be represented through *Information Entropy* (Claude E 1948). 5 am on a Monday has relatively low information entropy when compared to 6 pm on a Friday, given that one could more easily predict the place type of an individual in the first case, namely in some form of accommodation. Information entropy ( $E_T$ ) is defined in Eq. 1 where  $p_i$  is the probability of a given temporal band.

$$E_T = - \sum_i p_i \log_2(p_i) \quad (1)$$

Previous work (McKenzie et al. 2014) explored the amount by which the hourly temporal bands are unpredictable. Computing entropy across check-ins to all POI in the dataset showed that there is a statistical difference in the information that is presented between the hourly temporal bands (Table 2). This is important as the ability to determine the place where someone is can drastically increase depending on the time that she publishes content.

**Table 2** Information entropy for five lowest and five highest temporal bands

Low entropy			High entropy		
Day	Hour (AM)	Entropy	Day	Hour (PM)	Entropy
Monday	05:00:00	4.76	Thursday	07:00:00	5.97
Monday	04:00:00	4.87	Tuesday	07:00:00	5.96
Tuesday	04:00:00	4.93	Friday	06:00:00	5.95
Thursday	04:00:00	4.95	Friday	07:00:00	5.94
Tuesday	03:00:00	4.99	Saturday	12:00:00	5.93

### 4.3 Thematic Indicativeness

The words and language that people use when talking about the activities are indicative of the type of place they are doing the activity. Previous work in this area has shown that non-geographic terms and phrases can be geospatially indicative (Adams and Janowicz 2012; Mahmud et al. 2014). The results show that words in the English language can be tied to some region on the planet with varying levels of probability.

The thematic bands introduced in Sect. 3 define each place type in the Foursquare dataset as a distribution across topics. In short, the place types are defined by the language of the people that have visited them. Three examples of topics extracted from the unstructured natural language of the Foursquare tips are shown in Fig. 4 as word clouds of the topic’s most prevalent terms.

Using these thematic bands as the foundation, we use an LDA inference approach (McCallum 2002) to infer a distribution of these same topics for any new unstructured text-based document. For example, given content such as,

So glad I made it in to deposit my check at the ATM before they closed.

We, as humans, likely infer that the user is at a bank. From a computational perspective, an LDA model would need to construct a topic distribution for this text that would likely place a high probability on the topic related to banking (Fig. 4b), low probability on the topic related to Mexican food (Fig. 4a) and somewhere in the middle for the non-place type topic (Fig. 4c). It is also likely that the *bank* place type follows a very similar topic distribution to the topic distribution of the sentence above. *Jensen-Shannon distance (JSd)* (Lin 1991) (Eq. 2) is used to measure the dissimilarity between our newly created topic distribution ( $P$ ) and each of the topic distributions for all 421 place types ( $Q$ ). *KLD* (Eq. 3) represents the *Kullback–Leibler divergence* and the lowercase  $d$  in *JSd* signifies *Distance* instead of *Divergence*.  $M$  is equal to  $\frac{1}{2}(P + Q)$ . The smaller the dissimilarity value (bounded between 0 and 1), the more likely it is that our example content can be assigned to that place type. In this simplified example, the sentence above shows the least dissimilarity with the *bank* place type, and thus the user is said to be most likely at a bank. An implementation of this model is discussed in further detail in Sect. 5.

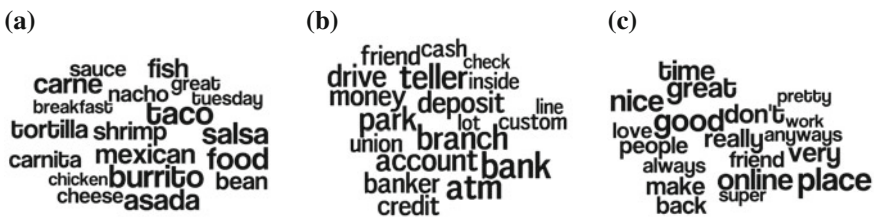


Fig. 4 Three example topics represented as word clouds of their most prevalent terms. **a** Terms related to Mexican food. **b** Banking related terms. **c** Non-place type specific terms

$$JSd(P \parallel Q) = \sqrt{\frac{1}{2}KLD(P \parallel M) + \frac{1}{2}KLD(Q \parallel M)} \quad (2)$$

$$KLD(P \parallel Q) = \sum_i P(i) \log_2 \frac{P(i)}{Q(i)} \quad (3)$$

## 5 Implementation: A Use Case

In the previous sections, we discussed the various bands of semantic signatures and the ways in which these bands contribute to determining the place where someone is. In this section, we bring the bands of the semantic signatures together to implement one approach that determines a user's place. An example use case is introduced, and the parameters are altered to show how sensitive the model is to changes. A first implementation of a formula is introduced to quantify the place-based privacy implications of the content.

### 5.1 Thematic Content

To start, let us imagine that an unknown individual publishes some small amount of unstructured content, e.g., a tweet. In this first iteration of the example, the content is both thematic and spatial but does not include any temporal property.

Excited for chicken tacos and delicious salsa in Beverly Hills. (1)

After stemming, a topic distribution for the text is inferred through an LDA topic inferencer based on the topic distributions (200 topics) learned from the 421 place types (thematic bands). A *JSd* dissimilarity value is then computed between the topic distribution for this text and each of the place type topic distributions. Note that this example uses a very small amount of text, so the inference model has a limited amount of data on which to infer the topic distribution. A greater amount of data would arguably lead to more accurate results. The top 10 least dissimilar place types are shown in Table 3.

The place types listed vary in their specificity. *Taco place* is a sub type of *Mexican restaurant* while *building* is a very generic place type. To put it another way, the descriptive content contributed as tips about taco places are narrower in their theme than the building place type which might include a wide range of themes related to places that exist within a building, e.g., restaurant types or car mechanics. Equation 4 shows how the thematic property of a place type ( $PT_{Theme}$ ) is quantified. Note that this function simply converts the dissimilarity value into a similarity value (higher value = better match).

$$PT_{Theme} = 1 - PT_{JSd} \quad (4)$$

**Table 3** Top 10 place types that are least dissimilar from the sample content (Quote 1)

Place type	JSd dissimilarity value
Mexican restaurant	0.267
Taco place	0.268
Food	0.301
Bar	0.302
Restaurant	0.309
American restaurant	0.317
Building	0.321
Miscellaneous shop	0.321
College cafeteria	0.329
Food and drink shop	0.330

### 5.2 Spatial Constraints

From a regional or *spatial* perspective, the content in Quote 1 indicates that the publisher is in *Beverly Hills*. We know from our gazetteer of places that there are four *Mexican restaurants* within the neighborhood boundary. Making the assumption that there is a certain region around an individual’s point location that they can sense (e.g., visually, auditory), we construct a grid over a region. We expect that one would be able to locate something or someone reasonably quickly within this region. Provided this assumption, we overlay a 500 × 500 meter cell grid over the Beverly Hills neighborhood in Los Angeles. Recording the presence or lack thereof of POI in each grid cell we find 115 out of 118 grid cells contain at least one POI. Of these, 2 grid cells contain at least one Mexican restaurant producing a ratio of 2/115 or 0.017.

Through these two data dimensions we are able to first determine the place type of the user and building off this constraint, spatially restrict the location possibilities. Using a rudimentary cell-based clustering technique we can further restrict the expected spatial locations of a content publisher.

### 5.3 Spatial Change

Building on the content of Quote 1, let us imagine that instead of sharing *Beverly Hills* as her location, this person mentions *East Los Angeles*. The textual content remains the same, so we have still determined that *Mexican restaurant* is the probable place type, but in this case, the number and spatial layout of place instances matching this criteria has changed. Overlaying the same 500 × 500 meter cell grid over *East Los Angeles* we find that 112 out of 136 cells contain at least one POI and of these cells, 36 contain at least one Mexican restaurant resulting in a ratio of 0.321. So while the place type remains the same, the difference in spatial layout of these two

**Table 4** Effort values for two neighborhoods, Beverly Hills and East Los Angeles

Neighborhood	Mexican restaurant cells	Ratio	Total cells	Effort value ( $\times 10^4$ )
Beverly Hills	2	0.017	136	2.5
East Los Angeles	36	0.321	118	979.3
Greater Los Angeles region (Full area)	2,328	0.088	98,461	20.8

The Greater Los Angeles region is shown for comparison

neighborhoods means that there is a substantially lower chance of someone locating the user in East Los Angeles compared to Beverly Hills.

While the ratio is informative, the raw cell count is important here as well. Tasked with finding the publisher of the content a user would have to travel to 36 different regions (cells) in East Los Angeles but only 2 in Beverly Hills. Stepping back to the entire greater Los Angeles region, there are 98,461 cells that overlap neighborhood boundaries, and of these, 26,311 contain POI. Of the cells containing at least one POI, 2,328 contain at least one Mexican restaurant, producing a ratio of 0.088. Taking this ratio by itself implies that on average it is harder to locate someone at a Mexican restaurant in East Los Angeles than in the greater Los Angeles area overall. Though in this case, one would have to travel to 2,328 different regions (cells) in order to find the content publisher.

A relative *effort* value bounded between 0 and 1 is proposed by multiplying the number of likely cells by the ratio and dividing by the total possible set of cells over the regions. Table 4 lists the resulting effort values for the neighborhoods previously discussed.

## 5.4 Content Change

Again, let us slightly alter the published content and observe the implications on location privacy. Keep in mind that the actual location of the user (Beverly Hills) and activity (eating Mexican appetizers) remains the same. If instead of posting about the specific type of appetizer, the user generalizes her content as shown in Quote 2, what impact does this have on our ability to locate her?

Excited for great chicken appetizers in Beverly Hills. (2)

A topic distribution for this new content is again inferred from the existing LDA topic model and JSd is used to calculate the dissimilarity between this topic distribution and all place type topic distributions. The top ten least dissimilar place types are shown in Table 5.

Importantly, *Mexican restaurant*, presumably the place type the user is currently enjoying their food, appears nowhere in the list. The best match is instead, *food*, which is the parent category of *Mexican restaurant*, as well as many other place

**Table 5** Top 10 place types that are least dissimilar from the sample content (Quote 2)

Place type	JSd dissimilarity value
Food	0.263
Restaurant	0.268
American restaurant	0.275
Miscellaneous shop	0.276
Cafeteria	0.287
Cafe	0.305
Building	0.310
Assisted living	0.312
College cafeteria	0.313
General entertainment	0.322

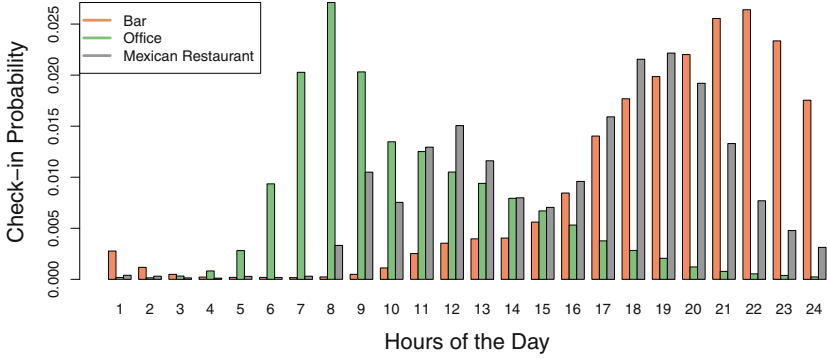
types. Instead of 4 possible locations in Beverly Hills, we are now faced with 163 possible locations. At least one *food* location exists in 44 of the 112 cells leading to a ratio of 0.393 and an *effort value* of 0.127. A similar adjustment is seen in East Los Angeles and for the greater Los Angeles region overall. Note that the broad activity of going out for food, even more specifically, appetizers, has not been lost through adjusting the text. By simply publishing a more generic term as part of her content, the publisher decreased her ability to be found in Beverly Hills dramatically.

## 5.5 Temporal Baseline

In addition to the textual and regional content specified in the examples above, one could imagine that someone might also tag their post with some type of temporal information. For example, a user might add the time *Friday at 7 pm* (e.g., as a meeting time) to the text.

In this example, the time is reported to a high granularity, permitting us to employ the 168 band temporal signatures in determining the place type probability. Taking the temporal signatures for each place type, we can directly compare the probabilities for Friday (Fig. 5) at 7 pm. For the purposes of this example, we have reduced our set of 421 place types to the three shown in this figure. Of these three, *Mexican restaurant* is the place type showing the highest probability at this time. Based on this information alone, we make the assumption that the user is at a *Mexican Restaurant* in Beverly Hills. This is in agreement with our text-based topic analysis discussed in Sect. 5.1.

This is not the entire story, however. While *Mexican restaurant* shows the highest temporal probability at 7 pm on a Friday, visually, it is followed quite closely by *bar* (Fig. 5). Computationally we can quantify this concern by referencing the *information entropy* for the hourly temporal signatures (a sample is shown in Table 2). Friday at 7 pm lists the fourth highest entropy value. The high entropy of this band tells us that in general, at 7 pm on a Friday night, people tend to be at quite a range of



**Fig. 5** Hour resolution temporal bands for *Bar*, *Office* and *Mexican Restaurant* on Friday

place types. Conceptually, this makes sense as this is the start of the weekend, and people could be engaging in a range of activities (e.g., watching a movie, at a bar, eating dinner, etc.). Knowledge of this high entropy reduces our certainty in determining the place type of the user and therefore has an impact on our overall ability to establish the spatial location of the user. The influence of temporal bands can be quantified using Eq. 5, where  $PT_{tp}$  represents the temporal probability of the given time band,  $\max(tp)$  is the maximum temporal band value, and  $PT_E$  is the information entropy of the given time band.

$$PT_{Time} = PT_{tp}/\max(tp) \times W + (1 - PT_E/\max(E)) \times (1 - W) \quad (5)$$

If we set the weight component  $W$  equal to 0.5 and assume a time of 7 pm on Friday, *Mexican restaurant* produces a  $PT_{Time}$  value of 0.382, while *Bar* lists a value of 0.345. Importantly, the information entropy values remain the same in this case. This allows us to compare place types across different temporal bands.

What would happen if instead of Friday at 7 pm, the user tweets out her message 1 h later? The information entropy for 8 pm on a Friday is 5.852 (compared to 5.932 at 7 pm). The order of temporal probabilities has shifted as well with *bar* now slightly more probable than *Mexican restaurant*, 0.022 and 0.019 respectively. These changes lead to revised  $PT_{Time}$  values for the two place types. *Mexican restaurant* has dropped to 0.351 while *Bar* has risen to 0.389. Though minute, a 1 h adjustment has had a significant impact on determining the place type. At 8 pm on Friday, the temporal bands now indicate that the user is likely at a bar.

## 5.6 A Combined Approach: Thematic and Temporal Bands

We now need to combine the two values calculated through referencing the thematic and temporal bands into a single value which indicates the most likely place type for



**Table 6** Statistical approach to determining place type based on temporal and thematic bands

Time	Thematic		Temporal		Combined	
Place type	<i>Mex</i>	<i>bar</i>	<i>Mex</i>	<i>bar</i>	<i>Mex</i>	<i>bar</i>
Friday 7 pm	0.733	0.607	0.381	0.351	<b>0.558</b>	0.542
Friday 8 pm	0.733	0.607	0.345	0.389	0.521	<b>0.543</b>

the user. In the case of Friday at 7 pm, both the temporal band and thematic band indicate that the user is likely at a *Mexican restaurant*. One hour later offers a different perspective with the textual content indicating a *Mexican restaurant* and the temporal component suggesting a *bar*. A single value can be calculated through Eq. 6. Note that the equation gives the option of weighting one component over another.

$$PT_{Prob} = PT_{Theme} \times W + PT_{Time} \times (1 - W) \quad (6)$$

With equal weights of 0.5, Table 6 shows the resulting place types depending on time and theme. The thematic properties of both Mexican restaurant and bar remain the same across time, while the temporal properties change based on the values computed in Eq. 5. The combined value is calculated through Eq. 6. Not surprisingly, the results suggest that the user is likely at a Mexican restaurant on Friday at 7 pm, since both the thematic and temporal values agree. More interestingly, at 8 pm, this method determines that the user is slightly more likely to be at a *bar*, even though the content suggests that she is likely to be at a *Mexican restaurant*.

## 6 Conclusions and Future Work

In this work we discuss the use of semantic signatures for exposing location information about a user through the content that she publishes. These semantic signatures, described through various spatial, temporal, and thematic bands mined from user-generated geosocial content, have shown to be an important basis on which the place type of an individual's location can be determined. Despite omitting or masking geographic coordinates, the methods presented in this work show that a person's location can still be revealed through comparing the signatures to non-geotagged content published by an individual. We propose a method to compute the location indicativeness of the signatures, i.e., the ability to locate somebody based on their published content.

Our initial findings suggest that protecting a user's geographic coordinates and other potentially revealing characteristics, such as ethnicity, is not sufficient as everyday digital footprints can give away the user's location as well. These findings, for instance, could be used to develop mobile applications that helps users, e.g., political activists, to make small changes to their content in order to better protect their geo-privacy.

Future work in this area will focus on expanding the range of semantic signatures. For example, the data collection for check-ins is currently being expanded to look at yearly data with the goal of exploiting seasonal effects on place type check-ins. Furthermore, hyperlocal data such as events could be used to enhance the robustness of these signatures. In addition, we hope to expand this work into a prototype application or browser plug-in that reports on the level of location privacy that is attainable based on the content as well as spatial and temporal information that someone publishes.

## References

- Adams B, Janowicz K (2012) On the geo-indicativeness of non-georeferenced text. In: ICWSM, pp 375–378
- Almuhimedi H, Schaub F, Sadeh N, Adjerid I, Acquisti A, Gluck J, Cranor LF, Agarwal Y (2015) Your location has been shared 5,398 times!: a field study on mobile app privacy nudging. In: The 33rd annual ACM conference on human factors in computing systems (CHI). ACM, pp 787–796
- Armstrong MP, Rushton G, Zimmerman DL (1999) Geographically masking health data to preserve confidentiality. *Stati Med* 18(5):497–525
- Batty M (1974) Spat Entropy. *Geogr Anal* 6(1):1–31
- Beldad A, Kusumadewi MC (2015) Heres my location, for your information: the impact of trust, benefits, and social influence on location sharing application use among indonesian university students. *Comput Hum Behav* 49:102–110
- Benisch M, Kelley PG, Sadeh N, Cranor LF (2011) Capturing location-privacy preferences: quantifying accuracy and user-burden tradeoffs. *Pers Ubiquitous Comput* 15(7):679–694
- Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *J Mach Learn Res* 3:993–1022
- Bohn J, Coroamă V, Langheinrich M, Mattern F, Rohs M (2005) Social, economic, and ethical implications of ambient intelligence and ubiquitous computing. In: *Ambient intelligence*. Springer, Heidelberg, pp 5–29
- Clarke CC (2015) A multiscale masking method for point geographic data. *Int J Geogr Inf Sci* 30(2):1–16
- de Montjoye YA, Hidalgo CA, Verleysen M, Blondel VD (2013) Unique in the crowd: the privacy bounds of human mobility. *Sci Rep* 3
- Dixon PM (2002) Ripley's K function. In: *Encyclopedia of environmetrics*
- Duckham M, Kulik L (2006) Location privacy and location-aware computing. *Dyn Mob GIS: Investigating Change Space Time* 3:35–51
- Dwork C (2011) Differential privacy. In: *Encyclopedia of cryptography and security*. Springer, pp 338–340
- Gambis S, Killijian MO, del Prado Cortez MN (2010) Show me how you move and i will tell you who you are. In: *Proceedings of the 3rd ACM SIGSPATIAL international workshop on security and privacy in GIS and LBS*. ACM, pp 34–41
- Goodchild MF (2015) Space, place and health. *Ann GIS* 21(2):97–100
- Hampton KH, Fitch MK, Allshouse WB, Doherty IA, Gesink DC, Leone PA, Serre ML, Miller WC (2010) Mapping health data: improved privacy protection with donut method geomasking. *Am J Epidemiol* 172(9):1062–1069
- Janowicz K (2012) Observation-driven geo-ontology engineering. *Trans GIS* 16(3):351–374
- Kounadi O, Leitner M (2015) Spatial information divergence: using global and local indices to compare geographical masks applied to crime data. *Trans GIS* 19(5):737–757
- Lin J (1991) Divergence measures based on the shannon entropy. *IEEE Trans Inf Theory* 37(1):145–151

- Lindqvist J, Cranshaw J, Wiese J, Hong J, Zimmerman J (2011) I'm the mayor of my house: examining why people use foursquare—a social-driven location sharing application. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, pp 2409–2418
- Mahmud J, Nichols J, Drews C (2014) Home location identification of twitter users. *ACM Trans Intell Syst Technol (TIST)* 5(3):47
- McCallum AK (2002) MALLETT: a machine learning for language toolkit. <http://mallet.cs.umass.edu>
- McKenzie G, Janowicz K (2014) Coerced geographic information: The not-so-voluntary side of user-generated geo-content. In: Eighth international conference on geographic information science
- McKenzie G, Janowicz K (2015) Where is also about time: a location-distortion model to improve reverse geocoding using behavior-driven temporal semantic signatures. *Comput Environ Urban Syst* 54:1–13
- McKenzie G, Janowicz K, Gao S, Yang JA, Hu Y (2015) POI pulse: a multi-granular, semantic signatures-based approach for the interactive visualization of big geosocial data. *Cartographica: Int J Geogr Inf Geovis* 50(2):71–85
- Seidl DE, Jankowski P, Tsou MH (2015) Privacy and spatial pattern preservation in masked GPS trajectory data. *Int J Geogr Inf Sci* 1–16
- Seidl DE, Paulus G, Jankowski P, Regenfelder M (2015) Spatial obfuscation methods for privacy protection of household-level data. *Appl Geogr* 63:253–263
- Shannon CE (1948) A note on the concept of entropy. *Bell Syst Tech J* 27:379–423
- Sweeney L (2002) k-anonymity: a model for protecting privacy. *Int J Uncertainty Fuzziness Knowl Based Syst* 10(05):557–570
- Weber RH (2010) Internet of things—new security and privacy challenges. *Comput Law Secur Rev* 26(1):23–30
- Zandbergen PA (2014) Ensuring confidentiality of geocoded health data: assessing geographic masking strategies for individual-level data. *Adv Med* pp 1–14
- Zhang S, Freundschuh SM, Lenzer K, Zandbergen PA (2015) The location swapping method for geomasking. *Cartography Geogr Inf Sci* 1–13

**Part III**  
**Analysis and Visualization of (Big) Spatial**  
**Data**

# Modelling Spatial Patterns of Outdoor Physical Activities Using Mobile Sports Tracking Application Data

Rusne Sileryte, Pirouz Nourian and Stefan van der Spek

**Abstract** The paper presents a workflow for collecting, structuring and processing geo-referenced recreational mobility data from a sports tracking application as to monitor recreational usage of urban spaces. The data collected include GPS trajectories of people walking, jogging, and running for recreational purposes in European cities. The presented workflow includes systematic steps for aggregating the trajectories and attributing them to a spatial network model called Urban Space Network. The nodes of this network are the navigable spaces or streets and its links are the connections between them. A method is proposed to find a fuzzy notion of recreational space usage, using the number of distinct application users whose trajectories have been accounted for the space in question. The fuzzified space usage values are then attributed to the nodes of the network. This model can be primarily used to observe actual patterns of space usage and has the potential to be used as ‘ground truth data’ for validating and calibrating network-based models of recreational mobility. Patterns revealed by the workflow can be used to study where outdoor physically active mobility happens and where it is absent. Thus the proposed workflow can provide spatial and objective insight useful in planning, management and governance of cities in promoting active mobility that is already a rather global trend in urbanism.

**Keywords** Mobile sports tracking application data · Mobility data · Urban space network · Recreational usage

---

R. Sileryte (✉) · P. Nourian · S. van der Spek  
Faculty of Architecture and the Built Environment,  
Delft University of Technology, Delft, The Netherlands  
e-mail: r.sileryte@tudelft.nl

P. Nourian  
e-mail: p.nourian@tudelft.nl

S. van der Spek  
e-mail: s.c.vanderspek@tudelft.nl

## 1 Introduction

Guidelines, policies, and regulations have already begun being created to encourage physical activities in cities (NYC DDC 2010). Therefore, a better understanding of where physical activities are conducted would enable more effective policy interventions to promote physically active lifestyles in different built-environment contexts. Probabilistic models on networks could bring the benefit of predicting the likely effects of changes on facilitating or hindering the phenomena.

A substantial amount of data related to outdoor physical activities can be obtained from mobile sports tracking applications. This data is constantly generated worldwide by smart device users willing to record their spatio-temporal activity. The available data is extremely big, provided voluntarily and in large numbers, public, and therefore not raising privacy issues, always available up-to-date, features the same world-wide method of collection and finally, it is constantly growing.

However, mobile sports tracking application data is collected for personal motives and the motivation of application providers is rather satisfaction of a user, than aggregate data collection. In fact, nor the data is available in a single click; neither application providers supply an interface for ready-made free access due to likely privacy issues. Furthermore, the available data is raw and so vast in size, that structuring, filtering and aggregation procedures need to be applied. The GPS (Global Positioning System) data is just a sequence of points in Euclidian space, which need to be mapped and analysed in a 'network space', while constructing an appropriate 'space network' for any of the cities is a non-trivial task itself.

Three cities have been chosen as case studies based on the availability of data provided by Eurostat, similarity of rate between city's population and the chosen sports tracking application users. The chosen ones are namely Vilnius (Lithuania), Valencia (Spain) and Gothenburg (Sweden) with a ratio of 2–3 spotted application users per 1000 inhabitants. This paper describes a workflow for observing usage of urban spaces for running and walking activities in various European cities based on the mobile sports tracking application data. The developed workflow exceeds the scope of a single project in that it aims to simplify and standardize the preprocessing of mobility data.

The following section delineates related work and clarifies how this research is distinguished from the similar works. The third one explains the procedure for acquisition of required data, which is followed by constructing an Urban Space Network. Next section explains a method to interpret the relative space usage and delivers the results. Finally, the conclusions are drawn preceded by the discussion and recommendations.

## 2 Related Research

Previous studies of physically active human mobility rely on manual data collection by directly observing chosen locations during certain short periods (Floyd et al. 2008), asking residents in surrounding areas of a park to complete 7-day physical activity logs that include the location of their activities (Kaczynski et al. 2008), comparing recipients places of residence with their physical activity registered by accelerometers (Cohen et al. 2006) or even using a telephone survey (Lopez 2004).

All the previously mentioned methods are performed by intensive human labour and can only be applied on relatively small-scale measurements. In addition, walkability has been mostly studied as a property of the entire neighbourhood rather than particular urban space or their networks. Yet a number of researches have explored definite relation between street walkability and the configuration of urban street network and its attributes such as transport nodes, land use, infrastructural elements, major attractors, aesthetic features, etc. (Hillier and Iida 2005; Gebel et al. 2007).

Mobile GPS data has been already used by various researches in order to investigate spatial mobility patterns in urban settings. Van der Spek et al. (2009, 2013) have carried out a research which aims to explain pedestrians' behaviour in various cities by deploying GPS tracking system supplemented with questionnaires. Piorkowski (2009) has pioneered in using mobile sports tracking application data for analytic purposes. He aimed on enhancing location privacy and designing better context-aware services. Ferrari and Mamei (2011, 2013) have used Nokia Sports Tracker application data to identify the areas and temporal routines of a city most used for a given sports activity, highlight cultural and climate-related differences among cities and show differences in the routine behaviour of various demographic and social communities. Oksanen et al. (2015) aim to extract frequently used routes from massive public workout data in order to define the most popular routes as a suggestion for application users.

The goal of this research, in respect to the previously described ones, is to use mobility data in tandem with other open data sources, for modelling recreational usage in a network rather than Euclidean space through an automated procedure, which later allows utilising the model as a ground truth for the further investigations of the desired phenomena.

## 3 Required Data

### 3.1 *Mobile Sports Tracking Application Data*

Sports tracking applications cannot provide direct access to their databases due to privacy issues; however, some of them display public workouts on dedicated websites. In that case, users are consent to publicly displayed (but not distributed) personal data. After considering a number of applications, Endomondo was chosen due

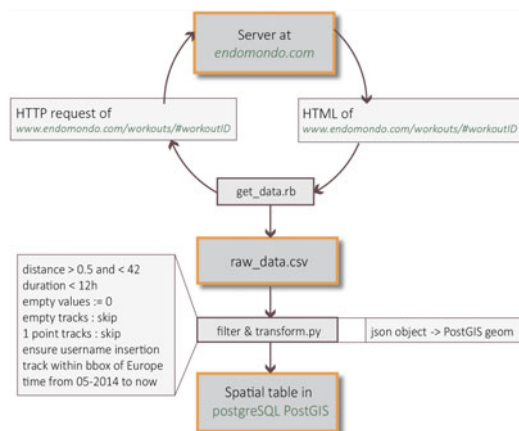
to its popularity rate and relatively convenient data access. Workouts are available to be viewed on [www.endomondo.com/workouts/+workoutID](http://www.endomondo.com/workouts/+workoutID) unless specified to be private by a user.

Every workout is a JSON (Java Script Object Notation) object embedded into an HTML (Hyper Text Mark-up Language) code of a page. Additional to the GPS trajectory, other available attributes include type of the workout (running, walking, etc.), date and time, user name (id), distance, duration, average and maximum speed, burnt calories, hydration, altitude and weather data. A user can choose to make any of these attributes private, edit the values or delete the workout permanently at any time (Endomondo 2015). Tracking is based on a GPS receiver and therefore is dependent on the characteristics of each individual device.

A tutorial in Barsukov (2014) has been used as a basis for the data acquisition framework. The adapted scheme of data acquisition is shown in Fig. 1. A local script sends an HTTP (Hyper Text Transfer Protocol) request to the server for a workout with a chosen ID and either gets a negative response (in case the workout is listed as private or it has been deleted) or a positive response and an HTML code of a page, in which case the algorithm continues exploring the data. If GPS trajectory is available and listed as ‘Running’ or ‘Walking’, the required fields are output into a text file, which is later filtered based on multiple criteria and transformed from a JSON object into a PostGIS geometry.

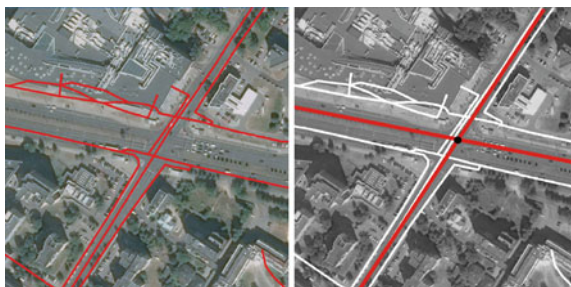
Data samples were collected every 8 days in a period of May 2014–May 2015, aiming to have sufficient data throughout the full year and a variety of weekdays as well as occasional public holidays. Data acquisition process took approximately 1248 h and resulted in more than 3.5 million valid GPS tracks of almost a million distinct users within the territory of Europe. The collected data is evenly distributed throughout a day, all seasons and weekdays.

**Fig. 1** Workflow for the acquisition of mobile sports tracking application data





**Fig. 2** Original OSM street network expressed in polylines (*left*) and actual perceived urban space needed for the active recreational travel analysis (*right*)



### 3.2 *OpenStreetMap*

Lately the road network provided by the OpenStreetMap (OSM) is often chosen to form the backbone of urban networks because of its universal coverage and standard defined for all modes of transport. Besides, due to its open access nature and volunteered contribution, OSM can have a very good level of completeness (Mooney 2015) and it includes representation of paths for non-motorised means of transport, which is essential for the analysis of the jogging and walking movement patterns.

A single polyline in OSM dataset usually describes a single path; however, in some cases it can also form a boundary polygon and represent an area, which stands for various parks, squares and even wider boulevards within which no further paths are drawn. The elimination of one of the entity types would result into missing network connections, which would result into misleading snapping of GPS tracks and wrong evaluation of network configuration. Thus, the inconsistency of entity types needs further attention while processing the dataset.

Girres and Touya (2010) have listed a number of possible problems regarding the OSM street segment geometry and topology, including duplicate overlapping or missing segments, intersection nodes, etc. In addition to these, the high level of detail presented in the street network is redundant and even confusing for the later applied algorithms. While a single street in OSM can be represented by multiple lines, which stand for different car lanes, bicycle lanes, footpaths and sidewalks, all these lines are still perceived as a single space by a person engaged into an active recreational activity (Fig. 2). Therefore, the OSM street network has been later processed and coupled with additional datasets in order to overcome the identified problems and provide a neat Urban Space Network.

### 3.3 *European Urban Atlas Road Land-Use Data*

The Urban Atlas (UA) is a joint initiative of the European Commission Directorate-General for Regional Policy and the Directorate-General for Enterprise and Industry with the support of the European Space Agency and the European Environment

Agency. Its aim is to provide pan-European comparable and freely accessible land use and land cover data for Large Urban Zones with more than 100,000 inhabitants. The resulting vector maps provide land-use classification for 21 different land-use classes with minimum overall accuracy of 85 % and positional accuracy of  $\pm 5$  m (Urban Audit 2007).

The ‘Roads and Associated Land’ class is represented by a single polygon, which comprises a city road network. The associated lands are: slopes of embankments; areas enclosed by roads, without direct access; fenced areas along roads; noise barriers; rest areas, service stations and parking areas; railway facilities; foot- or bicycle paths parallel to the traffic line; green strips and alleys (with trees or bushes). Since road lanes, cycle lanes, pedestrian paths, complicated crossroad lanes and street crossings are all covered by a single polygon, it becomes easier to determine a single space than in case of OSM dataset.

However, in order to use the polygon as a network, it has to be converted into polyline features. It also does not contain paths meant for non-motorised means of transport, and lacks most of the bridges. Due to these reasons, the UA road-land-use polygon needs to be both processed and combined with OSM data to satisfy research needs.

## 4 Urban Space Network

### 4.1 Definition of an Urban Space Network

Generally, a street network is defined as a system of interconnecting lines that represent a system of roads for a given area (Mora and Squillero 2015). In case of this research, Urban Space Network (USN) is a network of interconnected public urban spaces, which are navigable for humans but not necessarily for vehicles. It can be defined as a network whose edges represent a single human-navigable space (i.e. street, footpath, parkway, square, etc.), and its vertices are intersections of such spaces in which there are more than two choices of moving direction. Thus, the conventional street network is merely a subset of the USN.

Specifically, the USN is a topological skeleton of the navigable urban spaces. This topological construct can be represented as a (dual) graph whose nodes and links represent spaces and connections between them respectively. From a cognitive perspective, having navigable spaces as the nodes has a number of advantages for later studies, i.e. the possibility of modelling cognitive costs of going from one space to another. However, the basic reason why recreational activities need to be modelled and analysed in a network space instead of Euclidean space is the assumption that human movement in cities is steered by the built as well as natural environment and its implied movement restrictions. E.g. while the two banks of a river might be very close to each other in an Euclidean space, they might be extremely far away in a

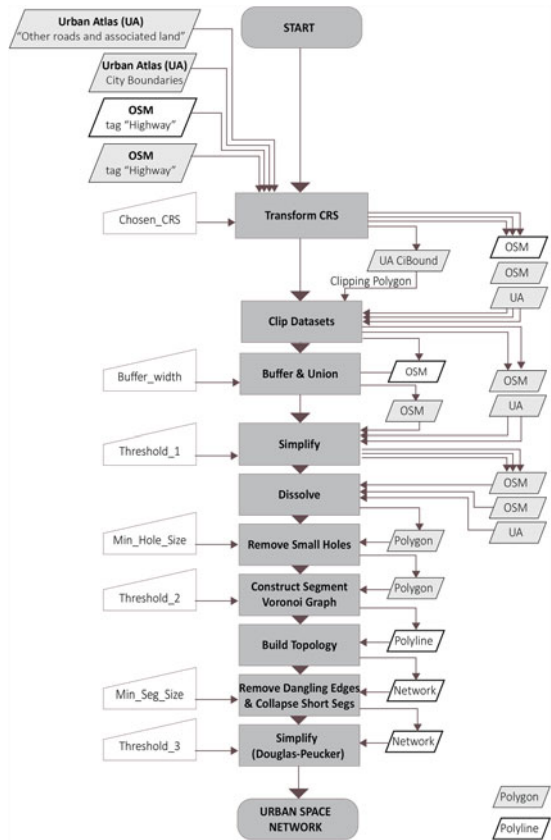
network space and therefore the same built environment factors that determine the usage of one bank may have no influence on the other one.

Furthermore, the USN must be generalised, i.e. contain a single edge for a single perceived space and a single node of intersection. It has to be noted that pedestrians, in contrast to vehicles, are not compelled to use designated paths, e.g. a piece of road between 2 crossings together with all its sidewalks, bicycle and car lanes, and other associated land is considered one navigable space if there are no possibilities to navigate from it to another one.

### 4.2 Dataset Integration

The biggest mismatch between the UA and OSM datasets is different type of entities (polygon in case of UA and polyline and polygon in case of OSM). There are also geometrical mismatches or cases when streets in one dataset do not appear in the

**Fig. 3** The framework of Urban Space Network generation, integrating OSM and UA datasets



other. In order to overcome these issues and correct the topological errors apparent in the OSM dataset, a new polygon-based approach has been developed as in Fig. 3.

The first step of dataset integration is to ensure that all of them belong to the same coordinate system. For this research, along with the default WGS 84, Europe Albers Equal Area Conic (ESRI:102013) has been chosen for visualisation and calculations since it is adapted to fit Europe and uses metric unit system, thus no further recalculation from degrees to meters is needed.

Another important step is network generalization and simplification. Automated generalisation has long been a research effort of cartographers (Jiang and Claramunt 2004; Savino 2011; Li et al. 2014). While the previously mentioned researches mainly treat road networks formed by a single dataset and generalisation for scaling purposes, in case of this research an additional challenge is created by using multiple datasets and pedestrian routes, which do not follow such strict patterns as road lanes.

In order to unify the type of entities OSM polylines are buffered and that way transformed into a single polygon. The buffer width is decided based on the general level of detail set for the network's generalisation. When both datasets have the same type of entity, they can be dissolved into a single polygon (Fig. 4). However, beforehand they are simplified using a well-known Douglas-Peucker algorithm with the threshold of 1 m in order to reduce computation time. A number of holes, which do not form a substantial gap between the paths, are cleaned by removing polygon rings smaller than a chosen threshold.

After the datasets are united into a single polygon, its centreline needs to be extracted in order to return to the polyline type of entity. The centreline of a polygon is also an approximation of all the neighbouring paths into a single network edge. The Boost library, which provides free peer-reviewed portable C++ source



**Fig. 4** Polygon-based approach for integrating UA and OSM datasets: *grey polygons* represent OSM line features buffered by a chosen distance, *white polygon* represents OSM pedestrian area; *green polygon* comes from UA dataset of land use type 'Roads and associated land'



**Fig. 5** Part of the resultant USN of Valencia overlaid with Google Earth image

libraries, has been used. The Boost.Polygon.Voronoi has been used to compute a Segment Voronoi (Delaunay) Graph, which takes line segments as an input; therefore, no geometry densification is needed.




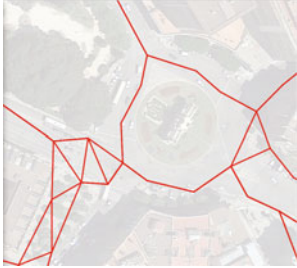
The post processing is needed in order to decrease the complexity of the network and that way save computation time as well as to facilitate the interpretation of space usage values. This step includes building network topology, removing dangling edges, collapsing short segments and simplifying polylines. The example of resultant Urban Space Network in Valencia is shown in Fig. 5.

### **4.3 Validation**

The benefits of the developed USN construction method are validated comparing it with a network obtained using a more commonly used approach for network generalization and simplification: that is by iteratively using topological cleaning tools followed by a vertex-snapping algorithm. The method, differently than the proposed one, is based on polylines; therefore, initially, centrelines have to be extracted from the UA road polygon and OSM pedestrian areas. Table 1 shows the differences between the two approaches considering a number of relevant aspects.

While both methods have their own benefits and drawbacks, the polygon-based method fits the purpose of this research better, since it provides a simpler outcome with less redundant connections, easily removable artefacts and a single network edge per single perceived space.

**Table 1** Comparison between the polygon-based and the polyline-based dataset integration and network generalisation methods

	Polygon-based	Polyline-based
Topological validity	The outcome network is always topologically valid	The outcome needs to be cleaned from topological errors: mainly overlaps, pseudo-nodes and duplicate geometries
Junction simplicity	Junctions need to be further processed by collapsing short segments 	Bigger junctions (more than 4 ways) tend to create artefacts 
Redundant segments	Centerline extraction algorithm creates redundant dangles 	Snapping algorithm results into redundant connections 
Geometric distortions	Geometric distortions do not exceed the buffer width	Polyline geometry can get severely distorted while moving all vertices of a polyline into different directions
Attributes	No attributes preserved	Attributes are preserved
Execution time	The crucial time needed for both methods is the extraction of Segmented Voronoi (Delaunay) Graph edges, which lie inside the polygon; the buffering time in polygon-based method is comparable with the snapping time in polyline-based method	



## 5 Space Usage

### 5.1 Filtering GPS Trajectories

In the initial state acquired GPS tracks are rather a set of coordinates, which are not in any way related to the USN (Fig. 6), therefore in order to define the usage measure, GPS tracks need to be processed filtered and snapped to the underlying network.

Filtering of GPS points is needed in order to remove outliers, which appear in GPS trajectories due to various reasons: lack of satellites in sight due to environment obstructions, ‘cold start’ or signal multipath. Filtering outliers has been detached from the initial filter that takes place while writing data into the database in order to reduce total filtering time and be able to process only the relevant GPS tracks. However, this is a trade-off between filtering time and loss of individual GPS point attributes. Consequently, such methods as proposed by Schuessler and Axhausen (2009), Auld et al. (2013), Biljecki (2010), which suggest removing the outliers from GPS data based on the unrealistic altitude, sudden speed and acceleration jumps or sudden changes in heading become unavailable.

In case of this research the under-filtering is less of a problem than over-filtering due to the snapping algorithm, which relies on a sequence of points. In addition, scarce data should not be lost during the outliers filtering. Therefore, the definition of an outlier has been formulated as following: it is a point that lies from both of its neighbours further than three times the median while the distance between the neighbours is less than the smaller distance between the point and each of its neighbours. Median refers to the median distance between two consecutive points calculated for each GPS track individually.

The heuristics of using three medians comes from the evaluation of a sample set of 100 randomly chosen GPS tracks from different cities, which can be visually confirmed as not having outliers. The calculation is based on the ratio between the

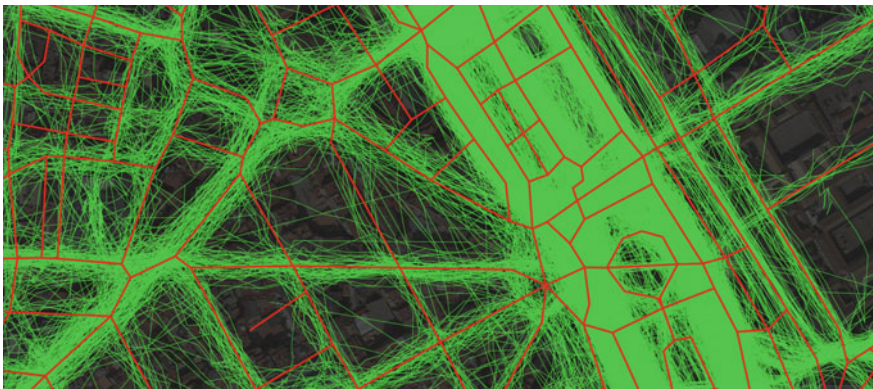


Fig. 6 GPS track (transparent green) on a USN (single line red) prior to snapping

median distance between two consecutive points and a maximum deviation from the median in each of the test tracks. The average value of 100 tested ratios appeared to be 2.2215 with standard deviation of 0.8429. Thus, if the ratio between a point and its neighbours is higher than the mean ratio plus the standard deviation, the point can be considered as suspicious.

## 5.2 *GPS Track Snapping on an Urban Space Network*

The reason for the geometric mismatch between the GPS tracks and the space network lies both in the inaccuracy of the GPS measures (Ranacher et al. 2015) and the data used to construct the USN. Most of the map-matching algorithms tend to deal with the GPS tracks of vehicle movements, which are in many aspects different from the workout data. E.g. runners as in contrast to vehicles, do not necessarily stay on a designated path, they can change moving direction at junctions as well as in the middle of a path, do not have any movement restrictions or predictable moving speed. Moreover, nothing is known about the characteristics of a GPS device, positioning data quality, satellites in range or the frequency of GPS fixes. Due to these reasons, most of the advanced algorithms cannot be implemented and therefore only geometrical and topological data is used for snapping.

The GPS snapping algorithm has been developed based on the algorithms proposed by Marchal et al. (2004), Yang et al. (2005), Quddus and Washington (2015). It is a topological algorithm, which relies on the multiple hypothesis' technique. It allows to keep track of several positions or paths at once and to select eventually which candidate is the best. The first point is snapped to the two closest segments of the extracted piece of the whole network. Later, the best-fit edge is decided by checking the following points and choosing the best matching one. The path is augmented through topological connections of the best fitting edge, always choosing two of them based on a single point and deciding the better one based on a sequence of points up until the last GPS point is reached. The sample results of map matching algorithm can be seen in Fig. 7.

The accuracy of the map-matching algorithm has been computed by visually comparing the GPS track with the assigned USN edges of 25 randomly selected samples in Vilnius city, which all together make up almost 5000 GPS points. Mapping accuracy has been computed as a number of correctly assigned network edges over the number of all edges considered (assigned, over-assigned and under-assigned) and results into 85 % of overall mapping accuracy, which is reasonable for a geometrical/topological, map matching algorithm. The standard deviation of GPS points to the network edge they are snapped to is 15.859 m.

Moreover, over-assignment is more frequent than under-assignment. This happens often due to lack of edges in the network, i.e. recreational activities happening in spaces which are not represented by any edge in the network. This can happen



**Fig. 7** GPS snapping algorithm: *bright line* indicates original GPS track; *dark line* indicates USN edges to which the GPS track has been snapped



because of two reasons—either the lack of an existing path in the OSM or UA data or the absence of a path as such, e.g. running in out-door stadium, in meadows or private lands.

### 5.3 Value of Recreational Usage

After snapping GPS tracks every space in a network gets an attribute of a number of distinct application users spotted therein. While the overall goal is to model the recreational space usage, i.e. give an indication to every space of how much the particular space is used for recreation, the actual counts cannot clearly represent the measure. Moreover, literal quantification of recreational usage is impossible, since it is a rather qualitative notion.

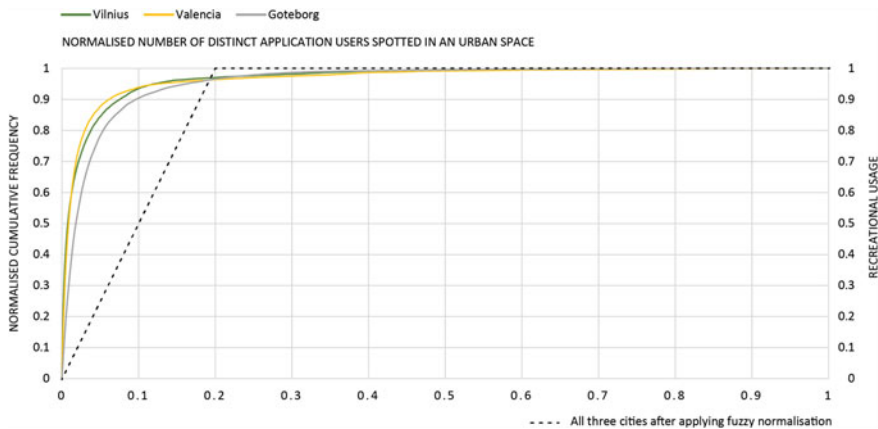
In order to quantify a qualitative measure, a fuzzy notion of likeliness has been used (Klir and Yuan 1995). It describes how likely it is that a space is used for active recreational travels and is measured in the range of 0–1, where 0 means no usage and 1 means that a space is definitely used. All values in between indicate how much a space is used compared to the other ones. It is important to note that values of recreational usage are not numerical but rather of an ordinal nature. Fuzzy normalisation primarily serves for visualisation purposes enabling more intuitive and comprehensible overview of space usage.

In order to perform the fuzzy normalization, first the cumulative frequency for each space usage value is calculated. The number of distinct users spotted in a space over the whole study period is denoted as  $u$ . The set of spaces with  $k$  or less users is defined as  $S(k) = \{u | u \leq k\}$ ; as to which the cumulative frequency is  $f^C(k) = |S(k)|$ . Then normalized space usage and its normalized cumulative frequency are defined

respectively as:  $k^n = k/k_{max}$ , in which  $k_{max}$  is the maximum number of spotted distinct users for a single space in the whole range of spaces; and  $f_n^C(k) = (f^C(k) - f^C(0))/(f^C(k_{max}) - f^C(0))$  in which  $f_n^C(k)$  denotes the normalized cumulative usage frequency of  $k$  or less users.

Any space which has a number of spotted application users above 0, is regarded as ‘somewhat used for recreation’. The distribution of values differ per city due to the different proportions of network size and number of application users and because of different distributions of recreational activity, which are dependent on individual characteristics of the built environment. Higher number of attractive spaces shares out the total number of the users, while lower amount of attractive spaces concentrates the users within them.

However all normalized frequencies have similar distributions and approximately even out at one point corresponding to 20 % of the maximum usage. Simply put, in all three cases only 3 % of all the network spaces have a number of spotted users higher than 20 % of the maximum registered. Therefore, this point has been used as a reference for the likeliness coefficient. For example, in case of Vilnius the maximum number of spotted users in an urban space is 592, which means that all spaces, which have 118 or more users, are regarded as ‘used for recreation’. Accordingly, a space, which has 59 users, is considered to have recreational usage value of 0.5. Figure 8 illustrates the dependency between the number of users spotted in a single space and its recreational usage.



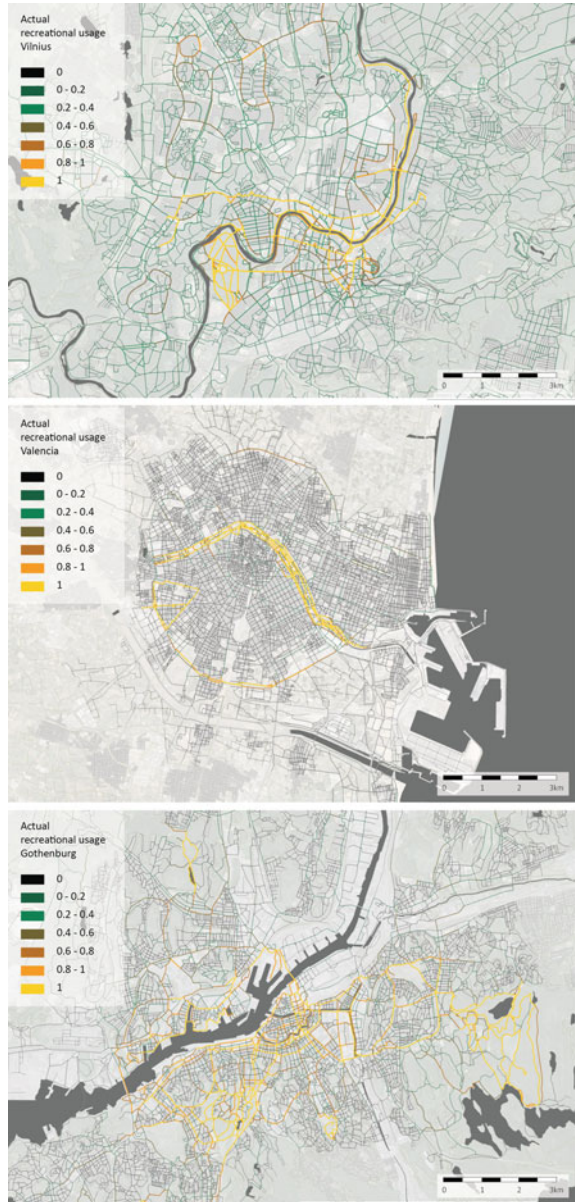
**Fig. 8** The *solid lines* correspond to the normalised cumulative frequency of urban spaces, while the *dashed line* corresponds to the recreational usage of those spaces, when the normalised number of spotted distinct application users is the same. All normalisations are done according to the cities’ own minimum and maximum values

## 6 Results

The resulting maps for all three case study cities can be seen in Fig. 9.

While examining the maps closer, it can be noticed that large recreational areas in all of the three cities attract the most of the recreational activities. While these

**Fig. 9** Visualisation of USN coloured according to its recreational usage values as defined after the fuzzy normalisation; case studies of Vilnius, Valencia and Gothenburg



results are expected, the interesting things can be noticed while closer examining non-recreational areas. For example, in case of Vilnius, some heavy traffic streets are used for recreation more than the nearby green zones, while in case of Valencia almost all recreational activities are concentrated in the parks or green alleys, leaving densely urbanised areas excluded. In case of Gothenburg, the most recreational spaces almost evenly spread throughout the city, interconnecting with each other and forming their own ‘recreational network’.

The constructed network can be overlaid with a number of related maps in order to visually inspect the relationships between different phenomena. However, more importantly, network nodes can also be attributed with a number of measures, such as space greenness, network centrality, land use, etc. in order to use quantitative methods to find associations between the values.

Finally, by looking at the maps, it can be noticed that it is not only the attractiveness of a single space that enables presence of recreational activity but also its position in a broader network of spaces; not in a sense of being in an attractive area but in a sense of being connected to other attractive spaces. In other words it is the position of a space in the network space that matters more than its position in the Euclidean Space. Therefore, these findings suggest that it is the network-based analysis that must play bigger role than the neighbourhood-based analysis. This highlights the role of USN as an essential construct in this research.

## 7 Discussion and Future Work

First of all, a collaboration between sports tracking application and a researcher would significantly improve the efficiency of data acquisition. Furthermore, knowing such characteristics as user age group, occupation, education, etc. might give a better overview of data validity and allow deeper investigation of recreational travel patterns. Currently, user group analysis is not possible due to the privacy matters.

Furthermore, the running and walking activities have been considered equally, while they might also have different movement patterns. In addition, various other types of recreational travels could be added among which recreational cycling, orienteering, roller skiing, skateboarding, etc. Generally, the collected data is limited to only one sports tracking application which limits the set of tracked individuals to those who have knowledge of a foreign language, possession of a smart phone, ability to use the application and, of course, having given a consent to be tracked. Therefore, it must be acknowledged that acquired data represents only a certain subset of all recreational travels conducted in a city, which may cause related bias to the research results.

Even though the integration of UA and OSM datasets improves the completeness of a USN, a number of paths and connections remain unknown. This problem could be tackled by upgrading the GPS network-snapping algorithm. The missing paths could be added to the constructed USN based on the clusters of GPS tracks.

This would also improve the mapping accuracy of the algorithm itself. Furthermore, some heavy traffic roads should rather be considered as barriers, so that only certain connections through them would be possible. Finally, buffering sometimes may cause connection of spaces, which actually do not reach in reality due to topography, water features, etc.

## 8 Conclusions

The conducted research has investigated how mobile sports tracking application data can be used to model and visualise the recreational usage of an Urban Space Network. An automatic and non-labour intensive method has been devised for data acquisition, management and processing. Collected GPS tracks have been filtered from blundering fixes and snapped to a USN with 85 % mapping accuracy. GPS tracks when aggregated per single network edge form a measure, which is later normalised using fuzzy normalisation methods, and represents how much a space is used for recreation compared to the other ones.

Before processing the mobility data, a systematic workflow has been developed for constructing an Urban Space Network using OSM data complemented with UA road land use data. The method relies on integration of datasets, generalisation and simplification through buffering linear features, combining all polygons and using Segmented Voronoi (Delaunay) Graph to extract polygon centreline, which, after minor processing and additional simplification is used as a representation of a USN. The constructed network is relevant for the desired type of analysis and differs from conventional street networks in that it includes paths for both motorised and non-motorised means of transport, which run through urban fabric as well as parks and urban forests. A particular characteristic of the USN is that it has low granularity, however, well-preserved space connectivity.

The visualisation of results has proved that analysing recreational usage in a network space instead of Euclidean space brings clearer insight and provides a basis for understanding and explaining the usage patterns and their associations with built environment effects. Finally, testing all processes and algorithms in parallel for three different case studies has ensured that the collected data as well as the developed methods would not be dependent on a specific urban structure and can be repeated for any of the European cities with sufficient application users.

**Acknowledgments** This paper is based on the Master thesis of the first author, written at the Technical University of Delft. The authors would also like to thank the thesis co-reader Dr. Hugo Ledoux for his considerate review and suggestions.

## References

- Auld J, Williams C, Mohammadian A (2013) Prompted recall travel surveying with GPS. In: Transport Chicago Conference, Zugegriffen, vol 15
- Barsukov N (2014) Generating running route maps. <http://barsukov.net/programming/2014/07/26/endomondo-code.html>, cited 1 Dec 2014
- Biljecki F (2010) Automatic segmentation and classification of movement trajectories for transportation modes. Master's thesis, TU Delft, Delft University of Technology
- Cohen DA, Ashwood JS, Scott MM, Overton A, Evenson KR, Staten LK, Porter D, McKenzie TL, Catellier D (2006) Public parks and physical activity among adolescent girls. *Pediatrics* 118(5):e1381–e1389
- Endomondo (2015) Endomondo sports trackers. <http://www.endomondo.com>
- Ferrari L, Mamei M (2011) Discovering city dynamics through sports tracking applications. *Computer* 44(12):63–68
- Ferrari L, Mamei M (2013) Identifying and understanding urban sport areas using nokia sports tracker. *Pervasive Mobile Comput* 9(5):616–628
- Floyd MF, Spengler JO, Maddock JE, Gobster PH, Suau LJ (2008) Park-based physical activity in diverse communities of two us cities: an observational study. *Am J Prev Med* 34(4):299–305
- Gebel K, Bauman AE, Petticrew M (2007) The physical environment and physical activity: a critical appraisal of review articles. *Am J Prev Med* 32(5):361–369
- Girres JF, Touya G (2010) Quality assessment of the french openstreetmap dataset. *Trans GIS* 14(4):435–459
- Hillier B, Iida S (2005) Network and psychological effects in urban movement. In: *Spatial information theory*. Springer, pp 475–490
- Jiang B, Claramunt C (2004) A structural approach to the model generalization of an urban street network. *GeoInformatica* 8(2):157–171
- Kaczynski AT, Potwarka LR, Saelens BE (2008) Association of park size, distance, and features with physical activity in neighborhood parks. *Am J Public Health* 98(8):1451
- Klir G, Yuan B (1995) Fuzzy sets and fuzzy logic, vol 4. Prentice Hall, New Jersey
- Li Q, Fan H, Luan X, Yang B, Liu L (2014) Polygon-based approach for extracting multilane roads from openstreetmap urban road networks. *Int J Geogr Inf Sci* 28(11):2200–2219
- Lopez R (2004) Urban sprawl and risk for being overweight or obese. *Am J Public Health* 94(9):1574–1579
- Marchal F, Hackney J, Axhausen K (2004) Efficient map-matching of large gps data sets: tests on a speed monitoring experiment in zurich. *Arbeitsbericht Verkehrs und Raumplanung* 244
- Mooney P (2015) An outlook for openstreetmap. In: *OpenStreetMap in GIScience*. Springer, pp 319–324
- Mora AM, Squillero G (2015) Applications of evolutionary computation. In: *18th European conference, EvoApplications 2015*, vol 9028. Springer
- NYC DDC (2010) New york city active design guidelines: promoting physical activity and health in design. <http://centerforactivedesign.org/dl/guidelines.pdf>, cited 29 Nov 2014
- Oksanen J, Bergman C, Sainio J, Westerholm J (2015) Methods for deriving and calibrating privacy-preserving heat maps from mobile sports tracking application data. *J Transp Geogr* 48:135–144
- Piorkowski M (2009) Sampling urban mobility through on-line repositories of gps tracks. In: *1st ACM international workshop on hot topics of planet-scale mobility measurements*, ACM
- Quddus M, Washington S (2015) Shortest path and vehicle trajectory aided map-matching for low frequency GPS data. *Transp Res Part C: Emerg Technol* 55:328–339
- Ranacher P, Brunauer R, Van der Spek SC, Reich S (2015) GPS error and its effects on movement analysis. arXiv preprint [arXiv:150404504](https://arxiv.org/abs/150404504)
- Savino S (2011) A solution to the problem of the generalization of the italian geographical databases from large to medium scale: approach definition, process design and operators implementation. Ph.D. thesis, Universita di Padova

- Schuessler N, Axhausen K (2009) Processing raw data from global positioning systems without additional information. *Transp Res Record: J Transp Res Board* 2105:28–36
- Urban Audit (2007) State of european cities report. Tech. rep, Study contracted by the European Commission
- Van der Spek S, Van Schaick J, De Bois P, De Haan R (2009) Sensing human activity: GPS tracking. *Sensors* 9(4):3033–3055
- Van der Spek SC, Van Langelaar CM, Kickert CC (2013) Evidence-based design: satellite positioning studies of city centre user groups. *Proc ICE-Urban Design Plann* 166(4):206–216
- Yang JS, Kang SP, Chon KS (2005) The map matching algorithm of GPS data with relatively long polling time intervals. *J East Asia Soc Transp Stud* 6:2561–2573

# Estimating the Biasing Effect of Behavioural Patterns on Mobile Fitness App Data by Density-Based Clustering

Cecilia Bergman and Juha Oksanen

**Abstract** Crowd-sourced data of high spatial and temporal resolution can provide a new basis for mobility analyses given that its various types of biases distorting the results are identified and adequately handled. In this paper, trajectory patterns that can affect the validity of mobile fitness app data are examined by means of cycling trajectories ( $n = 50,524$ ) from the Helsinki Metropolitan Area, in Finland. In addition to mass events and group journeys, we evaluated the biasing effect of routes that have been repeatedly recorded by the same application user. Based on the results, repeatedly recorded commuting routes may skew fitness application data more than group patterns. Many of the changes in the frequencies and length distributions at different temporal granularities before and after extracting the ‘bias patterns’ were statistically significant. Also the skewed distribution of tracks among users (i.e. contribution inequality) became more even. The biases induced by behavioural patterns ought to be considered when evaluating the validity of fitness app data in analyses of general mobility behaviour and when designing value-added applications based on the data. Considering the trade-off between privacy and data accuracy regarding dissemination of sensitive crowd-sourced movement data, the findings emphasise the importance of preserving the possibility to detect individual-level phenomena in order to produce valid analysis results.

**Keywords** Crowdsourcing · Big Data · Fitness apps · Cycling · Trajectory similarity · Clustering

---

C. Bergman (✉) · J. Oksanen  
Department of Geoinformatics and Cartography, Finnish Geospatial Research Institute,  
National Land Survey of Finland, Kirkkonummi, Finland  
e-mail: cecilia.bergman@nls.fi

J. Oksanen  
e-mail: juha.oksanen@nls.fi



# 1 Introduction

Crowdsourcing is expected not only to provide insights into phenomena for which no official data is being collected, but also to augment and supplant conventional data sources (Pucci et al. 2015; Tam and Clarke 2015). Consequently, addressing the inherent biases that affect the quality and value of the new datasets has become crucial (Shearmur 2015). Mobile sports tracking, or fitness application trajectories, are an example of crowd-sourced movement data, which has been proliferating in recent years due to advances in positioning technology. Considering especially the emerging interest in using fitness app data in an urban planning context the representativeness of the data remains an important issue. For example, with respect to cycling, the main focus in planning is on cycling as transportation, while tracking as an activity is typically attached to fitness-oriented cycling (Griffin and Jiao 2015). In addition to the representativeness of utilitarian cycling, concerns have been expressed about the digital divide and how it, together with other bias-causing factors, can reflect both spatial and socio-economic biases related to, e.g. gender, age and wealth (Bell et al. 2014; Griffin and Jiao 2015; Oksanen et al. 2015; Romanillos et al. 2015).

A limited amount of attention has been paid on other behavioural patterns that might introduce biases into mobile fitness app data. Previous studies have shown that the extremely uneven distribution of recorded workouts between users can locally distort heat maps that show the popular places for engaging in sports (Oksanen et al. 2015) and thereby affect routing (Bergman and Oksanen 2016). Furthermore, the interactive filtering of a heat map proposed by Sainio et al. (2015) enabled us to visually recognise mass events, such as competitions that were unique with respect to certain characteristics, for instance length. Identifying such patterns is important in order to understand the data, derive valid conclusions regarding cycling behaviour and potentially exclude them from further analysis. The novel value-added services that are being developed by aggregating the data for insights into the collective behaviour of the application users, such as the popularity of different routes, are affected by the bias patterns as well.

The objective of this paper was to automatically recognise the pre-identified patterns that can bias the dataset and thus affect its utility, and to assess the effect of the patterns on certain general characteristics of the dataset. The investigated patterns included both mass events along with other group journeys and routes that have been recorded by the same user more than once. While referring to the aforementioned categories, terms ‘group patterns’ and ‘individual patterns’, respectively, will be used in the rest of the paper. An efficient method based on the idea of progressive clustering (Andrienko et al. 2007; Rinzivillo et al. 2008) was used to identify similar trajectories. Although we concentrated on cycling data recorded by users of the Sports Tracker mobile application, the methods are also applicable to data from other fitness applications as well as other sports that do not need to be restricted to a network. The rest of the paper is organised as follows. In Sect. 2, we provide a concise overview of existing work on trajectory similarity analysis and

present the method employed to search for patterns in the dataset. Section 3 introduces the dataset and provides novel insights about its frequency and length distributions at different temporal levels of detail. In Sect. 4, we present the results and evaluate the importance of extracting the identified patterns from mobile sports tracking data. Finally, we discuss the meaning of the results with respect to the utility of mobile sports tracking data and end the paper by offering some conclusions.

## 2 Trajectory Similarity Analysis

### 2.1 Previous Research

Discovering new knowledge based on the trajectories of moving objects and their semantic enrichment is crucial for smart mobility applications and an important topic in GIScience. Various patterns, such as stops and moves (Spaccapietra et al. 2008), periodic (Cao et al. 2007) and relative motion, that is, individual–group dynamics (Laube et al. 2005), frequent trajectories (Savage et al. 2010), outliers and causal interactions (Liu et al. 2011), and POIs (Liu and Seah 2015) have been mined from GPS trajectories. In addition to human mobility, patterns of animal movement, e.g. migration (Damiani et al. 2015), as well as natural processes, e.g. hurricanes (Dodge et al. 2011), have also been investigated. In general, the methods used for trajectory mining can be divided into various categories depending on, for example, the nature of the movement, which can be either free or network restricted, and whether the aim is to extract patterns within single trajectories or from a set of trajectories.

Similarity analysis is an intrinsic part of many methods of pattern extraction and data mining, such as clustering, i.e. the grouping of trajectories based on their similarity to one another. Similarity can either be partial (Lee et al. 2007) or complete, as in this study, where an entire trajectory is compared to other trajectories. Although most studies have focused on geometric similarity, the importance of temporal and attribute proximity has also been stressed in many contexts (e.g. Nanni and Pedreschi 2006). Moreover, in certain situations it may be more appropriate to understand similarity as a similar variation in movement parameter profiles (Dodge et al. 2012) or in the geographic context (Buchin et al. 2014). When the exact route is irrelevant, trajectories can also be grouped based on their origin and destination (Andrienko et al. 2007). All in all, determining an appropriate measure of similarity is very much dependent on the context (Gudmundsson et al. 2012).

Several methods, which have their roots in time-series analysis, have been developed to measure the similarity of trajectories. Euclidean Distance (ED)-based approaches are popular because they are simple and efficient to calculate, and provide a metric similarity measure. Other well-known, more complex methods, such as Dynamic Time Warping (DTW), Longest Common Subsequence (LCSS)

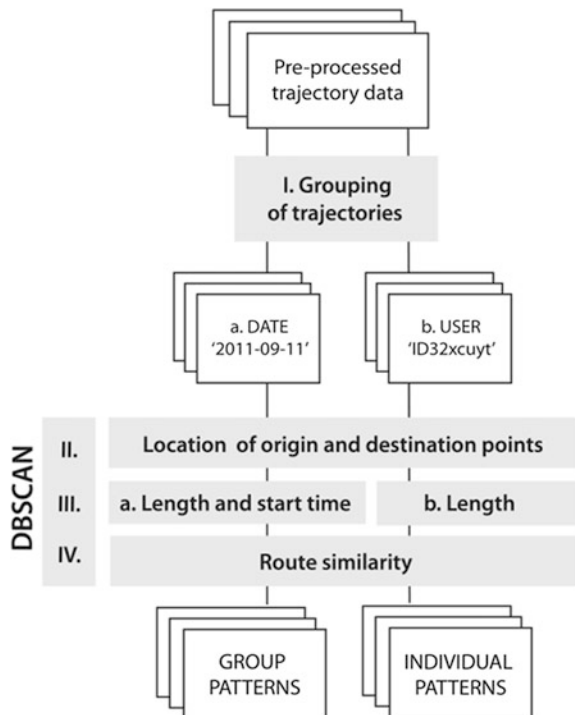
and Edit Distance with Real Penalty (ERP), are however more robust with respect to noise, outliers and time-shifts (for an overview, see e.g. Dodge 2011; Gudmundsson et al. 2012; Long and Nelson 2013).

## 2.2 Discovering Similarity Patterns in Mobile Fitness App Data

In this study, we were interested in recognising GPS trajectories that are co-located spatially and, in certain cases, also temporally. We treated each recorded track as one discrete trajectory, which we then compared to other trajectories as a whole. Although in this study we only required that similar trajectories approximately follow the same path, the analysis needs will ultimately determine which trajectories should be considered similar.

The workflow of discovering similarity patterns is presented in Fig. 1. The advantages of progressive clustering include a simple distance function for each step and restricting the computationally more expensive step (IV) to a potentially interesting subset of the data (Andrienko et al. 2007; Rinzivillo et al. 2008). Renso and Trasarti (2013) further point out that such iterative mining, where at each step a

**Fig. 1** The workflow, which is based on progressive clustering, used in the present study. After pre-processing, trajectories were grouped by date (*Ia*) and by user (*Ib*). Trajectories in each group were then clustered based on the origin and destination points such that both points were required to belong to the same cluster (*II*). The identified clusters were further clustered by route length and start time (*IIIa*), or just route length (*IIIb*). In the final step, clustering was performed based on route similarity (*IV*)



new constraint is introduced to remove uninteresting data, may enhance our understanding of the data. DBSCAN-algorithm (Density-Based Spatial Clustering Applications with Noise; Ester et al. 1996) of scikit-learn (Pedregosa et al. 2011) allowed us to form clusters of arbitrary shape without knowing their number in advance, while requiring only two parameters: the radius of the neighbourhood of an object,  $Eps$ , and the minimum number of objects in the neighbourhood,  $MinPts$ .

In the first step, the pre-processed trajectories (see Sect. 3) were partitioned based on both the user (Ib) and the date (Ia); hereby, it was assumed that group patterns take place during the same day. In step II, the trajectories in each group were clustered by DBSCAN based on the spatial similarity of their origin and destination points. Clusters of similar trajectories had origin and destination points that coincided spatially so that they were density reachable from each other. In step III, time was constrained depending on the pattern. In mass events, the recording of tracks should start at approximately the same time or within a specified time window, since in large events participants may be divided into different starting blocks. Group journeys were not handled differently, although they could be separated from events as they proceed simultaneously throughout the entire journey (IIIa). No time constraint was used with the user-specific trajectories (IIIb). Following the specifications proposed by Liu et al. (2012), we calculated  $Eps$  by dividing the median length of the cluster’s trajectories by a value that determined how large a divergence from the median length we would accept. This was a robust method considering the fact that, due to GPS error, the accepted divergence should be higher for routes of 100 km than for those of 20 km. Finally, the trajectories in each identified group were clustered based on the similarity of their routes in step IV. The average distance between trajectories was used as a measure of (dis)similarity (see e.g. Nanni and Pedrexchi 2006) as follows.

$$D(T_1, T_2) = \left( \sum_{i=1}^{N_{fix}} d(T_1(i), T_2(i)) \right) / N_{fix} \tag{1}$$

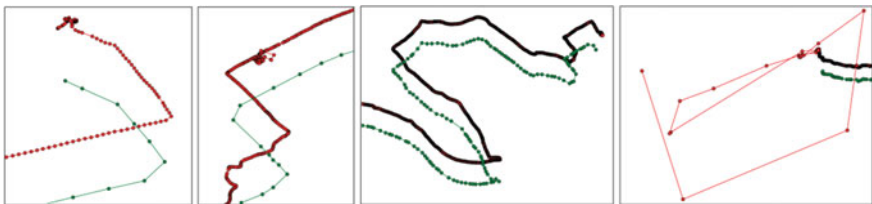
where  $T_1$  and  $T_2$  are two trajectories for which we generated a fixed number of ordered equidistant points ( $N_{fix}$ ) along the path and  $d()$  is the Euclidean distance between the  $i$ th generated points along the two trajectory lines. Apart from being computable in linear time and also applicable with respect to trajectories of complex shape, this measure provided a metric dissimilarity value that was easy to understand.

### 3 Dataset

Mobile fitness applications can be situated between social media, emphasising their role as social networks, and, in the words of Kitchin (2014), ‘sousveillance’, by which he refers to their function as a type of self-monitoring mechanism. Typically,

application users can themselves define whether the workouts they record are private, i.e. only the user can view and access the track data, or public, i.e. everybody can view the track and its additional metadata. In this study, we used public track data recorded by the users of the Sports Tracker mobile application (<http://www.sports-tracker.com>). The users of Sports Tracker can decide, on a track-wise basis, whether or not they want to allow a track to be publicly viewable.

The dataset used in this study consisted of workouts recorded between April 2010 and November 2012 in the Helsinki Metropolitan Area, in southern Finland. We obtained the data from Sports Tracking Technologies Ltd. (currently Amer Sports Digital Services Ltd.), which pseudonymised the user identifiers of the tracks before data delivery. The original dataset included 62,843 GPS trajectories, which consisted of time-ordered location points recorded with an interval of one second. After removing trajectories that either had no timestamp or that had lasted for less than two minutes, we were left with 50,524 trajectories recorded by 3,723 users. Similar to many other online communities that are based on volunteered collaboration (Yang et al. 2016), the dataset was characterised by contribution (or participation) inequality; here half of the tracks were recorded by 5 % of the users. Further pre-processing of the trajectories was necessary for an efficient similarity analysis. In addition to the positional error intrinsic in GPS data, it was apparent that stops during which time the recording had continued affected the length of the trajectory. Based on experiments, stops were effectively removed by the following process which yet preserved the details of the road network (Fig. 2). First, we calculated speed and acceleration rate between all consecutive points and removed points where the speed was below 2 m/s or the acceleration rate was more than  $1 \text{ m/s}^2$ , which indicated jumping of the GPS signal. Second, a five-point median filter was used to further remove outliers and reduce the size of the dataset; the five points were replaced with a median point. Finally, the total magnitude of the heading change was calculated for each five ( $n$ ) point window; when the total change in degrees exceeded  $60^\circ(n-1)$ , the five consecutive points were removed. For instance, Zhang et al. (2013) have previously noted the vitality of heading change when using GPS data to identify stops.

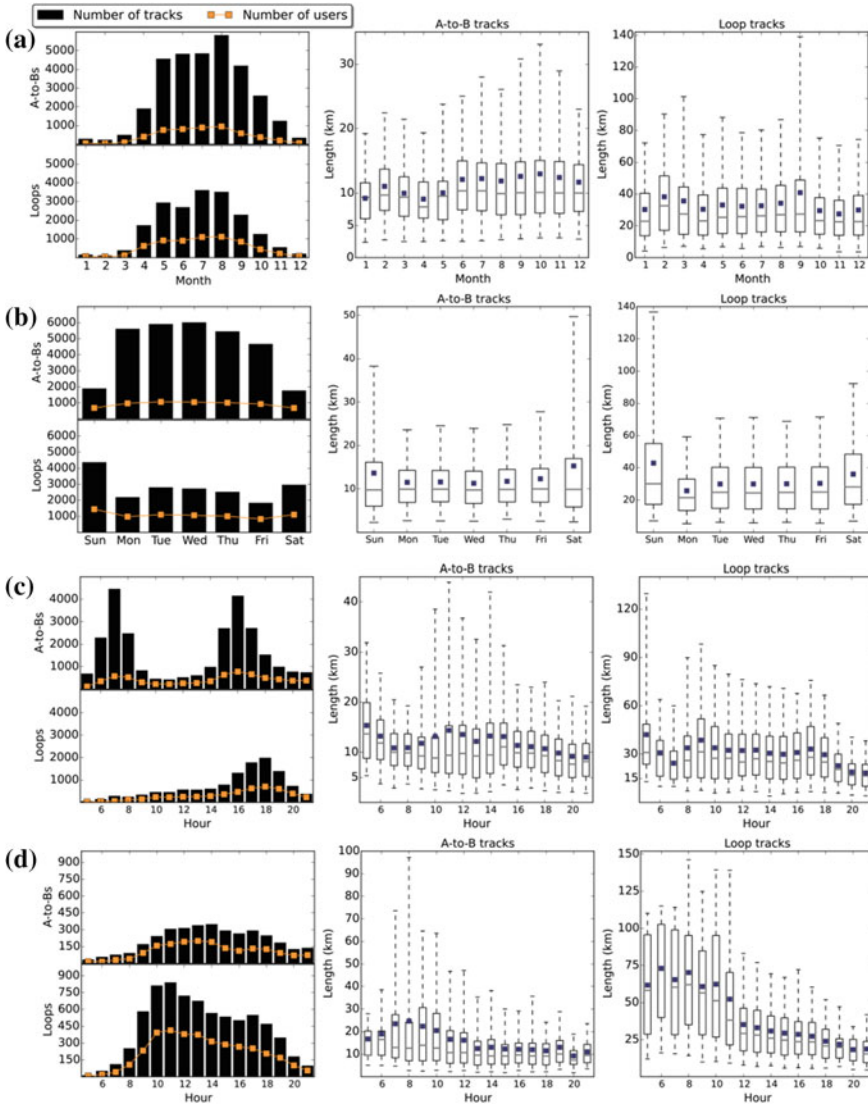


**Fig. 2** Examples of noisy trajectories before (*red*) and after (*green*) pre-processing. An offset of approx. 300 m has been added to one of the trajectories for visualisation purposes

To enhance understanding of the data and illustrate the results, the dataset was disaggregated by time as well as based on the geometric shape of the trajectories. If the ratio of the total length to the straight line distance between the origin and end points of the trajectory exceeded four, then the track was classified as a 'loop track'; otherwise, it was classified as an 'A-to-B track'. With a clear majority of the trajectories, the threshold ratio that determined the class of each track was either very large (loop tracks) or small (A-to-B track); for example, with only 3.3 % of the tracks, the ratio was in the range of three to eight. With this coarse definition, 61.8 % of the tracks belonged to the category of A-to-B tracks and 38.2 % of them were loop tracks. A-to-B tracks were on average 12.0 km long (median 9.9 km), while loop tracks were on average 33.4 km (25.7 km) long.

The frequency of tracked workouts varied between the summer and winter months, with the workouts being at their lowest points from December to March. In both cases, monthly length distributions were rather homogeneous, especially in the summer months, with only September standing out from the loop tracks (Fig. 3a). Tracking was more popular during weekdays (Mon–Fri) than at weekends (Sat–Sun) (Fig. 3b). Loop tracks dominated at weekends, especially on Sunday when the recorded tracks were also longer than on the other days. Clear peaks in the number of tracks could be observed in the morning and in the afternoon during weekdays (Fig. 3c). Similarly, at weekends loop tracks showed two—although not as obvious—peaks: one before noon and the other after 17 o'clock (Fig. 3d). At weekends, recorded tracks were longer before noon than in the afternoon and evening.

We also considered redefining the classes as utilitarian (A-to-B tracks) and recreational (loop tracks) trips. However, a purely geometry-based reclassification of the tracks as utilitarian and recreational trips would by no means be perfect. Not all of the tracks that started at one point and ended at another point were utilitarian by nature; a leisure rider can, for example, ride to a particular destination, stop the tracking there and then return back by train. Also, not all tracks that started and ended at the same point should be classified as recreational tours. A cyclist can ride to, e.g. a supermarket or football pitch, put the tracking on hold and then continue recording the same track after shopping or training, respectively. A straightforward solution would be to divide the track into two tracks if the original track includes a break that exceeds certain duration. However, recreational tours can also include breaks. All of these examples were identified from the dataset and demonstrate the heterogeneity of mobile fitness app data. We therefore use the terms A-to-B tracks and loop tracks in this paper. The classification does not play a role in the clustering method; it is only used to illustrate the results.



**Fig. 3** The days and length distributions of A-to-B and loop tracks for the different **a** months, **b** days, and **c** hours Mon–Fri and **d** Sat–Sun based on the time when the tracking was started. In the *boxplots*, the *lower edge of the box* represents the 25th percentile and the *upper edge* the 75th percentile. The median is represented with a *line* and the mean with a *box*. The whiskers extend to the 5th and 95th percentiles



## 4 Results and Analysis

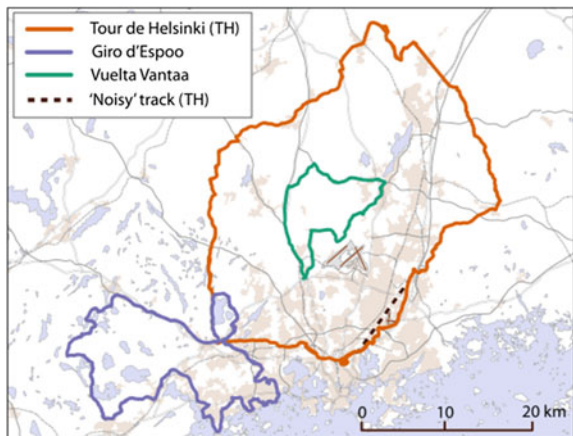
### 4.1 Performance Evaluation

The results for density-based clustering depend on the chosen parameter values. In all steps (see Fig. 1), the *MinPts* was two. In step II, the neighbourhood should be large enough to take into account both GPS inaccuracy and the variation in terms of where the tracking is turned on or off; for example, participants may start recording mass events already before the start line. Moreover, some people prefer to start and stop tracking further away from their home to preserve their location privacy. The *Eps* was 500 m in step II, whereas in step III we calculated it based on the data, as explained in Sect. 2.2. Here, we accepted a 10 % divergence from the median length. Regarding the required proximity of trajectories in time, we defined the *Eps* as 30 min (step IIIa). All parameters in steps II and III were intentionally rather large yet such that they efficiently restrained the number of candidates that needed to be compared in step IV.

Euclidean distance as a measure of similarity is not robust with respect to the outliers and noise in the data, some of which could be ‘cleaned’ in pre-processing. Based on our comparison of the clusters after step III and step IV, we left out only one of the 79 trajectories that represented the Tour de Helsinki 2012—the largest identified mass event—from the final cluster due to ‘noise’ (Fig. 4). We used the same *Eps* in step IV as in step II, that is, 500 m, but in order to test the sensitivity of the results to the *Eps*, we repeated the calculations also with six other distances ranging from 50 to 2,000 m.

Table 1 shows the effects of each clustering step on the identified clusters. Both the standard deviation of the length of trajectories in a cluster and the number of clusters significantly decreased in step III when the trajectories were grouped based on their length and starting time (Table 1a). When we opted not to consider temporal similarity during the clustering process, the largest reduction in the number of

**Fig. 4** Identified mass events of cycling in the Helsinki Metropolitan Area. The ‘noisy’ track was likely because of tracking being paused approx. 20 km before the end line. The map contains data from the Topographic database by the National Land Survey of Finland 01/2016





**Table 1** The effect of density-based clustering steps on the results in cases where the aim is to recognise **a** group patterns, and **b** individual patterns represented with cluster-wise statistics

		(a) Group patterns			(b) Individual patterns		
		II	III	IV	II	III	IV
Number of clusters		635	172	170	4,293	4,180	3,694
Number of tracks in clusters		3,728	1,003	567	38,329	32,062	25,776
Cluster size	Average	2.4	3.3	3.3	8.9	7.6	7.0
	Median	2.0	2.0	2.0	3.0	3.0	3.0
	Std.	4.8	8.5	8.5	18.2	16.7	13.9
Cluster-wise length (km)	Average	28.0	37.4	38.5	19.1	19.7	16.9
	Median	22.4	24.6	24.8	13.0	13.3	11.9
Cluster-wise std. of length (km)	Average	7.0	0.13	0.13	4.2	0.66	0.27
	Median	2.4	0.08	0.08	0.85	0.26	0.19
Std. of length/avg. length (%)	Average	24.3	0.5	0.5	18.4	3.4	2.4
	Median	16.7	0.3	0.3	8.9	2.3	1.6

clusters occurred only in the last step when we divided the identified groups of similar length into smaller groups based on their route (Table 1b). Nevertheless, the mean cluster-wise standard deviation of length decreased already in the third step, which was an indication of recreational loop tracks of varying length. Most typically though, a single user had recorded workouts of approximately equal length but varying routes. Thus, the number of tracks that appeared in clusters decreased in the last step (Table 1b), as did the average length of the trajectories in a cluster, indicating that longer rides in particular differed in their routes.

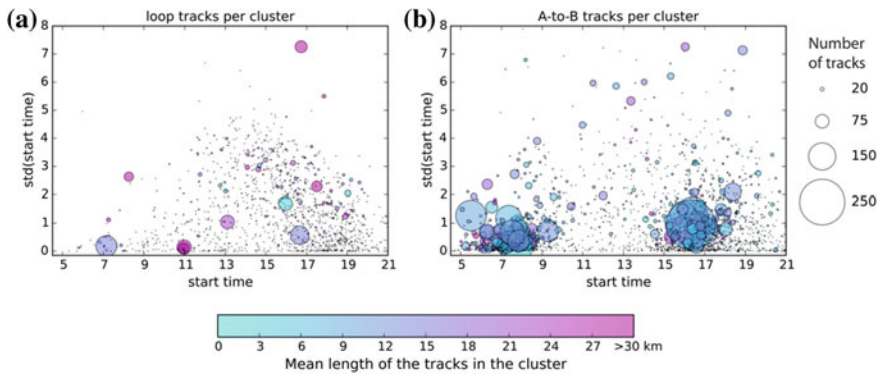
## 4.2 Identified Clusters

The results confirmed the existence of patterns identified in previous studies. We discovered nearly 4,000 clusters, 48 % of which included only two trajectories. The number of individual patterns (Table 2b, d) was more than 45 times greater than that of group patterns (Table 2a, c). Nearly all of the group journeys consisted of a group of two cyclists, with the exception of mass events (Table 2a). Also, the majority of the repeatedly recorded loop tracks included only two occurrences (Table 2b). In all cases, except with clusters consisting of only two trajectories (Table 2c), the size distribution was positively skewed. Figure 5 shows the standard deviation of the start time as a function of the temporal occurrence of clusters. The largest clusters seem to represent commuting, as they typically took place during the peak hours and the deviation within the clusters was small. All three mass events started at 11 o'clock and can be identified by their minimal standard deviation (Fig. 5a). Notice that Fig. 5 does not separate clusters of group patterns and individual patterns.

**Table 2** Cluster characteristics

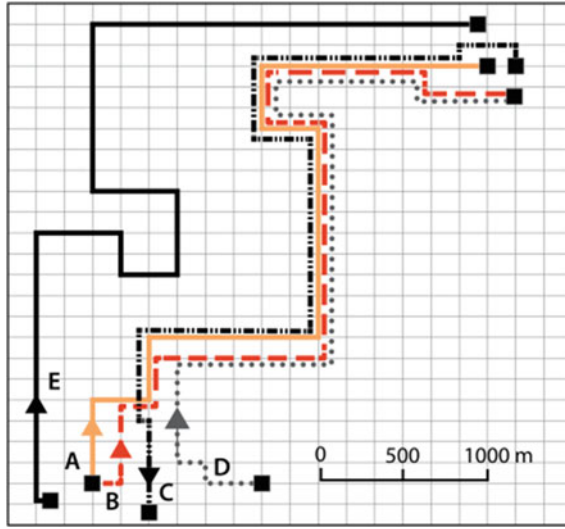
		Loop tracks		A-to-B tracks	
		(a) Group patterns	(b) Individual patterns	(c) Group patterns	(d) Individual patterns
Number of clusters (% of clusters with two members)		119 (92)	1,204 (59)	51 (100)	2,506 (39)
Number of tracks in clusters		465	4,665	102	21,111
Cluster size	Average	3.9	3.8	2.0	8.4
	Median	2.0	2.0	2.0	3.0
	Std.	10.2	6.7	0.0	16.0
Cluster-wise length (km)	Average	47.4	28.6	17.7	11.2
	Median	33.5	22.1	10.0	9.5
Cluster-wise std of length (m)	Average	157	389	71.1	207
	Median	92.3	291	42.1	161

The total number of clusters is different than in Table 1b (IV), because in a few clusters there were both loop and A-to-B tracks. This was partly because of tracks at the Velodrome which would require different pre-processing than tracks on the road network



**Fig. 5** Identified clusters of **a** loop tracks and **b** A-to-B tracks. The x-axis denotes the average start time and the y-axis the standard deviation of start times in the cluster

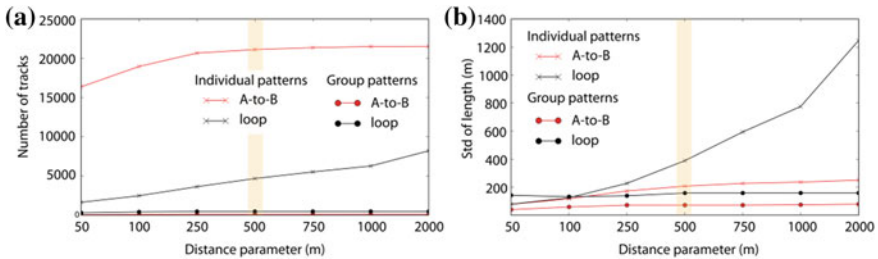
The standard deviation of the trajectory lengths in a cluster was smaller for group patterns (Table 2a, c) than individual patterns (Table 2b, d). Because of time constraints, the clusters of group patterns typically represented similar routes, and in nine cases the recorded tracks were exactly identical also from a temporal standpoint. This can mean that the route was recorded by one user who, for instance, shared the GPS track with another user. In all cases, the standard deviation of the length was small, which indicated the validity of the clustering method being used (Tables 1 and 2). However, it is important to recognise that the routes in a particular cluster were not necessarily identical, but could partly follow different paths as long as their mean distance was within the specified threshold, origin and destination



**Fig. 6** Schematic illustration of the principles of calculating track similarity. Tracks *A* and *B* belong to the same cluster. Tracks *C* and *D* do not fall into the cluster because their origin points are more than 500 m apart from those of tracks *A* and *B*. Notice that with track *C* the movement is in the opposite direction. Track *E* does not belong to the cluster due to its dissimilar length and route

points were co-located, and route lengths were similar (Fig. 6). As Fig. 5 shows, the identified clusters were also mostly homogeneous from a temporal standpoint ( $std < 2$ ).

Clustering of individual patterns was more sensitive to the distance parameter used in calculating route similarity (step IV) than group patterns. Especially with repeatedly tracked loop routes both the number of tracks in clusters (Fig. 7a) and the mean cluster-wise standard deviation of length (Fig. 7b) increased along with a greater distance parameter, whereas in other cases they remained more stable.



**Fig. 7** **a** Number of tracks in clusters and **b** average cluster-wise standard deviation of length with different distance parameters (*Eps*) in step IV. The *Eps* used in the present study (500 m) is marked with *brown* stripe

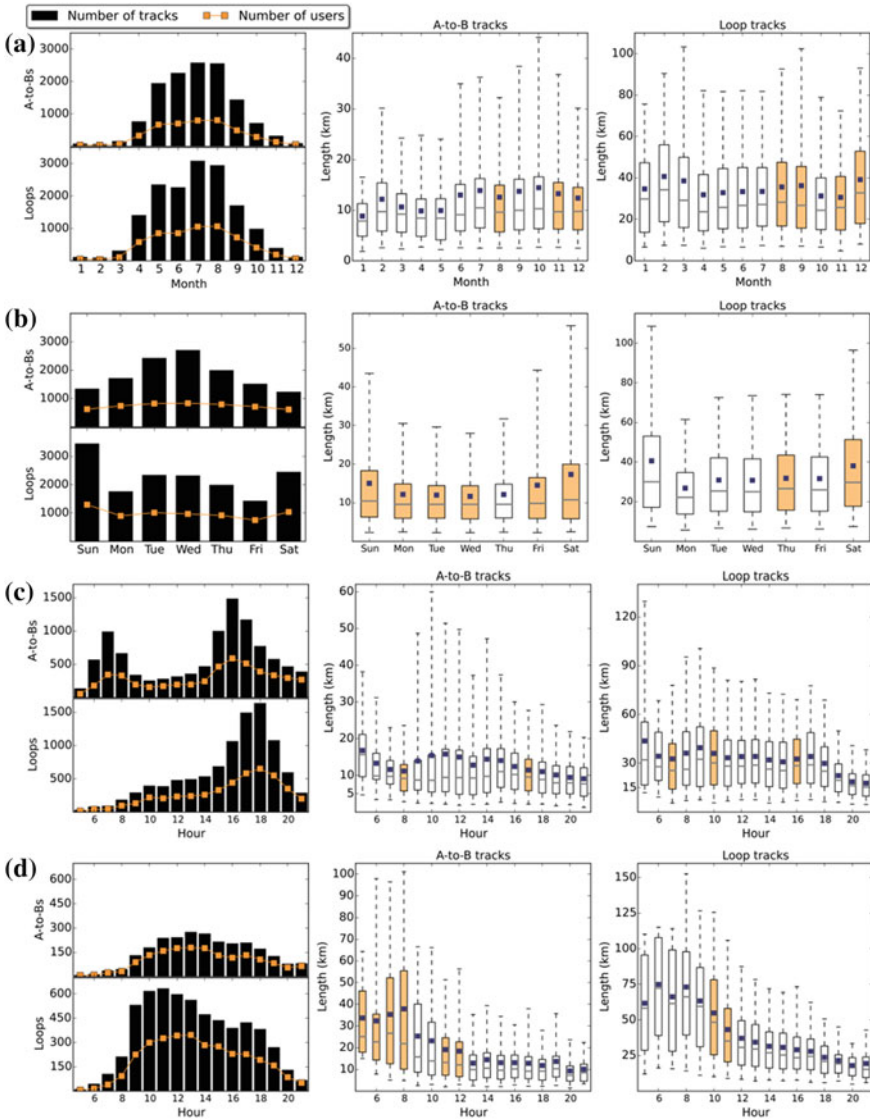
### 4.3 *Effects on the Dataset*

When considering further uses of the data for analyses, a crucial question must be addressed: What are the bias patterns and how should they be handled in data processing? If the aim is to provide insights into the popularity of cycling along the network, then we can completely exclude mass events where cyclists cannot influence the route, which may not even be especially popular except for during the event. Group journeys and commutes as well as other similar paths repetitively recorded by one person might be, for example, replaced by one representative trajectory. However, if we want to be able to find the most popular paths at different times of the day or during different months, it becomes problematic to decide which tracks should be retained. On the other hand, if the aim is to assess the effects of the cycling infrastructure and other factors on route choices, the spatial trajectories become important. Even small variations in routes can be important if we are specifically interested in, for instance, how construction work in an area has affected cyclists' behaviour.

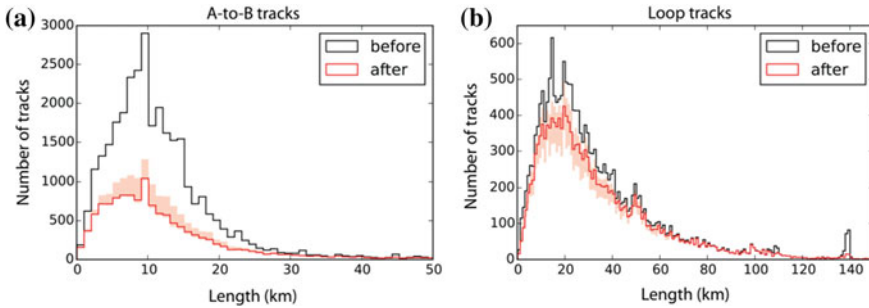
To evaluate the effect of the identified bias patterns, we replaced each identified cluster with an attribute vector that represented the median characteristics (length, speed, day, month and start time) of the cluster. Several methods have been developed to identify the representative trajectory geometry of a set of trajectories (e.g. Buchin et al. 2013; Etienne et al. 2015), but we did not investigate the spatial effects in this particular study.

After excluding the bias patterns, the number of trajectories was 28,543 and the share of loop tracks had increased to 54.9 % (compared to 38.2 % previously). The average length of the loop tracks increased to 33.9 km (compared to 33.4 km previously), whereas the median length increased to 26.6 km (compared to 25.7 km previously). The mean length of the A-to-B tracks was also higher, 13.1 km (compared to 12.0 km previously), whereas the median length was close to that of the original dataset, 9.7 km (compared to 9.9 km previously). Moreover, the following changes that are presumably indicators of Tour de Helsinki and other mass events were noticed in the distributions of loop tracks: a narrower length distribution in September (Fig. 8a); on Sunday (Fig. 8b); and at 10–11 o'clock at weekends (Fig. 8d). The non-parametric Mann-Whitney-Wilcoxon test confirmed the significance of these changes. Even though most statistically significant changes (shaded boxes in Fig. 8) were characterized by a small population size, for example, changes in the morning and in the afternoon were significant not only during weekdays, but, interestingly, also at weekends. As can be seen, the morning peak reduced more than the evening peak at weekdays, which can indicate that especially in the morning people's route choices are less diverse (Fig. 8c).

Figure 9 shows length histograms before and after replacing the identified clusters with their median attributes. Before the replacement, the loop tracks showed a peak every ten kilometres, which was less evident afterwards. Although many peaks indicated bias patterns, such as those representing the mass events Tour de Helsinki (140 km), Giro d'Espoo (111 km) and Vuelta Vantaa (79 km in 2012),



**Fig. 8** The frequencies and length distributions of A-to-B and loop tracks during different **a** months, **b** days and **c** hours Mon–Fri and **d** Sat–Sun based on the time when tracking was started after extracting the bias patterns. A shaded box means that the change in the distribution compared to that in Fig. 3 was statistically significant at a significance level of 0.05 based on Mann-Whitney-Wilcoxon test



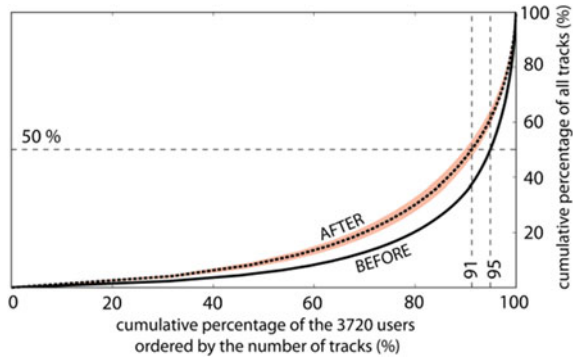
**Fig. 9** The length distributions of **a** A-to-B tracks and **b** loop tracks before and after extracting the bias patterns. The bin in both figures is 1 km. The shaded region represents the variation with *Eps* (step IV) between 100–2000 m

several peaks were preserved, e.g. at 20, 50 and 100 km (Fig. 9b). Based on the Kolmogorov-Smirnov (K-S) test, the change in distributions before and after extracting the bias patterns was significant ( $p < 0.01$ ) for the A-to-B tracks (statistic = 0.055, p-value  $4.12e-24$ ) as well as for the loop tracks (statistic = 0.019, p-value 0.0051). With parameters from 50 to 250 m the change was statistically significant only for the A-to-B tracks at a significance level of 0.05. It turned out that an average distance of 50 m was too small to identify the tracks of mass events, and hence all further sensitivity analyses were started from 100 m.

We used the chi-square test to evaluate the changes in frequencies of the monthly, weekly and diurnal distributions (Figs. 3 and 8). Because the sample sizes were different, we scaled the original binned counts to be comparable with those after the extraction. The p-values were small ( $p < 1e-100$ ) with all parameter values in all other cases except for the diurnal distribution of weekends (Figs. 3d and 8d). However, also here in the case of both A-to-B tracks and loop tracks we could reject the null hypotheses that the two binned data (before and after) stem from a common distribution pattern at a significance level of 0.05 with all parameters greater than 50 m, and at a significance level of 0.001 with all parameter values greater or equal than 250 m. The change of loop tracks at weekends was least significant. Although the changes in the frequencies of track counts during weekdays were significant, the bimodal and unimodal distributions of A-to-B tracks and loop tracks, respectively, were preserved.

Replacing clusters of individual patterns by one ‘track’ affected also on the contribution inequality. The share of users responsible of half of the recorded tracks increased to almost 9 %. Furthermore, the Gini coefficient which measures the inequality visualized with the Lorenz curve (see Yang et al. 2016) in Fig. 10 decreased from 0.75 to 0.65.

**Fig. 10** The distribution of tracks between users before and after represented by Lorenz curves. The *shaded* region represents the variation with *Eps* (step IV) between 100–2000 m



## 5 Discussion

To the best of our knowledge, no existing study has evaluated in detail the latent similarity patterns that may affect the validity of mobile sports tracking data. Beecham and Wood (2014) extracted group journeys from bike sharing data, providing insights on their spatial, temporal and demographic characteristics; however, their data was rather different with respect to the role of the bike-sharing system and the trip types, and it consisted—in spatial terms—only of the origin and destination points. Based on their findings, 3 % of bike-sharing trips were made in groups of two or more people, whereas in the present study only 1 % of mobile fitness app data tracks were identified as group journeys, including mass events. Furthermore, Andrienko et al. (2013) developed methods to analyse the internal trajectory patterns of individual events and commuter patterns. Event identification for its part has been a popular topic in terms of mining social media data, such as georeferenced images (Sun and Fan 2014) and microblog streams (Sakaki et al. 2010). Likewise, mobile phone data have been used to detect social events characterised by unusual activity (Traag et al. 2011).

Many previous studies have concluded that, besides recreational cycling trips, mobile fitness apps are used to record commuting journeys (Bell et al. 2014; Griffin and Jiao 2015; Oksanen et al. 2015). Also the contribution inequality has been reported before but not investigated in greater detail in the context of the present study. It turned out that the extent of bias induced by individual patterns, and specifically commuters, was greater than the bias caused by mass events and other group journeys. Hence, after excluding the bias patterns, loop tracks constituted the majority of the data. Visually, the frequency distributions before and after resembled each other in general, but statistically there were significant differences. Mass events can have a significant effect on spatial as well as length patterns, and therefore, it is often advisable to exclude them from the data. The effect of mass events was further enhanced by their temporal coincidence. Cyclists actively tracking their commutes can, depending on the analysis purpose, be handled via different means. For example, Oksanen et al. (2015) discussed the concept of



popularity and suggested the number of users and number of tracks revised by a diversity index as alternatives to the number of recorded tracks. Their study demonstrated that it would be feasible to calibrate the biased Sports Tracker dataset with adequate reference data, such as official bicycle counts, for a richer view of cycling volumes along the network of routes in the Helsinki area. However, when no reference data with sufficient spatial coverage is available, or when the analysis requires information about the routes, then we should consider extracting the latent patterns inducing biases in the dataset.

Because the method used was restricted to whole matching, that is, to comparing the similarity of trajectories as a whole, certain patterns that would also require discovering similar subtrajectories were excluded. These included journeys that were only partially made in groups as well as participants in mass events who did not finish the tour. Another limitation of this study is the use of only one dataset from a single region. That a relatively small group of users was causing biases in the dataset raises questions related to the generalizability of the results. The biasing effect of mass events can vary between mobile fitness apps because some applications are more oriented to competitive athletes than others. This may also be reflected in the popularity of the application among commuters. However, we argue that all similar types of datasets can be biased by commuters who repetitively track their route to work. Furthermore, similar uneven distributions of workouts between users have been reported in studies related to other fitness applications and areas (Ferrari and Mamei 2013; Vickey and Breslin 2012). Our findings suggest that this inequality is to large extent related to commuting. Whether the dataset covers all tracks or only those that are publically viewable does not necessarily make a big difference.

## 6 Conclusions

The aim of the study was to estimate the biasing effect of mass events and other group journeys as well as routes repeatedly recorded by the same user on mobile fitness app data. Density-based clustering was used in multiple steps to efficiently extract the afore-mentioned behavioural patterns. Together, the group and individual patterns had a statistically significant effect on the characteristics of the dataset, especially with respect to its frequencies and length distributions at different temporal granularities.

Excluding individual patterns resulted in a more even distribution of tracks between users. Based on the results, contribution inequality was strongly yet not exclusively related to commuting behaviour. Considering the trade-off between privacy and data accuracy regarding dissemination of sensitive movement data, the results highlight the importance of recognising individual-level phenomena. If tracks recorded by the same user cannot be associated to each other due to the information loss caused by anonymization, significant biasing behavioural patterns might stay hidden.



There is obviously a need for analysis methods that are robust to latent bias patterns. It should be noted, though, that extracting these patterns does not release us from the necessity of using several data sources to assess the validity of such datasets and obtain insights that are representative and useful. Although the results clearly show that, for example, contribution inequality can indicate the prevalence of biases and therefore should not be ignored, a more thorough assessment of their generalizability would require inspection of datasets from a wider area, both geographically and application-wise.

**Acknowledgments** We gratefully thank Sports Tracking Technologies Ltd. (currently Amer Sports Digital Services Ltd.) for providing us the workout tracking data. This work was carried out as a part of the projects MyGeoTrust and SUPRA (Revolution of Location-Based Services: Embedded data refinement in Service Processes from Massive Geospatial Datasets) funded by Tekes, the Finnish Funding Agency for Technology and Innovation (grants 40302/14 and 40261/12).

## References

- Andrienko G, Andrienko N, Wrobel S (2007) Visual analytics tools for analysis of movement data. *ACM SIGKDD Explor Newsl* 9(2):38–46
- Andrienko N, Andrienko G, Barrett L, Dostie M, Henzi P (2013) Space transformation for understanding group movement. *IEEE Trans Visual Comput Graphics* 19(12):2169–2178
- Beecham R, Wood J (2014) Characterising group-cycling journeys using interactive graphics. *Transp Res Part C: Emerg Technol* 47:1–13
- Bell B, Evans J, Mason C, Schliwa G (2014) Can cycling apps be used to inform smart infrastructure planning? <http://efr.pbworld.com/publications/default.aspx?id=80> Accessed at 7 Dec 2015
- Bergman C, Oksanen J (2016) Conflation of OSM and sports tracking data for automatic bicycle routing. *Trans in GIS*. doi:10.1111/tgis.12192
- Buchin K, Buchin M, van Kreveld M, Löffler M, Silveira RI (2013) Median trajectories. *Algorithmica* 66(3):595–614
- Buchin M, Dodge S, Speckmann B (2014) Similarity of trajectories taking into account geographic context. *J Spat Inform Sci* 9:101–124
- Cao H, Mamoulis N, Cheung DW (2007) Discovery of periodic patterns in spatiotemporal sequences. *IEEE Trans Knowl Data Eng* 19(4):453–467
- Damiani ML, Issa H, Fotino G, Heurich M, Cagnacci F (2015) Introducing ‘presence’ and ‘stationarity index’ to study partial migration patterns: an application of a spatio-temporal clustering technique. *Int J Geogr Inf Sci*. doi:10.1080/13658816.2015.1070267
- Dodge S (2011) Exploring movement using similarity analysis. Dissertation, University of Zürich
- Dodge S, Weibel R, Laube P (2011) Trajectory similarity analysis in movement parameter space. In: *Proceedings of GISRUK, Plymouth, UK, 27–29 April 2011*
- Dodge S, Laube P, Weibel R (2012) Movement similarity assessment using symbolic representation of trajectories. *Int J Geogr Inf Sci* 26(9):1563–1588
- Ester M, Kriegel HP, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd* 96(34):226–231
- Etienne L, Devogele T, Buchin M, McArdle G (2015) Trajectory Box Plot: a new pattern to summarize movements. *Int J Geogr Inf Sci*. doi:10.1080/13658816.2015.1081205
- Ferrari L, Mamei M (2013) Identifying and understanding urban sport areas using Nokia Sports Tracker. *Pervasive Mobile Comput* 9(5):616–628

- Griffin GP, Jiao J (2015) Where does bicycling for health happen? analysing volunteered geographic information through place and plexus. *J Transport Health* 2(2):238–247
- Gudmundsson J, Laube P, Wölle T (2012) Computational movement analysis. In: Kresse W, Danko DM (eds) *Handbook of geographic information*. Springer, Heidelberg, pp 725–741
- Halkidi M, Batistakis Y, Vazirgiannis M (2001) On clustering validation techniques. *J Intell Inform Syst* 17(2–3):107–145
- Kitchin R (2014) *The Data Revolution: Big Data, Open Data, data infrastructures and their consequences*. SAGE Publications Ltd
- Laube P, Imfeld S, Weibel R (2005) Discovering relative motion patterns in groups of moving point objects. *Int J Geogr Inf Sci* 19:639–668
- Lee JG, Han J, Whang KY (2007) Trajectory clustering: a partition-and-group framework. In: *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, Beijing, China, 11–14 June 2007
- Liu Y, Seah HS (2015) Points of interest recommendation from GPS trajectories. *Int J Geogr Inf Sci*. doi:[10.1080/13658816.2015.1005094](https://doi.org/10.1080/13658816.2015.1005094)
- Liu W, Zheng Y, Chawla S, Yuan J, Xing X (2011) Discovering spatio-temporal causal interactions in traffic data streams. In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, San Diego, CA, 21–24 Aug 2011
- Liu Q, Deng M, Shi Y, Wang J (2012) A density-based spatial clustering algorithm considering both spatial proximity and attribute similarity. *Comput Geosci* 46:296–309
- Long JA, Nelson TA (2013) A review of quantitative methods for movement data. *Int J Geogr Inf Sci* 27(2):1–27
- Nanni M, Pedreschi D (2006) Time-focused clustering of trajectories of moving objects. *J Intell Inform Syst* 27(3):267–289
- Oksanen J, Bergman C, Sainio J, Westerholm J (2015) Methods for deriving and calibrating privacy-preserving heat maps from mobile sports tracking application data. *J Transp Geogr* 48:135–144
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O et al (2011) Scikit-learn: machine learning in python. *J Mach Learn Res* 999888:2825–2830
- Pucci P, Manfredini F, Tagliolato P (2015) Mapping urban practices through mobile phone data. Springer International Publishing
- Renso C, Trasarti R (2013) Understanding human mobility using mobility data mining. In: Renso C, Spaccapietra S, Zimányi E (eds) *Mobility data*. Cambridge University Press, pp 127–148
- Rinzivillo S, Pedreschi D, Nanni M, Giannotti F, Andrienko N, Andrienko G (2008) Visually driven analysis of movement data by progressive clustering. *Inform Vis* 7(3–4):225–239
- Romanillos G, Austwick MZ, Ettema D, De Kruijf J (2015) Big data and cycling. *Transport Rev*. doi:[10.1080/01441647.2015.1084067](https://doi.org/10.1080/01441647.2015.1084067)
- Sainio J, Westerholm J, Oksanen J (2015) Generating heat maps of popular routes online from massive mobile sports tracking application data in milliseconds while respecting privacy. *ISPRS Int J Geo-Inf* 4(4):1813–1826
- 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*, Raleigh, NC, 26–30 April 2010
- Savage NS, Nishimura S, Chavez NE, Yan X (2010) Frequent trajectory mining on GPS data. In: *Proceedings of the 3rd International Workshop on Location and the Web—LocWeb'10*, Tokyo, Japan, 29 Nov 2010
- Shearmur R (2015) Dazzled by data: big data, the census and urban geography. *Urban Geogr* 36(7):965–968
- Spaccapietra S, Parent C, Damiani ML, de Macedo JA, Porto F, Vangenot C (2008) A conceptual view on trajectories. *Data Knowl Eng* 65(1):126–146
- Sun Y, Fan H (2014) Event identification from georeferenced images. In: Huerta J, Schade S, Granell C (eds) *connecting a digital europe through location and place*. lecture notes in geoinformation and cartography. Springer International Publishing, pp. 73–88

- Tam S-M, Clarke F (2015) Big data, official statistics and some initiatives by the Australian Bureau of statistics. *Int Stat Rev* 83(3):436–448
- Traag V, Browet A, Calabrese F, Morlot F (2011) Social event detection in massive mobile phone data using probabilistic location inference. In: privacy, security, risk and trust (PASSAT) and IEEE Third International Conference on Social Computing (SocialCom), pp. 625–628
- Vickey TA, Breslin JG (2012) A study on twitter usage for fitness self-reporting via mobile apps. AAAI Spring Symposium—Technical Report, SS-12-05, pp.65–70
- Yang A, Fan H, Jing N, Sun Y, Zipf A (2016) Temporal analysis on contribution inequality in OpenStreetMap: a comparative study for four countries. *ISPRS Int J Geo-Inform* 5(1):5
- Zhang L, Dalyot S, Sester M (2013) Travel-mode classification for optimizing vehicular travel route planning. In: Krisp JM (ed) *Progress in location-based services, Lecture notes in geoinformation and cartography*. Springer, Berlin Heidelberg, pp 277–295

# Enhancing Exploratory Analysis by Summarizing Spatiotemporal Events Across Multiple Levels of Detail

Ricardo Almeida Silva, João Moura Pires, Maribel Yasmina Santos  
and Nuno Datia

**Abstract** There are many spatiotemporal events with high levels of detail (LoDs) being collected in many phenomena. The LoD of analysis plays a crucial role in the user's perception of phenomena. From one LoD to another, some patterns can be easily perceived or different patterns may be detected. Standard practices work on a single LoD driven by the user in spite of the fact that there is no exclusive LoD to study a phenomenon. Our proposal aims to support users in carrying the inspection and comparison tasks of a phenomenon across multiple LoDs, without having to look at raw data, and to handle the spatiotemporal complexity. This paper presents a framework to build *abstracts* at different LoDs where five types of abstracts are proposed. The framework makes no assumption about the phenomenon, the analytical task and the phenomenon's LoDs. The SUITE's prototype implements the proposed framework allowing users to inspect abstracts across multiple LoDs simultaneously, helping to understand in what LoDs the phenomenon perception distinguishes itself or in what LoDs "interesting patterns" emerge.

**Keywords** Multiple levels of detail • Spatiotemporal data • Visual analytics

---

R.A. Silva (✉) · J.M. Pires  
NOVA LINCS, DI, FCT, Universidade NOVA de Lisboa, Caparica  
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

N. Datia  
ISEL, Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa,  
Lisboa, Portugal  
e-mail: datia@isel.ipl.pt

## 1 Context and Motivation

Crimes, traffic accidents, forest fires, respiratory infections, human interaction with mobile devices (e.g., tweets), among others, are leading to the storage of lot of spatiotemporal events with high levels of detail (LoDs). We consider spatiotemporal events as data with the following structure:  $(S, T, A_1, \dots, A_N)$ , where  $S$  describes the spatial location of the event,  $T$  specifies the time moment, and  $A_1, \dots, A_N$  are attributes detailing what has happened. As an example of a spatiotemporal event we have a car accident occurred in some latitude and longitude, at eight o'clock, with two victims.

The underlying space-time complexity in spatiotemporal events makes the data analysis process very challenging which at a first glance may seem a chaos of events without any particular meaning. For example, an overview of geo-referenced car accidents occurred in USA<sup>1</sup> is displayed in Fig. 1. From it, we can perceive hotspots of accidents in metropolitan areas, which is common-sense knowledge. However, “interesting patterns” relating to space and time may be hidden in the vast amount of data that is usually displayed and analyzed.

Both spatial and temporal attributes of events can be expressed at different spatial and temporal LoDs. For example, they can range from grids with different cell sizes (e.g., cells of 2 km<sup>2</sup> or 4 km<sup>2</sup>) to cities or countries, and from seconds to months or years, respectively. The LoD reflects the units' size in which phenomena are aggregated/summarized, likely affecting the user perception about them (Andrienko et al. 2010; Laurini 2014; Silva et al. 2015b).

A change in the phenomenon's LoD can bring improvements for the analytical activity (Camossi et al. 2008; Laube and Purves 2011; Silva et al. 2015b). From one phenomenon's LoD to another, some patterns can become easily perceived and different patterns may be detected. Moreover, the volume and complexity of data can be reduced without affecting the user's analytical capability.

The LoD plays a crucial role during the analytical process and, often, there is no exclusive LoD to analyze a phenomenon (Keim et al. 2008; Andrienko et al. 2010). This key idea is illustrated in Fig. 2 by displaying the number of accidents in USA using time series across multiple LoDs. For example, at LoD 2.27 km<sup>2</sup> *Month*, a cyclical pattern for each year can be easily detected while at LoD *CountiesYear*, a decreasing trend is observed.

The identification of the proper LoDs to analyze a spatiotemporal phenomenon is a key issue for the users (Keim et al. 2008; Andrienko et al. 2011). However, standard practices provide Visual Analytics tools that work on a single LoD driven by the user (Maciejewski et al. 2010; Ferreira et al. 2013). To understand what LoD would be adequate to detect patterns, users have to probe hypotheses, which can be a challenging task.

---

<sup>1</sup>USA car accidents occurred between 2001 and 2013, which corresponds to about 450.000 geo-referenced accidents: <http://www.nhtsa.gov/FARS>.

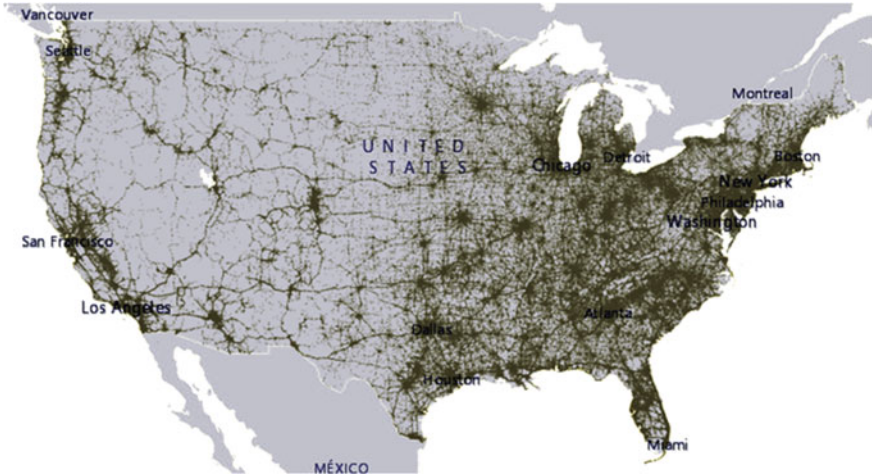


Fig. 1 An overview of the car accidents in USA

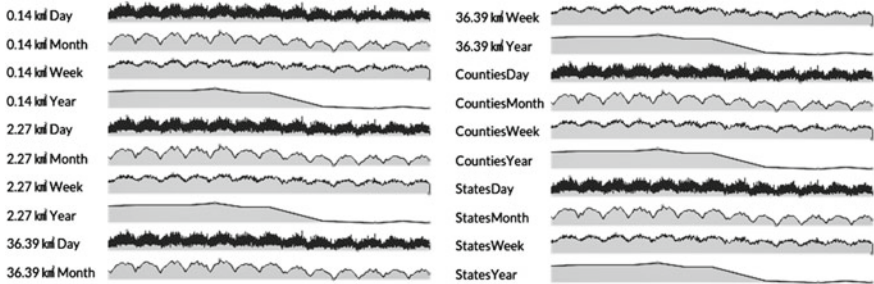


Fig. 2 The number of accidents of USA across multiple LoDs

Firstly, the users often face difficulties to specify in advance what an “interesting pattern” is. The importance of a pattern depends on the specific application, the analysis question, and its concordance with domain knowledge (Keim et al. 2008; Andrienko et al. 2010). Secondly, the LoDs in which “interesting patterns” can easily be perceived are often difficult to determine a priori (Sips et al. 2012). Approaches allowing users to study and explore phenomena across multiple LoDs are necessary (Camossi et al. 2008; Keim et al. 2008; Andrienko et al. 2010; Sips et al. 2012).

To meet this need, the mainly contribution of this work is a framework to build summaries, at different LoDs, about phenomena described by spatiotemporal events. As our framework does not make any assumption about the phenomenon, the analytical task and the phenomenon’s LoDs, it can be widely used to get an overview of the phenomenon under analysis. The framework establishes five types

of abstracts working with space and time together. The users can inspect those abstracts across multiple LoDs simultaneously, helping to understand either in what LoDs the phenomenon perception distinguishes itself or in what LoDs “interesting patterns” emerge. A Visual Analytics prototype was also developed that implements our contributions.

This paper is organized as follows. Section 2 presents related work about modelling spatiotemporal phenomena at different LoDs, and discusses proposals working on different LoDs. In Sect. 3, the background needed about the granularity theory is given in which the proposed framework is based on. Section 4 presents the framework for summarizing spatiotemporal events. Section 5 presents the Visual Analytics prototype. Section 6 concludes with some remarks about the work undertaken and guidelines for future work.

## 2 Related Work

To observe spatiotemporal phenomena at different LoDs, we need, on one hand, approaches that are able to model spatiotemporal phenomena at different LoDs, and on the other hand, we need analytical approaches that can work across different LoDs. Such approaches were researched and are further discussed in this section.

Several approaches for modelling spatiotemporal phenomena at multiple LoDs under different terminologies like multirepresentation, multiresolution and granular computing are presented.

Multirepresentation approaches (Parent et al. 2009) provide different point of view from a spatiotemporal phenomenon, allowing the observation of the same geographical space and/or interval of time at different perspectives.

Multiresolution approaches (Stell and Worboys 1998; Zhou et al. 2004) focused essentially in the generalization of spatial features which involves feature simplification, dimensionality reduction, and whether a spatial feature should exist in some spatial LoD or not (Weibel and Dutton 1999).

Granular computing approaches model phenomena at several LoDs based on granularities. There are several proposals for granularities definitions, namely temporal granularity (Bettini et al. 2000), spatial granularity (Camossi et al. 2006; Pozzani and Zimányi 2012) and a granularity definition applicable to any domain (Pires et al. 2014). Granularities can be related through relationships, allowing one to compare and relate granules belonging to different granularities. This is useful to model phenomena at different LoDs.

Camossi et al. (2006) propose a granular computing approach to index and aggregate spatiotemporal information at different LoDs in object-oriented database management systems. However, more than indexing spatiotemporal information at different LoDs, the granularity theory proposed in Pires et al. (2014) and Silva et al. (2015a, b) was devised to represent spatiotemporal phenomena at different LoDs in

a semi-automatic way, also including some concepts from multirepresentation and multiresolution (see Sect. 3). A more detailed discussion about modeling spatiotemporal phenomena at different LoDs can be found in Silva et al. (2015a).

Although there are several approaches to model spatial phenomena at different LoDs, few works on approaches that work across several LoDs were found.

Camossi et al. (2008) propose a spatiotemporal clustering technique applicable to different temporal and spatial LoDs in order to improve a clustering algorithm efficiency. The appropriate temporal and spatial LoD depends on a trade-off between the mining efficiency and the maximum detail desired, which is an input parameter. The choice of the temporal and spatial LoD is done iteratively through the LoDs available until the best trade-off is found. This approach is independent from the application domain and is focused on a particular analytical “task” (i.e., clustering). Furthermore, it follows a fully automatic approach to find the proper temporal and spatial LoD.

When a user is at an early stage of the analysis, semi-automatic approaches are desirable so that human judgment can be involved in the analytical process as stated by the Visual Analytics principles (Keim et al. 2008). In other words, approaches that combine automatic algorithms and interactive interfaces including the user in-the-loop, relying on users’ cognitive capabilities and domain knowledge. The human involvement in the analytical process is crucial as the appropriate LoDs may depend on the specific application, the analysis question, and its concordance with domain knowledge.

Some proposals working on several LoDs appeared from the Visual Analytics research area. Sips et al. (2012) propose a Visual Analytics approach called *Pinus*, aiming the detection of patterns at multiple temporal LoDs in numerical time series, specifically from environmental sciences. This approach makes no assumption about the temporal LoD. The *Pinus* visualization allows users to visually detect the temporal LoDs where interesting patterns emerge in the time series. Goodwin et al. (2016) propose a framework for analyzing multiple variables across spatial LoDs and geographical location. They use a novel interactive visualization to identify correlation in multiple variables allowing a user’s analysis in several spatial LoDs and geographical location simultaneously.

Sips et al. (2012) and Goodwin et al. (2016) recognize the importance of the LoD in the user’s analysis that is embedded in their proposals. Particularly, the former focuses on several temporal LoDs, and the latter handles multiple spatial LoDs. Furthermore, the former approach is independent from the user’s analytical task while the latter is focused on comparing multiple variables across geographical location and spatial LoDs.

Approaches working across several LoDs are needed and, as shown, they are starting to be developed. Yet, to the best of our knowledge, there are no approaches that work across several spatial and temporal LoDs, and that are independent from the analytical task and the domain applicable in the context of spatiotemporal events.



### 3 Granularity Theory

A granularity theory devised to model spatiotemporal phenomena at multiple LoDs was proposed in Pires et al. (2014) and Silva et al. (2015a, b). This theory provides the concepts needed to our framework that are explained below.

The granularity theory lies on the concept of granularity. Granularities perform divisions of a domain. Each division corresponds to a non-decomposable entity, mentioned as a granule. A granularity was formally defined as follows (Pires et al. 2014).

**Definition 1** (*Granularity*) Let  $\mathcal{IS}$  be an index set;  $D$  be a domain;  $2^{DS}$  the power set of the  $DS$ ; and  $GS$  be a subset of the power set of the  $DS$  apart from the empty set  $GS \subseteq 2^{DS} \setminus \{\emptyset\}$  such that any two elements are disjoint from each other. A granularity  $G$  is a bijective mapping:

$$G : GS \rightarrow \mathcal{IS} \quad (1)$$

A granularity is a set of granules where each one is composed by its extent  $g \in GS$  and its index value  $ind \in \mathcal{IS}$  denoted by  $g_{ind}$ . *States*, *Counties* are examples of spatial granularities and *Hours*, *Days* are examples of temporal ones.

Granularities can be related through relationships allowing one to compare and relate granules belonging to different granularities, useful to hold spatiotemporal data at different LoDs. A fundamental relation between granularities is the relationship *finer than*, i.e., a granularity  $G$  is *finer than*  $H$  if and only if each extent of granule of  $G$  is contained in one and only one extent of a granule of  $H$  ( $G \preceq H$ ). For example, *Counties* is finer than *States* ( $Counties \preceq States$ ) and *Hours* is finer than *Days* ( $Hours \preceq Days$ ).

A granularities-based model is composed by a set of atoms which are used to make statements about phenomena. An atom is a predicate symbol together with its arguments such that arguments are granules from granularities. Each predicate is defined by a signature. The signature declares the granularities and the respective granules that can be used in the arguments. Let  $P \in \mathcal{P}$  be  $n$ -ary predicate with a set of arguments denoted by  $Args(P)$ , and  $\mathcal{G}$  a set of granularities of the model. A predicate signature is of form  $P(\{(arg, G_{(P,arg)}) \mid arg \in Args(P) \wedge G_{(P,arg)} \subseteq \mathcal{G}\})$  declaring the set of valid granularities for each argument  $G_{(P,arg)}$ .

Let's consider that we want to describe crimes in USA. We introduce the predicate symbol *crime* to describe the spatial location, the time moment and the number of victims resulting from a crime:  $crime(when, where, victims)$ . The crime signature's predicate is defined as follows: (i)  $G_{(crime,where)} = \{Counties, States\}$ ; (ii)  $G_{(crime,when)} = \{Hours, Days\}$ ; (iii)  $G_{(crime,victims)} = \{Numbers, Binary\}$ . Note that, the granularity *Numbers* corresponds to  $\mathbb{N}$  and the granularity *Binary* is defined as: (0, "No Victims"), ( $[1, \varepsilon]$ , "With Victims") where  $\varepsilon$  is the maximum number of victims occurred in a crime. This way,  $Numbers \preceq Binary$ .

An atom is of form  $P(\tau)$  with  $\tau = \{(arg, g) | arg \in Args(P) \wedge g \text{ is a granule of a valid granularity } G_{(P,arg)}\}$ .  $\tau$  denotes the tuple of terms of an atom. An example of an atom describing an crime can be:  $crime(\{(where, Oakland), (when, 03/01/2015 \text{ 18 h}), (victims, 0)\})$  where the underlying granularities used are  $\{(where, Counties), (when, Hours), (victims, Numbers)\}$ .

Until now, the term LoD was informally used. However, the concept of LoD was formally defined (Silva et al. 2015b). One valid LoD of the predicate  $P$  consists of a set of argument pairs and a valid granularity. Looking at Fig. 3, the  $LoD_a \{(where, Counties), (when, Hours), (victims, Numbers)\}$  and  $LoD_b \{(where, Counties), (when, Days), (victims, Numbers)\}$  are two examples of valid LoDs of the  $crime$  predicate. The set of all valid LoDs of the predicate  $P$  is denoted by  $\mathcal{L}^P$ . From now on, a LoD will refer to one LoD belonging to  $\mathcal{L}^P$ .

Two LoDs can be related by the relationship *is more detailed than* formally defined in Silva et al. (2015b). The  $LoD_a$  is more detailed than  $LoD_b$  because  $County \leq States, Hours \leq Days$  and  $Numbers \leq Binary$ . All LoDs in  $\mathcal{L}^P$  form a partial order along with the relation *is more detailed than*. In our example, the partial order formed by the LoDs in  $\mathcal{L}^{crime}$  is illustrated in Fig. 3.

Using a predicate  $P$ , a phenomenon is described by a collection of atoms for each valid LoD. There might be equal atoms in some LoDs which are described in Silva et al. (2015b) in the form of:  $G_{Syn}(P(\tau), f)$ , where  $f$  is the number of atoms in the form  $P(\tau)$  such that  $f > 0$ . For the sake of simplification and assuming atoms of the predicate  $P$ , we will use the following notation:  $(\tau)/f$ . Notice that,  $(\tau)/f$  can also be mentioned as an atom according to the granularities-based model definition (Silva et al. 2015b).

Events are captured at lowest LoD regarding all valid LoDs of the predicate  $P$ . After that, and using the generalization concept introduced in Silva et al. (2015b), events can be generalized automatically for any coarser LoD. For example,  $crime(\{(where, Oakland), (when, 03/01/2015 \text{ 18 h}), (victims, 0)\})$  at  $LoD_a$  can be generalized to  $crime(\{(where, Oakland), (when, 03/01/2015), (victims, 0)\})$  at  $LoD_b$  as well as can be generalized to  $crime(\{(where, Michigan), (when, 03/01/2015 \text{ 18 h}), (victims, 0)\})$  at  $LoD_c$ .

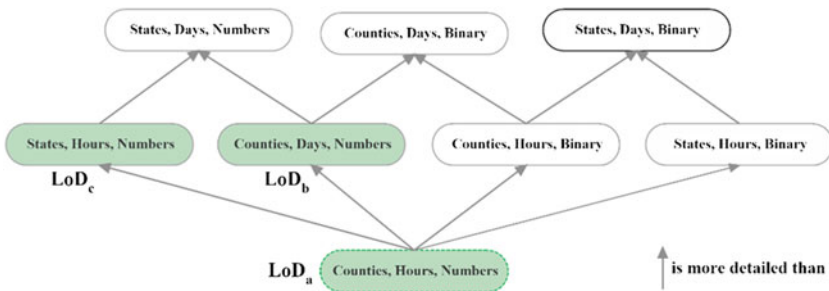


Fig. 3 The partial order of all valid LoDs of the crime predicate

Using the granularity theory, we can have spatiotemporal events at multiple LoDs. As discussed, the LoD has an important role for the user's analysis. Therefore, analyses across multiple LoDs simultaneously are needed.

## 4 SUITE: A Framework for Summarizing Spatiotemporal Events

The goal of this work is to propose a framework to build statistical summaries, at different LoDs, about phenomena described by spatiotemporal events. The users should be able to inspect and compare the phenomenon's perception across multiple LoDs, without having to look at raw data, and consequently, not needing to handle the spatiotemporal complexity.

The proposed framework is based on the granularity theory but, since we are assuming spatiotemporal events, we consider that each used predicate has one and only one argument *where* describing the spatial location of the event, and one and only one argument *when* specifying the time moment. Other arguments can be used to detail what has happened.

The signature of *event* follows the pattern,  $event((where, G_{(P,where)}), (when, G_{(P,when)}), Args)$ , and  $Args = \{(arg_1, G_1), \dots, (arg_n, G_n)\}$  represents the signature for the other arguments. We also assume that any valid spatial granularity does not have a temporal evolution (Silva et al. 2015a), i.e., the used spatial granularities remain stable along the considered temporal scope.

Let  $\alpha = \{(where, S), (when, T), \dots, (arg_n, G_n)\} \in \mathcal{L}^{event}$  be a *LoD* of *event*. An atom  $event((where, s), (when, t), args)/f$  represents, at a some *LoD* from  $\mathcal{L}^{event}$ ,  $f$  spatiotemporal events that happen located on a spatial granule  $s$ , at a temporal granule  $t$ , described by  $args = \{(arg_1, v_1), \dots, (arg_n, v_n)\}$ .

An atom is associated with one and only one *spatiotemporal grain*  $st$  composed by a spatial granule  $s$  and a temporal granule  $t$ ,  $st = (s \in S, t \in T)$ . In general, a set of *event* atoms is associated to each spatiotemporal grain  $st \in S \times T$ . This set may be empty, meaning that no event happened at the spatiotemporal grain  $st$ ; or the set has just one atom  $event((where, s), (when, t), args)/f$  meaning that  $f$  similar events happened at the spatiotemporal grain  $st$ ; or the set has many atoms, meaning that many different events happened at the spatiotemporal grain  $st$ .

This way, a granularities-based model (or just model),  $\mathcal{M}(event)^\alpha$ , regarding a predicate *event* at a *LoD* can be described as a set of indexed collections of atoms, each being indexed by a spatiotemporal grain from  $S \times T$ ,

$$\{st \rightarrow \{event((where, s), (when, t), args)/f\} | st = (s, t) \in S \times T\} \quad (2)$$

Let's consider a simple synthetic example of *event* with  $\{S_1, S_2\}$  as the set of valid spatial granularities,  $\{T_1, T_2\}$  the set of valid temporal granularities and just one additional argument with just one granularity  $G_{arg}$ . Assuming  $T_1 \leq T_2$  and

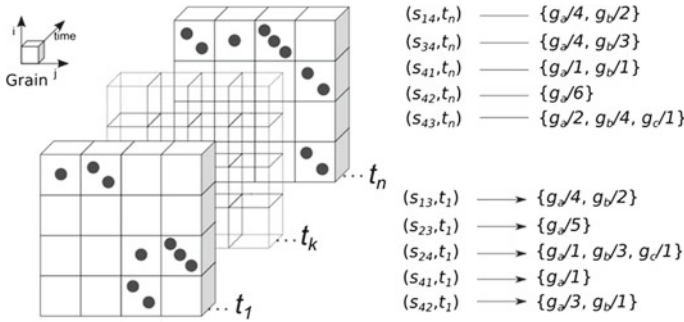


Fig. 4 Schematic representation of  $\mathcal{M}(\text{event})^\alpha$

$S_1 \leq S_2$  the most detailed *LoD* is  $\alpha = \{(where, S_1), (when, T_1), (arg, G_{arg})\}$ . Let's represent by  $s_{ij}$  the spatial granules from  $S_1$ ,  $t_k$  the temporal granules from  $T_1$  and  $G_{arg} = \{g_a, g_b, g_c\}$ . Figure 4 presents a set of available atoms indexed by each spatiotemporal grain  $(s_{ij}, t_k)$ . Each atom is written in a simplified form, such that  $event((where, s_{ij}), (when, t_k), (arg, g_{arg}))/f$  is just represented by  $g_{arg}/f$ . For instance, the set of atoms associated with  $(s_{13}, t_1)$  is  $\{g_a/4\}$  and  $\{g_b/2\}$ , and the set of atoms associated with  $(s_{24}, t_1)$  is  $\{g_a/1, g_b/3, g_c/1\}$ .

The USA car accident dataset, represented in Fig. 1, can be described at each *LoD*  $\alpha$  by an equation similar to (2), i.e.,  $\mathcal{M}(\text{accident})^\alpha$ , where each spatiotemporal grain  $st = (s, t)$  index a set of atoms representing the accidents which happened at that spatiotemporal grain. We can apply simple statistics to summarize  $\mathcal{M}(\text{accident})^\alpha$ .

For instance, some spatiotemporal grains  $st$  index empty sets while others index non-empty sets. The percentage of spatiotemporal grains with non-empty sets, named *occupation rate*, measures the average density of a model at a given *LoD*. Let's consider the spatial granularities  $grid(0.14 \text{ km}^2)$ ,  $grid(2.27 \text{ km}^2)$ ,  $grid(36.39 \text{ km}^2)$ , *Counties*, *States*, and the temporal granularities *Day*, *Week*, *Month*, and *Year*. Figure 5 shows the occupation rate for different combinations of spatial and temporal granularities. As we can see in Fig. 5, the occupation rate increases with coarser granules.

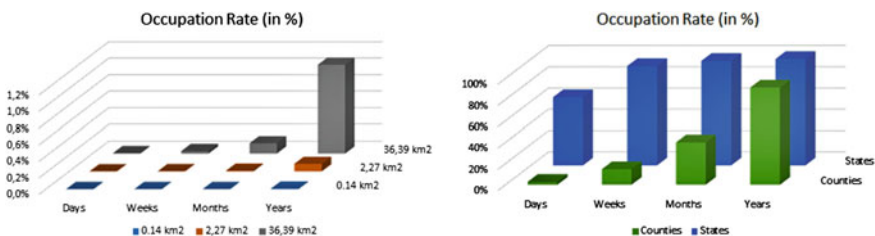
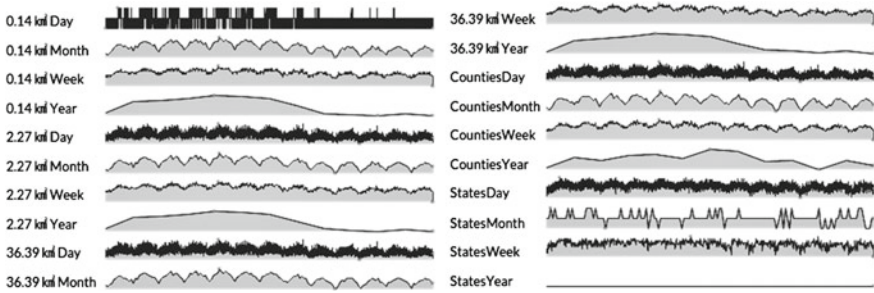


Fig. 5 The occupation rate for different combinations of spatial and temporal granularities



**Fig. 6** The occupation rate computed at each temporal granule

At each LoD, the context for the occupation rate, shown in Fig. 5, is global in the sense that it considers all spatiotemporal grains. The same computation can be done for each temporal granule  $t_i$ , considering all the spatiotemporal grains  $st = (s, t_i)$ . In that case, we get the temporal evolution for the occupation rate computed at each spatial context, as shown in Fig. 6. Each time series is displayed based on its maximum and minimum values. Bearing this in mind, at  $0.14 \text{ km}^2 \text{ Month}$  LoD a cyclical pattern is observed, for instance; and, at  $StatesYear$  LoD the one value is a constant which means that there is at least one accident in each state for each year. Another pattern can be seen at  $36.39 \text{ Year}$  LoD, for instance, which is showing a decreasing trend.

On the other hand, the occupation rate computation can also be done for each spatial granule  $s_j$ , considering all the spatiotemporal grains  $st = (s_j, t)$ , getting for each spatial granule the occupation rate across all the temporal granules. Figure 7 shows two maps where each “point” represents a spatial granule and its color is given by the occupation rate value according to the map’s legend (see Fig. 7).

The map at  $0.14 \text{ km}^2 \text{ Days}$  LoD shows an outlier, highlighted by a dash circle. In the “yellow” spatial granule, there are accidents occurring with some degree of frequency in comparison with the other granules. When we change the LoD to  $0.14 \text{ km}^2 \text{ Years}$ , the perception is changed and that outlier is no longer perceived.

To achieve this kind of perception, the proposed framework builds statistical summaries of each phenomenon’s LoD to support users in carrying inspection and comparison tasks of a phenomenon across multiple LoDs. Observing summaries across multiple LoDs can provide useful information to identify the proper ones to carry out a particular analysis. We will call those statistical summaries as *abstracts*.

## 4.1 Abstracts

Our framework was designed to build abstracts over  $\mathcal{M}(P)^\alpha$ . An abstract  $\mathbb{A}$  can be a number, a vector, or even a matrix measuring a particular feature of a

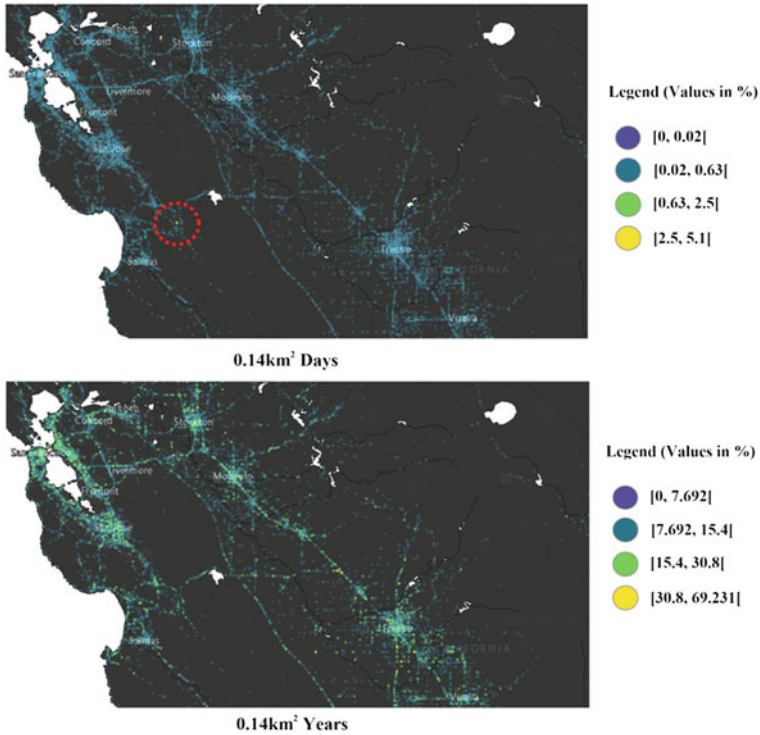


Fig. 7 The occupation rate computed at each spatial granule

phenomenon. Five types of abstracts with different contexts are introduced: (i) Global Abstract; (ii) Temporal Abstract; (iii) Spatial Abstract; (iv) Compacted Temporal Abstract; (v) Compacted Spatial Abstract.

**Definition 2 (Global Abstract)** Let  $\mathcal{M}(P)^\alpha$  be the set of granular syntheses indexed by each spatiotemporal grain. Thus, a function  $\mathbb{F}_{Global} : (\mathcal{M}(P)^\alpha) \rightarrow \mathbb{A}_{Global}$  produces a global Abstract such that  $\mathbb{A}_{Global}$  is one abstract  $\mathbb{A}$ .

For instance, in Fig. 5, we are displaying Global Abstracts i.e., the occupation rate for each LoD of the model  $\mathcal{M}(accident)^\alpha$ . Global Abstracts may hide some important variations in space and/or time. Hence, we introduce the possibility to create abstracts that are more “detailed”. One of them is the Spatial Abstract.

**Definition 3 (Spatial Abstract)** Let  $\mathcal{M}(P)^\alpha$  be the set of granular syntheses indexed by each spatiotemporal grain. Thus, a function  $\mathbb{F}_{Spatial} : (\mathcal{M}(P)^\alpha) \rightarrow \mathbb{A}_{Spatial}$  produces an abstract for each temporal granule such that  $\mathbb{A}_{Spatial} = \{(t, \mathbb{A}) | t \in T\}$ .

A Spatial Abstract is a summary based on  $\mathcal{M}(P)^\alpha$  for each  $t \in T$ . It allows us to look at the evolution of a summary over time, which is measuring a **spatial feature**

of a phenomenon. Figure 6 shows several Spatial Abstracts (i.e., the occupation rate), one for each LoD.

In the same way as the Spatial Abstract allows one to look at a summary over time, we introduce the Temporal Abstract to look at a summary over space. A Temporal Abstract is a summary based on  $\mathcal{M}(P)^\alpha$  for each  $s \in S$ . It allows us to look at a summary over space, which is measuring a **temporal feature** of a phenomenon.

**Definition 4** (*Temporal Abstract*) Let  $\mathcal{M}(P)^\alpha$  be the set of granular syntheses indexed by each spatiotemporal grain. Thus, a function  $\mathbb{F}_{Temporal} : (\mathcal{M}(P)^\alpha) \rightarrow \mathbb{A}_{Temporal}$  produces an abstract for each spatial granule such that  $\mathbb{A}_{Temporal} = \{(s, \mathbb{A}) | s \in S\}$ .

Two examples of Temporal Abstracts can be found in Fig. 7, in which the occupation rate is computed for each spatial granule  $s$ .

Moreover, each Spatial (or Temporal) Abstract can be further summarized into a single summary that we called Compacted Spatial (or Temporal) Abstract.

**Definition 5** (*Compacted Spatial Abstract*) Let  $\mathbb{A}_{Spatial}$  be a Spatial Abstract. Thus, a function  $\mathbb{F}_{CompactSpatial} : (\mathbb{A}_{Spatial}) \rightarrow \mathbb{A}_{CompactSpatial}$  produces a Compacted Spatial Abstract such that  $\mathbb{A}_{CompactSpatial}$  is one abstract  $\mathbb{A}$ .

For each Spatial Abstract (i.e., time series) displayed in Fig. 6, we can use an aggregation measure, like the average, to produce a Compacted Spatial Abstract. Other methods that come from descriptive statistics as well as methods to analyze time series can be used to build Compacted Spatial Abstracts.

**Definition 6** (*Compacted Temporal Abstract*) Let  $\mathbb{A}_{Temporal}$  be a Temporal Abstract. Thus, a function  $\mathbb{F}_{CompactTemporal} : (\mathbb{A}_{Temporal}) \rightarrow \mathbb{A}_{CompactTemporal}$  produces a Compacted Temporal Abstract such that  $\mathbb{A}_{CompactTemporal}$  is one abstract  $\mathbb{A}$ .

For each Temporal Abstract (i.e., map) displayed in Fig. 7, we can also use an aggregation like the average to produce a Compacted Temporal Abstract. Other methods that come from descriptive statistics or spatial statistics can be used to produce Compacted Temporal Abstracts.

## 4.2 Properties of Abstracts Functions

Abstracts are built through functions. Each function will be measuring certain feature of phenomena which in turn can employ different strategies using different information from the model  $\mathcal{M}(P)^\alpha$ . Bearing this in mind, we identified three properties that can further characterize the function that computes an abstract. These properties describe the way each spatiotemporal grain contributes to the Abstract computation, i.e., the way each  $\gamma = st \rightarrow \{event((where, s), (when, t), args)/f\}$  is integrated for the resulting abstract. They are: (i) neighbourhood dependency; (ii) spatiotemporal dependency; (iii) semantic dependency. These properties are further detailed.



**Neighbourhood dependency:** The contribution of each  $\gamma$  for the *Abstract* depends (or not) on the spatiotemporal neighbourhood. This neighbourhood dependency may be only temporal (e.g., depends only on the events that happen on the same spatial granules but on previous temporal granules); only spatial (e.g., depends only on the events that happen on the same spatial neighbour's granules but on the same temporal granule); or may be both spatial and temporal dependent.

When an *Abstract* computation is not dependent then the computation of  $\mathbb{F}(\{\gamma\})$ , where  $\gamma = st \rightarrow \{event((where, s), (when, t), args)/f\}$ , can be rewritten as  $\mathbb{F}(\{\gamma\}) = Agg_{\mathbb{F}}(\{\mathbb{F}'(\gamma)\})$ , where  $\mathbb{F}'$  computes the contribution of each  $\gamma$  and  $Agg_{\mathbb{F}}$  aggregates those contributions to get the final *Abstract*. This means that  $\mathbb{F}'$  can be a function of local computation not requiring information about others  $\gamma$ .

**Spatiotemporal dependency:** the contribution of each  $\gamma$  for the *Abstract* depends (or not) on the specific spatiotemporal grains  $st = (s, t)$  of  $\gamma$ . This spatiotemporal dependency may be only temporal (e.g. the contribution is different if the events happened at night or during day, or even varying with the season); only spatial (e.g., the contribution is different if the events happened at high mountains or at sea level, or even varying according to the spatial granule like the specific counties); or may be both spatial and temporal dependent.

When an *Abstract* computation is not spatiotemporal dependent then the computation of  $\mathbb{F}(\{\gamma\})$ , where  $\gamma = st \rightarrow \{event((where, s), (when, t), args)/f\}$ , can be rewritten. Consider  $\gamma'$  as  $\gamma' = st \rightarrow \{event(args)/f\}$  where we removed the information about  $s$  and  $t$  and leave the set  $\{event(args)/f\}$  indexed by  $st$  just to keep any neighbourhood information between spatiotemporal grains. Then,  $\mathbb{F}(\{\gamma\}) = \mathbb{F}(\{\gamma'\})$ .

When an *Abstract* computation is neither spatiotemporal dependent nor neighbourhood dependent,  $\mathbb{F}(\{\gamma\})$ , can be rewritten as  $\mathbb{F}(\{\gamma\}) = Agg_{\mathbb{F}}(\{\mathbb{F}'(\{\gamma'\})\})$ , where  $\mathbb{F}'$  computes the contribution of each set  $\{\gamma'\}$  independently of their spatiotemporal location, and  $Agg_{\mathbb{F}}$  aggregates those contributions to get the final *Abstract*.

**Semantic dependency:** the contribution of each  $\gamma$  for the *Abstract* depends (or not) on the semantic arguments of  $\gamma$ . When the *Abstract* is not semantic dependent then  $\gamma = st \rightarrow \{event((where, s), (when, t), args)/f\}$  can be simplified to  $\gamma = st \rightarrow \{event((where, s), (when, t))/f\}$ . For instance, if we are studying the car accidents an *Abstract* semantic dependent will consider the type of accident and/or the number of victims, while an *Abstract* semantic independent only considers the number of accidents.

When an *Abstract* computation is neither spatiotemporal dependent nor neighbourhood dependent nor semantic dependent,  $\mathbb{F}(\{\gamma\})$ , can be rewritten as  $\mathbb{F}(\{\gamma\}) = Agg_{\mathbb{F}}(\{\mathbb{F}'(\{f\})\})$ , where  $\mathbb{F}'$  computes the contribution of each bag  $\{f\}$ , and  $Agg_{\mathbb{F}}$  aggregate those contributions to get the final *Abstract*. The occupation



rate is an extreme example of such *Abstract* and can be defined based on:  $\mathbb{F}(\{\gamma\}) = \text{if } |\{\gamma\}| > 0 \text{ then } 1 \text{ else } 0$ :

$$\text{Agg}_{\mathbb{F}}(\{x\}) = \frac{\sum x}{|S \times T|} \quad (3)$$

### 4.3 Discussion

Our framework allows one to define or use many functions available in the literature that create summaries of data.

As presented, the functions computing abstracts may be semantic dependent. Such dependency is delimited by the predicate's signature regarding the arguments *args*. These arguments depend on the phenomenon itself. In the case of car accidents, one may collect information about the number of victims, whether some of the drivers present an alcoholic rate above the legally allowed, information about weather conditions, among others. A function computing an abstract can use this information. For instance, one can compute the occupation rate by weather conditions as Global Abstract; or we can use the Global Moran's I (Moran 1950) to build a Spatial Abstract that measures the correlation between spatiotemporal grains and the weather conditions.

Furthermore, the functions producing abstracts may be spatial and/or temporal dependent. In case of dependency, it is important to have a base knowledge for each spatial and temporal granule and that base knowledge should be relevant for the phenomenon in study. Some examples to describe a temporal granule are: the time of the day that each temporal granule exists (e.g., night or day), what kind of season it is in. Concerning the spatial granules, they can be characterized, for instance, as information about altitude, if is a rural or urban area, among others.

Moreover, the functions computing abstracts may be neighbourhood dependent. This dependency can make the functions likely to be more time-consuming when compared with the neighbourhood independent ones. Some examples are given: (i) the enhanced Jacquez  $k$  nearest neighbour test (Malizia and Mack 2012) can be an example of a Global Abstract neighbourhood dependent; (ii) the number of clusters of spatiotemporal grains identified by a 4D spatiotemporal density-based clustering approach proposed in Oliveira et al. (2013) can be an example of Global Abstract computed by a function neighbourhood and semantic dependent; (iii) Keogh et al. (2005) propose an algorithm to find the most unusual subsequence within a time series, which can be used as Temporal Abstract. Such abstract is computed by a function temporal neighbourhood dependent; (iv) based on the Fourier discrete transform, a function may compute a Temporal Abstract returning the  $n$  higher frequencies. Such function is temporal neighbourhood dependent.

In the absence of the neighbourhood dependency, functions making abstracts can work individually for each spatiotemporal grain as discussed. For this reason, parallel computing techniques can be employed.

Moreover, in the analysis of an abstract using different LoDs one needs to be aware of the *Modifiable Areal Unit Problem*, MAUP (Openshaw and Openshaw 1984), which means that patterns may be biased due to how raw data are aggregated. Since the SUITE's framework was proposed to allow analyses across multiple LoDs such problem may be identified or discarded sooner. For example, when a pattern is only visible in a specific LoD it can be further validated. One might conclude that the pattern suffers from MAUP and can be ignored or, if the phenomenon specifically operates there, it can be considered valid. Therefore, we argue that the analysis across multiple LoDs can attenuate the MAUP.

Finally, the functions computing abstracts may be used in different abstracts holding different properties and sometimes they depend on the phenomena in study. However, it is fundamental that those functions provide comparable abstracts. Ultimately, we aim to support users carrying inspection and comparison tasks of a phenomenon across multiple LoDs. To this end, comparable abstracts are fundamental to allow a fair comparison among phenomenon's LoDs.

## 5 SUITE's Prototype

The SUITE's prototype was developed, implementing the granularities-based model specific for spatiotemporal events. The server provides a set of RESTful Web services (Spring) implemented in Java and is relying on PostgreSQL as the Database Management System. The browser-based client handles user interaction and data presentation. It's coded in JavaScript, HTML5, and uses WebGL to display efficiently thematic maps.

The prototype receives as input a dataset of spatiotemporal events and a predicate signature, and then it generates automatically the set of atoms for each LoD of the corresponding predicate. The set  $\mathcal{L}^P$  is inferred based on the granularities defined for each argument and the relationship *finer than* that exists among them. During the generalization's computation, local summaries associated to each spatiotemporal grain are built. For example, the number of spatiotemporal events, the number of distinct atoms  $event((where, s), (when, t), args)/f$ , among others. These are used by the abstracts' functions in order to take advantage of the computation already performed when they are executed. Notice that, the functions developed to generate abstracts so far are displayed in Appendix. Moreover, new abstracts' functions can easily be added, since the prototype was developed in a modular way.

Figure 8 presents the three main panels of the SUITE Prototype user interface: Global, Spatial, and Temporal Abstract. Each panel is showing a thematic partial order for the computed LoDs of the dataset of car accidents in USA. Each node represents a LoD and is colored according to the chosen Abstract function, in this

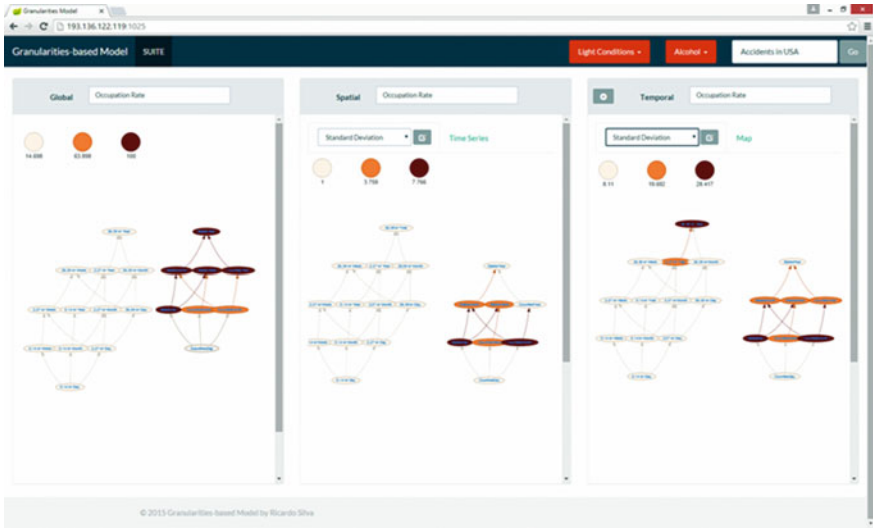


Fig. 8 An overview of the SUITE's prototype

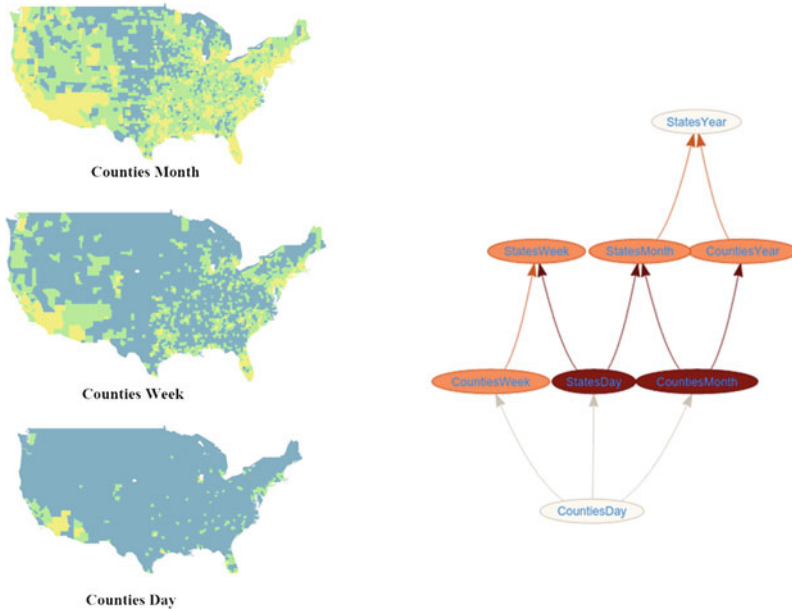


Fig. 9 The thematic partial order of the Compacted Temporal Abstract and the Temporal Abstract occupation rate across three LoDs

case the occupation rate. For the Spatial, and Temporal panels an additional function is required (in this case the standard deviation) to compute Compacted Spatial Abstracts and Compacted Temporal Abstract.

The thematic partial orders allow users to inspect and compare the results of abstractions functions, measuring a particular feature of a phenomenon. In Fig. 9 the thematic partial order of the Compacted Temporal Abstract displayed in Fig. 8 was zoomed in. We can see different classes of values of the corresponding Compacted Temporal Abstracts for the LoDs *CountiesDay*, *CountiesWeek* and *CountiesMonth*. The respective Temporal Abstracts are also displayed in Fig. 9, and as can be seen, different perceptions about the occupation rate were achieved.

Thematic partial orders should be further explored by analyzing the Spatial Abstracts or Temporal Abstracts. Notice that the time series displayed in Fig. 6 are Spatial Abstracts, and the maps shown in Figs. 7 and 9 are Temporal Abstracts coded into visual representations. These “perspectives” are shown together in Fig. 10 regarding the occupation rate considering only the accidents with drunk drivers involved. Notice that, in SUITE’s prototype users can conduct “semantic” filters based on the arguments *args* specified in the predicate signature.

Observing the Spatial Abstract value at *StatesMonth* LoD (left side in Fig. 10), we can see a change in the behavior of the occupation rate. It changes from a slightly constant behavior to a cyclic one, and this was not observed considering all accidents as shown in Fig. 11.

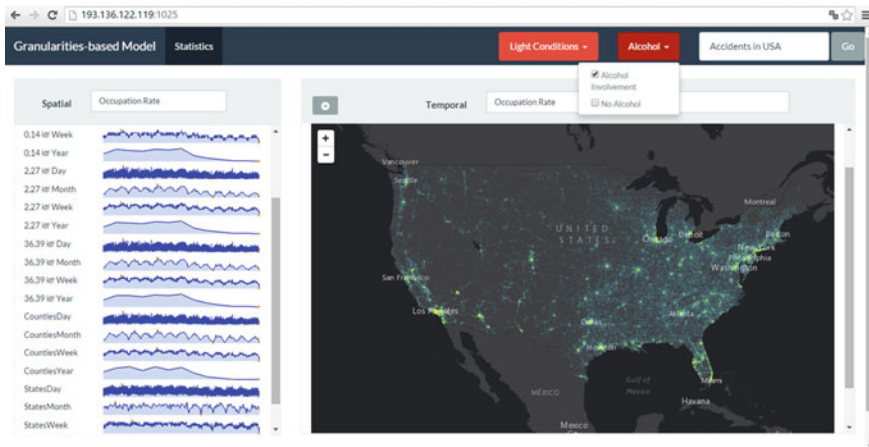


Fig. 10 The Temporal Abstract perspective and one Spatial Abstract at 36.39 km<sup>2</sup> Month using the occupation rate

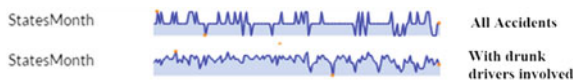


Fig. 11 Spatial Abstracts considering all accidents and with the accidents with drunk drivers

The SUITE’s prototype allows one to conduct analyses of abstracts across multiple LoDs, providing hints about in what LoDs the phenomenon perception distinguishes itself or in what LoDs “interesting patterns” may emerge.

## 6 Conclusions and Future Work

Standard practices provide tools that work on a single LoD driven by the user. However, the LoD plays a crucial role during the analytical process and, often, there is no exclusive LoD to analyze a phenomenon.

This paper presents a framework to build abstracts, at different LoDs, about phenomena described by spatiotemporal events. As our framework makes no assumption about the phenomenon, the analytical task and the phenomenon’s LoDs, it can be widely used to get an overview of the phenomenon under analysis. The framework establishes five type of abstracts working with space and time together as well as the properties to characterize the functions to compute them. This allows us to frame many proposals in the literature that create summaries of data in the proposed abstracts and properties.

The SUITE’s prototype implements the proposed framework allowing analyzes across multiple LoDs. Datasets of spatio-temporal events are automatically generated for the multiple LoDs, and for each one, abstracts can be computed. The prototype already implements a set of functions to compute abstracts (see Appendix) but new abstracts functions can be easily added.

Future work can be directed to further experimentation of the SUITE’s prototype, namely, the assessment of domain experts of the proposed framework, and to enrich the prototype with new functions to compute abstracts.

**Acknowledgments** This work has been supported by FCT—Fundação para a Ciência e Tecnologia MCTES, UID/CEC/04516/2013 (NOVA LINCOS) and UID/CEC/00319/2013 (ALGORITMI), and COMPETE: POCI-01-0145-FEDER-007043 (ALGORITMI).

## Appendix

Measure	Description	Global	Spatial	Temporal	Properties
Atoms Collision (%)	Percentage of granules with events, where atom collisions exits	●	●	●	
Occupation Rate (%)	Percentage of granules with events	●	●	●	
Bray-Curtis Similarity for Atoms	Calculates the similarity based on the counts of		●		▲

(continued)

(continued)

Measure	Description	Global	Spatial	Temporal	Properties
	atoms, between consecutive temporal grains				
Bray-Curtis Similarity for Synthesis	Calculates the similarity based on the number of granular synthesis, between consecutive temporal grains		●		▲
Correlation Index for Atoms	Correlation between the number of atoms of consecutive temporal grains		●		▲
Correlation Index for Synthesis	Correlation between the number of granular synthesis of consecutive temporal grains		●		▲
Dice Similarity (Binary)	Dice index (event/no event) between consecutive temporal grains		●		▲
Gower Similarity (Binary)	Similarity (event/no event) between consecutive temporal grains		●		▲
Jaccard Similarity (Binary)	Jaccard index (event/no event) between consecutive temporal grains		●		▲
Moran's I	Calculates the spatial autocorrelation among nearby locations, given a domain specific variable		●		▲ ●
Maximum Number of sequential occurrences	Calculates the maximum number of event happening for a spatial grain			●	▲

Legend

▲ Neighborhood dependent    🗺️ Spatio-temporal dependent    ● Semantic dependent

## References

Andrienko G et al (2010) Space, time and visual analytics. *Int J Geogr Inf Sci* 24(10):1577–1600  
 Andrienko G et al (2011) A conceptual framework and taxonomy of techniques for analyzing movement. *J Vis Lang Comput* 22(3):213–232  
 Bettini C, Jajodia S, Wang S (2000) *Time granularities in databases, data mining, and temporal reasoning*. Springer  
 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

- Camossi E, Bertolotto M, Kechadi T (2008) Mining spatio-temporal data at different levels of detail. In: The European Information Society. Springer, pp. 225–240
- Ferreira N et al (2013) Visual exploration of big spatio-temporal urban data: a study of New York City taxi trips. *IEEE Trans Vis Comput Graph* 19(12):2149–2158
- Goodwin S et al (2016) Visualizing Multiple Variables Across Scale and Geography. *IEEE Trans Vis Comput Graph* 22(1):599–608
- Keim D et al (2008) Visual analytics: definition, process, and challenges. In: Kerren A et al (eds) *Information Visualization, Lecture Notes in Computer Science*. Springer, Berlin Heidelberg, pp. 154–175
- Keogh E, Lin J, Fu A (2005) Hot sax: Efficiently finding the most unusual time series subsequence. In: *Fifth IEEE International Conference on Data Mining*, 8 pp
- Laube P, Purves RS (2011) How fast is a cow? Cross-scale analysis of movement data. *Trans GIS* 15(3):401–418
- Laurini R (2014) A conceptual framework for geographic knowledge engineering. *J Vis Lang Comput* 25(1):2–19
- Maciejewski R et al (2010) A visual analytics approach to understanding spatiotemporal hotspots. *IEEE Trans Vis Comput Graph* 16(2):205–220
- Malizia N, Mack EA (2012) Enhancing the Jacquez k nearest neighbor test for space–time interaction. *Stat Med* 31(21):2318–2334
- Moran PAP (1950) Notes on continuous stochastic phenomena. *Biometrika*, pp. 17–23
- Oliveira R, Santos MY, Moura Pires J (2013) 4D + SNN: A Spatio-Temporal Density-Based Clustering Approach with 4D Similarity. In: *2013 IEEE 13th International Conference on Data Mining Workshops (ICDMW)*, pp. 1045–1052
- Openshaw S, Openshaw S (1984) The modifiable areal unit problem
- 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
- Pozzani G, Zimányi E (2012) Defining spatio-temporal granularities for raster data. In: *Data Security and Security Data*. Springer, pp. 96–107
- 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 et al (2015b) Aggregating spatio-temporal phenomena at multiple levels of detail. In: *AGILE 2015*. Springer Science Business Media, pp. 291–308
- 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
- Weibel R, Dutton G (1999) Generalising spatial data and dealing with multiple representations. *Geograph Inform Syst* 1:125–155
- Zhou X et al (2004) Multiresolution spatial databases: making web-based spatial applications faster. In: Yu J et al (eds) *Advanced Web Technologies and Applications SE – 5, Lecture Notes in Computer Science*. Springer, Berlin Heidelberg, pp. 36–47

# Representation and Visualization of Imperfect Geohistorical Data About Natural Risks: A Qualitative Classification and Its Experimental Assessment

Cécile Saint-Marc, Marlène Villanova-Oliver, Paule-Annick Davoine, Cicely Pams Capoccioni and Dorine Chenier

**Abstract** Imperfections, often called ‘uncertainties’, exist in almost every spatio-temporal dataset, especially in historical data. They are of different types (unreliability, inaccuracy...) and concern every data dimension (space, time and theme). Based on previous work, this article proposes a synthesis qualitative classification of imperfection types. This classification has been assessed with domain experts (hydrologists, geophysicians and GIScientists working in a railway company) during an experiment, that gave positive results towards the use of this classification. Participants were also asked to evaluate the seriousness of each imperfection type in an analysis context. This evaluation has allowed to associate a quantitative index to each imperfection type and to visualize a quantity of imperfection attached to each spatial object in a map.

**Keywords** Imperfection · Uncertainty · Qualitative classification · Geohistorical data · User-based study

---

C. Saint-Marc (✉) · M. Villanova-Oliver · P.-A. Davoine  
Laboratoire d’Informatique de Grenoble, Grenoble, France  
e-mail: cecile.saint-marc@imag.fr

M. Villanova-Oliver  
e-mail: marlene.villanova-oliver@imag.fr

P.-A. Davoine  
e-mail: paule-annick.davoine@imag.fr

C. Saint-Marc · C. Pams Capoccioni · D. Chenier  
SNCF Ingénierie and Projets, La Plaine Saint-Denis, Paris, France

C. Pams Capoccioni  
e-mail: cicely.pams@reseau.sncf.fr

D. Chenier  
e-mail: dorine.chenier@reseau.sncf.fr



# 1 Introduction

Imperfections are present in almost every spatio-temporal dataset. It is defined as the fact that data representing reality differ from reality itself in various ways (Plew 2002). In the literature, the terms ‘uncertainty’, ‘unreliability’ or ‘data quality issue’ are employed as synonyms. We decided to use the term ‘imperfection’ in this article, as a broad term including all these representations.

Geographic information about the past, named *geo-historical* information, is especially concerned by imperfection, because it often comes from testimonies and archive documents, which deliver incomplete and sometimes uncertain information. Natural risk studies particularly require working over historical natural phenomena to understand at-risk processes in space and to better anticipate and prevent future events.

In a study we have made about the historical major flood of November 1999 in France and its impacts on the railway system, we have counted that 327 records out of 399 (82 %) in the dataset contained imperfections. Each dimension of data may be affected by imperfections: its location, its dating and its theme. In the dataset we studied, 36 % of the locations, 64 % of the dating and 14 % of the data values were imprecise or uncertain. Even if a dataset is uncertain or imprecise, it must often be included in spatial analysis because it is the only available resource. Thus the management of information imperfections in the field of natural risks is crucial.

Buttenfield (1993) identified two issues to communicate the imperfection of information. The first one was the lack of methods to measure and represent the numerous aspects of imperfection in geospatial data. Should imperfection be qualified or quantified? What was the best method? Where and how could it be included in information systems? The second issue was to figure it in communication mediums. Is it possible to visualize imperfection at the same time as data? Which methods may enable to interact with these visualizations in a comprehensive and useful manner? Moreover, may users agree to use this information, even if it is a complex and time-consuming additional parameter to integrate in their decisions (Kinkeldey et al. 2015).

Two ways exist to represent imperfections in geospatial data: to quantify it, for example in the form of statistical probabilities (e.g. Lowell 1997; Potter et al. 2012; Zoghلامي et al. 2012), or to qualify it with a set of appropriate terms (e.g. MacEachren et al. 2005; Thomson et al. 2005; Skeels et al. 2008; Arnaud 2009; Snoussi et al. 2012). Qualification allows the user to make the difference between types of imperfection whereas quantification has the advantage to be suitable for calculations by computers. Qualitative form can also be converted into quantitative form if needed (Thomson et al. 2005). This article focuses on qualitative classifications of imperfection and conditions for their visualization in a map.

The first section of the article deals with existing classifications of imperfection (here, ‘classifications’ stand for ‘definitions of different *types* of imperfection’) and a synthesized classification is proposed. The second section introduces the visualization methods to represent multiple types of imperfection simultaneously in a map

and explains how they can be used in the context of our synthesis classification. The third section describes a user study conducted with experts of the French railway company (SNCF), in order to validate our classification in a practical context. Finally the results are discussed and outlooks are given.

## 2 Representation of Spatio-Temporal Data Imperfection

Research in GIScience has shown interest in representing and visualizing imperfection for about 25 years now. This section deals with a synthesis of previous works related to the qualitative expression of imperfection. Based on these works, a synthesized classification is then presented.

### 2.1 *Classifications of Imperfection: A State of the Art*

#### 2.1.1 **Extent of This State of the Art**

Many works in GIScience interested in classifying imperfection types (Veregin 1989; Pornon 1992; Plew 2002; Thomson et al. 2005; Griethe and Schumann 2006; Pang 2008; Arnaud 2009; ISO 2013). For reasons of length, this article does not aim to review extensively all the literature about classification of imperfection. After an exploration of previous work, we have decided to describe three of them that review previous work before proposing their own classification, and that defended different viewpoints about imperfections (Skeels et al. 2008; Arnaud 2009; Snoussi et al. 2012). The first one focuses on spatial data use cases and has been developed with domain experts. The second one is GIScience-oriented with a social science approach, considering the format and human processing of spatio-temporal data. The third one is a high-level classification, designed for computer science and not specifically for geographic information, but which uses a more exhaustive terminology.

#### 2.1.2 **A Review of Reviews**

A first synthesized classification was proposed by Arnaud (2009) in the domain of natural risk geohistorical data. The author has reviewed typologies of imperfection in GIScience and Social Science. Mainly inspired by the classification by Thomson et al. (2005), it described 10 types of imperfection, ordered by origins of imperfection: information-related factors, human-related factors and media-related factors (Fig. 1). Each of these categories counted two to five imperfection factors.

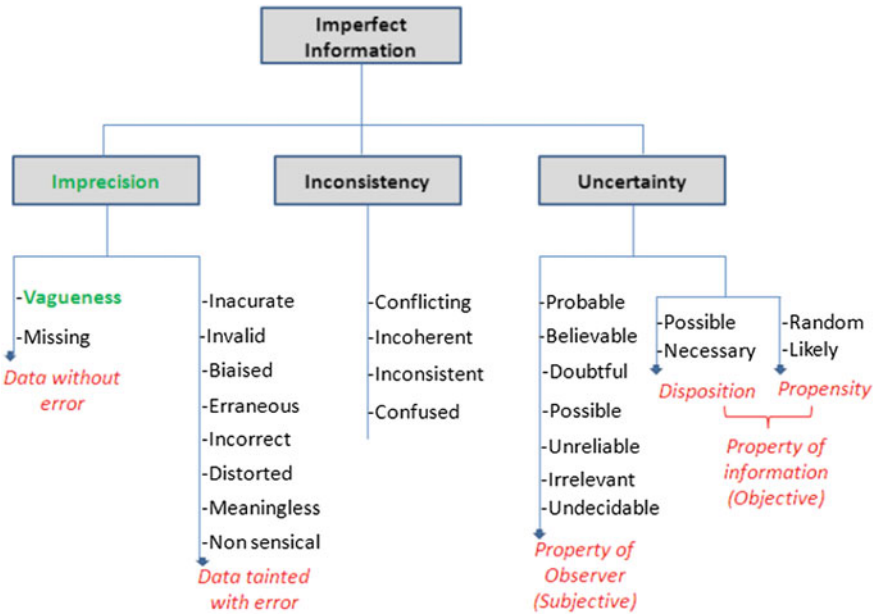
Uncertainty factors		Social sciences		GIScience			
		Roland-May (2000)	Lang et alii (2005)	Buttenfield (1994)	Thomson et alii (2005)	Skeels et alii (2007)	Griethe et alii (2006)
Information-related factors	Imprecision (fuzzy dimension)						
	- semantic	x					
	- spatial	x	x	x			
	- attributory			x	x	x	
	- granularity			x			
	Objectivity (exactness)			x	x		
	Validity				x		
	Incompleteness				x		
	Exhaustivity			x		x	
Human-related factors	Subjectivity of authors	x			x	x	x
	Logical consistency (agree/disagree)			x	x	x	
Medium-related factors	Technical errors in lineage (genealogy)	x	x	x	x	x	x
	Reliability/credibility of sources		x	x	x	x	x
	Interrelatedness of sources				x		

Fig. 1 Classification of uncertainty factors (translated from Arnaud 2009)

Seven of these factors are classified into two main categories: factors which characterize the information, in a similar way as metadata (*incompleteness*, *subjectivity of authors*, *logical consistency*, *objectivity* or *exactness*) and quality criteria, relative to the data itself (*imprecision*, *reliability of data sources*, *genealogy errors*). Arnaud (2009) specified that these factors are related and influence each other. For example, *subjectivity of authors* may induce issues in the *logical consistency* of information or in the *precision* of information. An interest of this classification is its inspiration from social sciences, which enriched the classification by taking into account data sources and human factors.

The second classification was proposed by Smets (1997) in the field of artificial intelligence. We consider here a version of this classification that was enriched by Snoussi et al. (2012). It is presented in Fig. 2. An interest of this classification is that it divides imperfections in three main categories: *imprecision* (real values are located in a set of values, finite or infinite—e.g. ‘at the north of’, ‘during 20th century’), *inconsistency* (contradiction between pieces of information) and *uncertainty* (partial knowledge that does not allow deciding if the value is true or false). Imprecision terms are classified in two groups: data with and without errors, while uncertainty terms are separated between objective and subjective assessment.

This second classification is quite exhaustive, presenting 24 types of imperfection. The whole terminology is organized in the form of a decision tree, which might help to pick up the right term corresponding to the encountered imperfection. The main drawback of this proposal is that it is based only on a reduced number of previous works, essentially in the domains of computer sciences and artificial intelligence. As a consequence, even if Snoussi et al. (2012) have applied it to



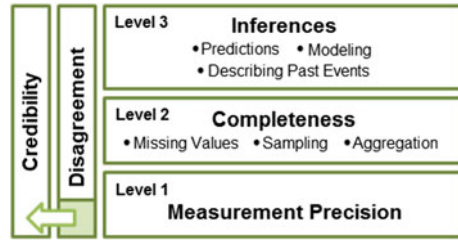
**Fig. 2** Taxonomy of imperfect information (Snoussi et al. 2012). Terms written in bold green are those that were added to the previous classification by Smets (1997)

qualify spatio-temporal phenomena in the field of natural risks, it has not been aligned with previous work in GIScience. Thus it is not obvious that the proposed terms are well chosen in comparison with the usual vocabulary of spatio-temporal data and so, that it is suitable for users who have to process geographic information.

The third classification we analyzed was proposed by Skeels et al. (2008). Its main interest is to focus on the data user. The authors have reviewed imperfection types used in GIScience and Scientific Visualization (SciVis) fields. Five common factors of imperfection are identified: *approximation* (estimated measurement of a phenomenon), *prediction/modelling* (if it is not sure whether or not a phenomenon will occur or occurred in the past), *disagreement* (or inconsistency), *incompleteness* and *credibility* (of the data or the data source). The authors interviewed 18 participants, of several knowledge domains, who used to deal with imperfections in their work. The interviews have contributed to refine the classification.

Their interviews have resulted in classifying the multiple types of imperfection as levels (or layers) of imperfection (Fig. 3). The lowest level corresponds to imperfection due to *imprecision of measurements*. The middle level concerns *incompleteness* in the data, caused by sampling, data aggregation or missing values (when it is not sure whether they are really missing or not). The highest level refers to imperfection in inferences made with the data, referring here to the way data is processed to make decisions (modelling, prediction or inferences about the past). The last two imprecision factors, *disagreement* and *credibility*, may concern each of

**Fig. 3** Classification of levels (or layers) of uncertainty (Skeels et al. 2008)



the three levels. According to the authors, *disagreement* and *credibility* are practically linked because if an inconsistency is observed, the credibility of the dataset is often undermined. One or several levels may concern a dataset or a project. Users considered the levels segmentation as crucial and problematic at the same time, because “uncertainty within one level, even if well quantified at that level, rarely can be adequately transformed or accounted for at another level when the decision-making process requires a transition between levels” (Skeels et al. 2008). In spite of this limit, levels segmentation remains important in order to formalize the level at which an imperfections exist in a dataset.

The main strength of this study is that users were involved in the making of the classification. It leads to a functional classification, corresponding to the different steps of project management. The segmentation in levels may be interpreted as a wish of the users to identify the sources of the imperfection they observe in their data.

The three classifications reported here show different points of view regarding the classification of imperfection. Interesting features of each classification may be combined to build a synthetized classification.

## 2.2 *Synthesis of Imperfection Classifications*

To synthetize these approaches, we based our work on the classification presented in Snoussi et al. (2012): we kept the exhaustiveness and the tree-based formalization. Some types considered in GIScience were missing: *validity* (or currency/timing) and *interrelatedness of sources*, used by Arnaud (2009), and *aggregation* (or summarized information) used by Skeels et al. (2008).

Moreover, classifications of Arnaud and Skeels et al. have shown that users are interested in the origin or level of imperfections. Their proposals could be combined in three stages, which might be applied to data sources, the data itself or to results of data processing:

- Stage 1: Data production. May generate imprecision (vagueness or imprecision with errors), incompleteness, missing data, interrelatedness of sources, inconsistency, subjectivity and reliability issues.

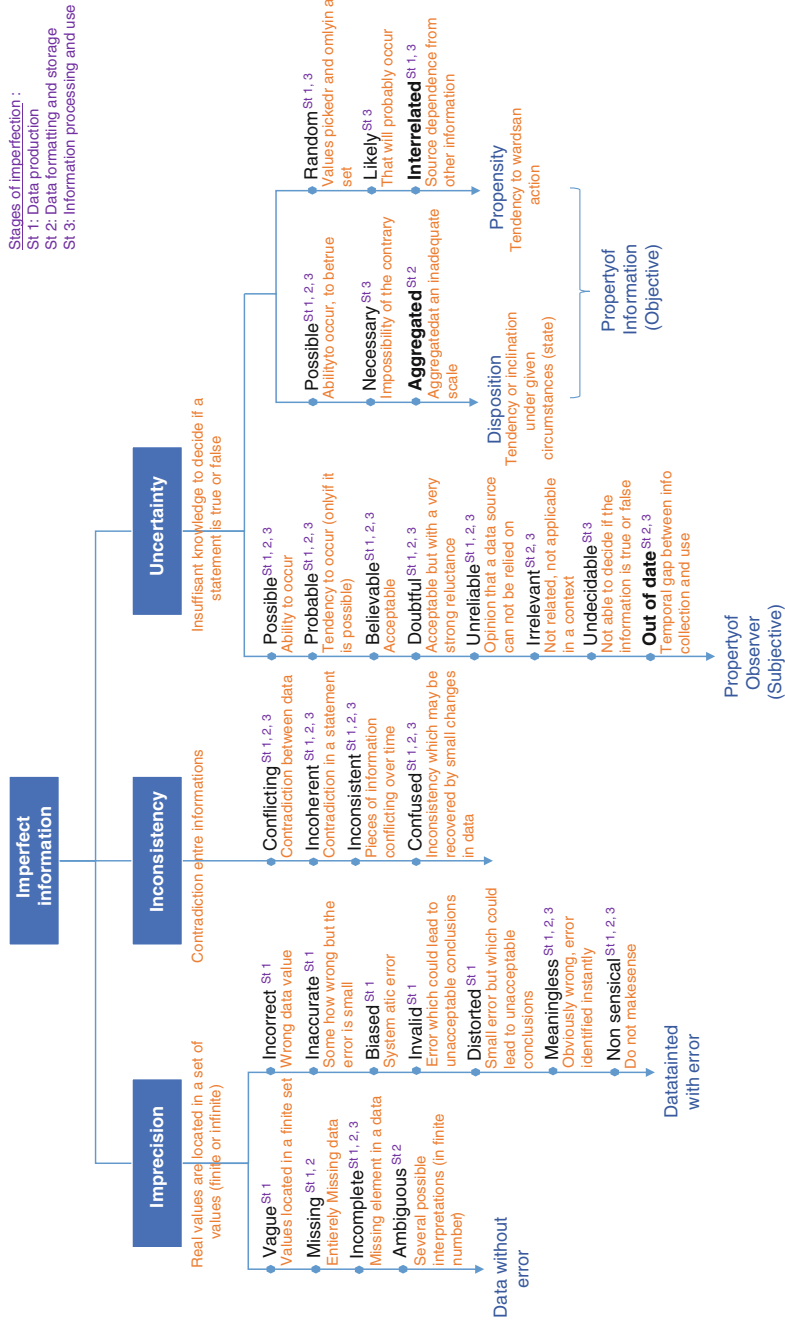


Fig. 4 Synthesis qualitative typology of imperfections

- Stage 2: Data formatting and storage. May generate incompleteness, missing data, imprecision due to aggregation, ambiguity, inconsistency, subjectivity and reliability issues.
- Stage 3: Information processing and use. May generate incompleteness, objective uncertainty of results, consideration of irrelevancy, out-of-date issues, inconsistency, subjectivity and reliability issues.

The resulting classification is presented in Fig. 4. Stages are indexed to imperfection types. Stage indexes may be helpful for users to find the best term to qualify an imperfection in their data or, on the contrary, to identify possible origins of an observed imperfection.

Once the different types of imperfection that can occur in a geospatial dataset are qualified, finding a suitable way to visualize them in a map is another issue. Integration in the map faces challenges, caused by the complexity of this information and the fact that it adds an additional information to the content of the map, what means an additional information to process for the user.

### 3 From the Qualitative Classification to Its Visualization

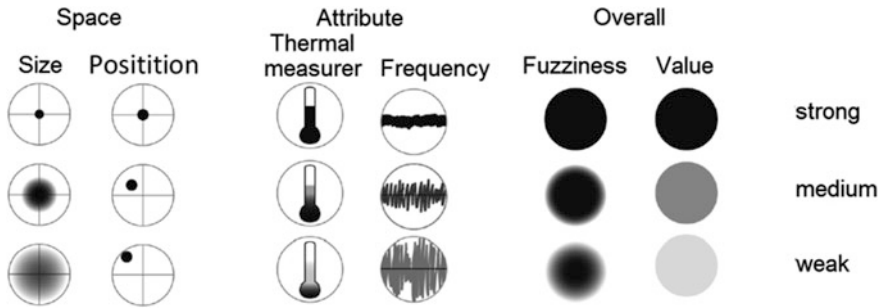
The suitability of a classification partly depends on its future use and the way it may be visualized in a map. This step has raised some questions and justified the design of the experiment we conducted to assess our classification.

#### 3.1 *Inspiration from Previous Work*

In spatio-temporal data, every dimension of the information may be concerned by imperfection. Adding imperfection dimensions in maps may cause design issues: visual clutter, occlusions between data or overload of information (Potter et al. 2012).

Many authors have been interested in finding the best representation for one type of imperfection which concerns a given data dimension (Buttenfield 1988; Olston and Mackinlay 2002; Cliburn et al. 2002; MacEachren et al. 2012). However, when several dimensions of data are concerned at the same time by imperfections of different qualitative types, it is challenging to visualize them all in one map.

A few authors have proposed methods to visualize several types of imperfection in a same map thanks to written labels (Buttenfield 1988; Arnaud 2009, p. 201), color mixing methods (O'Brien and Cheshire 2014), animated symbols and interactivity (Arnaud and Davoine 2011). A few others have dealt with the co-visualization of multiple data dimensions concerned by an imperfection type: concentric rings around imperfect areal objects (Bodin 2002) or collage of several iconic symbols next to each other (Seccia et al. 2014). Different kinds of imperfection and different imperfect data dimensions are sometimes aggregated in an



**Fig. 5** Tested symbols to represent confidence levels in spatial and attribute dimensions, and an overall confidence in the data (Seccia et al. 2014)

overall index of imperfection, which can be visualized on the map. Lay users tend towards basing their decisions on the overall imperfection information (Seccia et al. 2014).

In the literature, ordinal scales of visual variables or iconic symbols are mostly used to visualize imperfections, such as in Fig. 5. This representation implies a kind of quantification of the qualitative imperfection terms. So, for visualization purposes, a transition is needed from a qualitative to a quantitative point of view. The quantification of imperfection enable to apply aggregation methods in order to summarize imperfect indicators (Thomson et al. 2005).

### 3.2 *Quantification and Aggregation of Individual Imperfections*

In order to convert qualitative imperfection in a quantity, we propose to give a rate out of ten to each subtype of imperfection, according to the seriousness of its supposed impact on decision making. For example, a value considered *doubtful* will be rated as being more serious for analysis (8 out of 10) than a *credible* value (2 out of 10). The way quantitative indexes are aggregated is important. A poorly-calculated composite indicator, which compiles individual indicators, may send a misleading message, be misinterpreted and it may be misused if the construction process is not transparent (Saisana and Tarantola 2002). A report published by the Organization for Economic Cooperation and Development (OECD) have given insights on the process of constructing composite indicators (Nardo et al. 2008).

In the context of imperfect spatio-temporal data, two successive steps of aggregation of imperfections can be performed: firstly, the seriousness rates could be aggregated for each data dimension independently (which means spatial, temporal and descriptive attribute considered separately), and secondly, the ratings calculated for each data dimension could in turn be aggregated together, resulting in an overall index of imperfection.



**Table 1** Example of aggregation of ratings with three calculation methods

Scenario 1		Scenario 2	
Imperfection	Rating	Imperfection	Rating
Type 1	9	Type 1	9
Type 2	2	Type 2	2
		Type 3	7
Arithmetic mean	<b>5.5</b>	Arithmetic mean	<b>6</b>
Weighted arithmetic mean w/weights equal to ratings (favors high scores)	<b>7.7</b>	Weighted arithmetic mean w/weights equal to ratings (favors high scores)	<b>7.4</b>
Geometric mean (favors low scores)	<b>4.2</b>	Geometric mean (favors low scores)	<b>5</b>

Results of calculations are written in bold font.

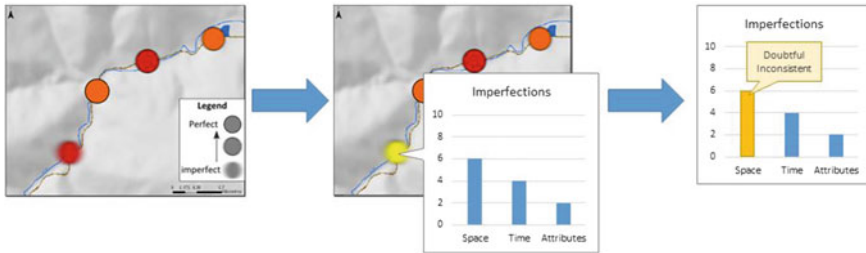
For the first level, either a weighted arithmetic mean or a geometric mean would be suitable. Indeed, they both enable compensability between good and bad rates (Nardo et al. 2008). Geometric mean favors low results for composite indicators that include low values, whereas compensability is higher with arithmetic mean. Examples of calculation results with arithmetic, weighted arithmetic and geometric means are presented in Table 1. As it does not seem to be appropriate to lower the displayed imperfection, arithmetic mean seems to be the most adapted in the context of visualizing imperfection. At the second level, an equal-weighted arithmetic mean would seem suitable, considering that each data dimension has the same importance.<sup>1</sup>

In the case where multiple imperfect datasets should be crossed to make an analysis, imperfection propagation methods could be used to transfer initial imperfection levels to the final result. We do not deal with this complex field of research in this article because it is not a major topic in the case of geohistorical data, where multiple data sources are usually simply gathered and compiled in a single layer of information. However, it would be an important topic of future work for a use of this geohistorical datasets in risk analysis.

### 3.3 Proposal of Visualization Method

Following the information-seeking mantra in visualization (Schneiderman 1996), users should be given the ability to visualize first the overall aggregated imperfection rating in the map display, then to access detailed rates of imperfection affecting each data dimension, thanks to interactivity, and finally to detail further, up to the terms qualifying imperfection types (Fig. 6). To visualize imperfection

<sup>1</sup>This hypothesis of equal importance for each data dimension needs to be assessed with users. It is currently a work in progress.



**Fig. 6** Process of visualizing imperfect information in a map: first overview of data-related overall imperfection thanks to fuzziness (*left*), then details about imperfections affecting data dimension (*middle*), finally details about qualitative aspects of imperfection (*right*)

affecting several dimensions of data, previous work propose to use glyphs, such as bar charts (Pang 2008; Potter et al. 2012). We propose to implement a graph visualization in order to show the amount of imperfection, in an extrinsic popup window.

If imperfection types are intended to be represented by quantitative ratings, these ratings should be objectively defined. To know if our qualitative classification seemed usable in a practical context and if qualitative types could be objectively associated with ratings, we conducted evaluation interviews with domain experts.

## 4 Evaluation of the Proposed Classification by Domain Experts

### 4.1 Objectives and Hypothesis

The opinion of expert users about our final classification (see Fig. 4) has to be known, in order to assess its usability. Our first hypothesis is that this detailed taxonomy is too rich and complex compared to the users' needs in their everyday professional activities. We first aim at defining which terms of the classification would be hard to understand or too much specific. Second, we try to identify which terms are already used by experts and are the most important for them. If the classification has to be simplified, then we will keep these important terms and synthesize others.

From a visualization perspective, we wanted to associate a quantified index to imperfect data, in order to give an overview of its imperfection before detailing on-demand. Therefore, we propose to participants to rate the seriousness of each imperfection type. Our hypothesis is that the rates would be rather homogeneous for each type of imperfection (e.g. doubtful, inaccurate...), what would enable us to associate each type of imperfection with an objective rate, so an objective quantitative measure.

## 4.2 Experimental Design

### 4.2.1 Participants

We interviewed twelve experts of the French railway company, working in domains related to our use case (cartography of past floods that impacted the railway system). Eight were women and four were men. They were aged between 20 and 39 years old. Among them were one archivist, one geotechnical engineer, three hydraulics engineer, one hydro-geologist, five GIS specialists and one computer scientist (Fig. 7). All were used to work with geographic data, using it several times a week, except one hydraulics engineer who used it only once a week on average.

### 4.2.2 Material and Method

Participants ran in a four steps experiment, which last about 30 min in total. First, general information were given about the proceeding of the experiment and the three main categories of imperfection of our classification were explained. We ensure that participants understood the explanations by asking them before continuing.

The second and third steps consisted in exercises to perform. Twenty-eight small cards were given to participants. On each one was written the name of a type of imperfection in black, followed by a short description of its meaning in a smaller blue lettering. In the first exercise, participants had to sort all the cards in three categories: 'I encounter this type of imperfection in my work', 'I understand this type but I do not encounter it' and 'I do not see what this type corresponds with'.

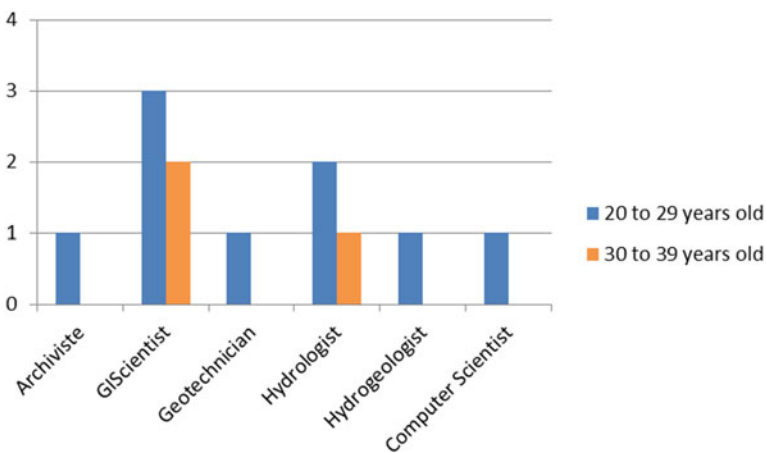


Fig. 7 Profiles of participants involved in the study

In the second exercise, the cards previously sorted in the category ‘I do not see what it corresponds with’ were moved aside, in order to prevent that comprehension issues had impacts on the result of exercise 2. Participants had to rate remaining cards on a discrete scale out of 10 points, by placing cards on a frame graduate from 0 to 10. Their grade represented the seriousness of the impact that this kind of imperfect data might have on the analysis performed with it. During both exercises, participants could ask any questions they wanted. Pictures of the card positions were taken at the end of each exercise, to process the results later.

The last part consisted in a structured interview to collect the opinion of participants: if they find useful to consider imperfection in their studies, what they thought about the classification, if they could plan to use it in their work and finally what their opinion was about the experiment.

## 4.3 Results

### 4.3.1 Exercise 1

Exercise 1 aimed to identify imperfection types that experts used to meet during their work and if some types were difficult to understand. We tested two hypothesis, thanks to a Khi-square test:

- Hypothesis 1: some imperfection types are significantly less understood than others.
- Hypothesis 2: some imperfection types are significantly encountered more often than others.

#### *Hypothesis 1: Imperfection types which are less understood*

Raw results showed that less understood types (by less than 3 participants, or less than 25 %) were: *Distorted* (41.7 % of the participants), *Random* (33 %) and *Likely* (25 %).

To test if this difference of understanding is significant, the number of answers ‘I encounter this type’ and ‘I understand this type’ were summed up in a same category. The applicability of Khi-square test is questionable because theoretical numbers of answers were not all superior to five. With an alpha error-risk of 5 %, the test showed that the difference of understanding is not significant. Hypothesis 1 is rejected.

#### *Hypothesis 2: Imperfection types which are more encountered*

Most of the types are encountered in experts’ work (about 58 %). Imperfection types which were often encountered (by more than 75 % of the participants) are: *Out of date*, *Credible*, *Doubtful*, *Imprecise*, *Incomplete*, *Missing* and *Vague*. Types that are also more encountered than the statistical mean are: *Interrelated*, *Incoherent*, *Incorrect*, *Unreliable*, *Irrelevant*, *Possible* and *Probable*.

Imperfection types that are mainly understood but not encountered are: *Ambiguous, Conflicting, Confused, Meaningless, Inconsistent, Invalid, Non sensical* and *Aggregated*.

To test if some types are significantly more encountered than others, we grouped the categories 'I understand this type' and 'I do not see what this correspond to'. With an alpha error-risk of 5 %, the test is significant: some imperfection types are really more encountered by experts in their work than others.

### 4.3.2 Exercise 2

Exercise 2 aimed at quantifying the impact of various imperfection types on analysis performed with this kind of imperfect data. First, we had to test if there was a significant difference between ratings of different types. If so, it would be possible to comment them.

#### *Analysis of the relationship between types and ratings*

We computed univariate statistics for every series of rates per imperfection type (Fig. 8). Most of ratings dispersions (represented by standard deviations) were low, except for nine types that we detail hereafter. For the types with low dispersion, we could see large differences between rating means, showing that a significant relation was possible.

We performed an ANOVA test on imperfection types and their ratings. With an alpha error-risk of 5 %, the relation is significant: ratings are significantly different between imperfection types. So it is justified to give rates to imperfection types.

#### *Ratings distribution*

Types that were judged homogeneously serious by participants were: *Invalid, Incorrect, Incoherent, Distorted, Ambiguous, Conflicting, Random* and *Undecidable*. Types that were judged homogeneously not to be a problem were: *Likely* and *Credible*.

Opinions diverged a lot (standard deviation >3) about: *Non sensical, Irrelevant, Necessary* and *Biased*. They were also quite high (standard deviation >2.5) for: *Meaningless* and *Unreliable*, in high ratings, and *Probable, Vague* and *Confused*, in low ratings.

Qualitative interviews have given explanations about the types *Non sensical, Irrelevant* and *Meaningless*. Indeed some participants explained that these types were put aside of analysis de facto because they are explicitly false or uninteresting, so they did not have any impacts on the quality of analysis. They gave them a rate of 0 out of 10. Other participants rated them as if imperfect data was integrated in the analysis and so, they gave high seriousness ratings. For the other types for which standard deviations were high, participants seemed to disagree but there was no clear explanation.

For others imperfection types, participants quite agreed to say that the following types were serious for the analysis performed with this kind of imperfect data: *Missing, Inconsistent, Incomplete, Aggregated, Doubtful* and *Dependant*. On the



Fig. 8 Ratings distribution

**Table 2** Final objective rates representing the seriousness of imperfection types

	H.	Grade		H.	Grade
Incorrect	++	10	Inconsistent	+	7
Invalid	++	10	Dependent	+	6.5
Non sensical	0	10	Incomplete	+	6
Meaningless	0	10	Summarized	+	6
Irrelevant	0	9	Doubtful	+	6
Incoherent	++	8.5	Old	+	5
Distorted	++	8.5	Inaccurate	+	5
Conflicting	++	8	Possible	+	4
Random	++	8	Likely	++	3
Ambiguous	++	8	Vague	--	2.5
Biased	--	7.5	Confused	--	2.5
Unreliable	--	7.5	Probable	--	2.5
Missing	+	7	Necessary	--	2.5
Undecidable	++	7	Believable	++	1

Homogeneity of grades (H.)

++ Homogeneous grades

+ Rather homogeneous

-- Very heterogeneous

0 Very heterogeneous but clarified

other hand, participants also quite agreed to say that the following types were medially serious: *Out of date*, *Imprecise* and *Possible*.

### *Synthesis of rating*

When the ratings were homogeneous or quite homogeneous, imperfect types could be quantified by giving a rate equal to their ratings mean, rounded out to 0.5. For the types *Non sensical*, *Irrelevant* and *Meaningless*, regarding the explanations given by participants, we gave a high rating because if these type are shown in our map, that would mean they had been included and not put aside. We gave them a rate equal to the third quartile, in order to keep a high note but not the maximum, which seemed to keep a track of the diversity of the rates. Finally, for the types for which opinions were very different (6 types) the objectiveness of rating was less certain. Those that had a rate between 5 and 10 were given the rate of 7.5 and those that had a rate between 0 and 5 were given a rate of 2.5.

Results are presented in Table 2. Final ratings are located between 1 and 10. The mean is equal to 6.4 and median is equal to 7. This show that participants preferred rates above the mean. Imperfections were seen as globally serious for the analysis.

### 4.3.3 Practical Use in an Expert Context

We asked participants if they found it useful to take imperfection into account in their studies. The answer was very positive for all the participants except one, who

said that in theory it is useful, but in practice, it is enough to evaluate if a data is acceptable or not.

Participants were asked if they would use this classification either during their work or to communicate about imperfections to their clients. Answers were mainly positive (Fig. 9). Only two participants said they would probably not use it in their work and three others they would not use it to communicate to their clients. They explained that the classification was too complex (either for them or the client) or too detailed to be efficiently used. One said that in front of a client, talking about imperfection presented a risk of discrediting results.

For their own production, two participants who answered yes told that it would depend on the use case: they would use it only if the final result is a key issue. One participant thought that the classification could be more exhaustive, but she/he did not give concrete examples.

For their clients, three participants, who were all GIScientists, said that it would be useful if the client is the data provider, in order to give her/him a feedback. Two participants expressed doubts about the capacity of decision makers to understand some imperfection terms. One told that not every type is useful to be presented to decision makers: for example, inconsistent data would be dismissed before showing him the results. Regarding other contexts of use, participants proposed to use the classification to think about their work practices in groups, to evaluate data suppliers and share this knowledge with colleagues or to assess the reliability of their results.

Then participants were asked if they thought they would use a simplified version of the classification, if it existed. Answers without and with simplification of the classification were compared (Fig. 10). Participants were more ready to use a simplified classification in their work. Their answers evolved less positively to communicate imperfection to clients. It could mean that the classification complexity is not the explaining factors to its use for clients. Two participants gave a more negative answer with a simplified version. One of them was less interested in using a simplified classification in his work because she/he thought it would be less accurate. The other one who was less interested in using it with clients has not justified his/her answer and its overall comment was rather positive towards a simplified classification.

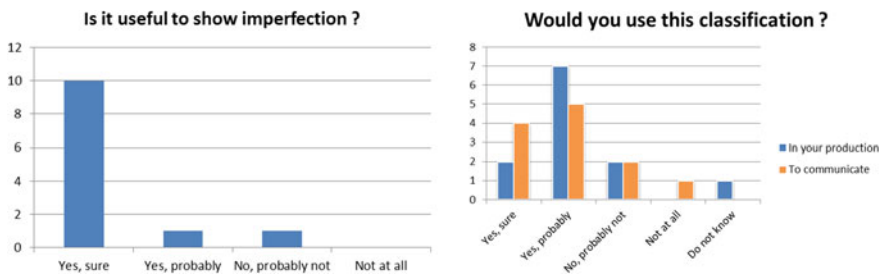
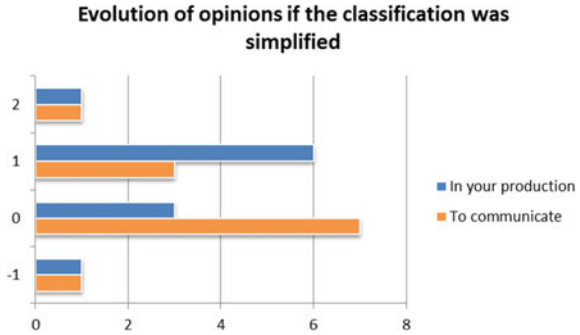


Fig. 9 Opinions of participants about a practical use of the classification



**Fig. 10** Difference of answers between the original and a simplified version of the classification. Positive results show a more positive opinion



#### 4.3.4 Overall Opinions of Participants About the Classifications

Finally, opinions about the classification were divided. Participants seemed pleased to learn something new and appreciated the internal reflection about the quality of their data that was triggered by the exercises. Three participants found the imperfection terms difficult to understand. Another one said that terms should be changed in more common ones. The majority of participants (7 out of 12) found the terms too much redundant. They quoted *possible*, *doubtful*, *probable* and *likely* as very near terms. To simplify the classification, one participant proposed to make small groups of near terms. Another one proposed to keep only the three main categories.

## 5 Discussion and Outlook

From a literature review, we have proposed a synthesis classification of imperfections affecting geohistorical data about past floods. The evaluation of this classification has shown that expert users understood the imperfection qualifying terms and mainly thought that it would be usable in their work and to communicate about imperfection with their clients. Expert users were more interested in using a simplified version of the classification in their work.

In order to include imperfections in a map, we have proposed to display an overall index of the amount of imperfection in a spatial object on the map and to give access to qualitative types of imperfection by user interaction. So, imperfection terms had to be quantified, in order to aggregate them as an overall index of imperfection. This quantification could be made through ratings, which represent the seriousness of each imperfection type for the analysis performed with this kind of data. The experiment has shown that it is possible to give an objective rating to the majority of imperfection types. The rates can be aggregated in two-step. Firstly, they can be aggregated by data dimension affected by imperfection, thanks to a weighted arithmetic mean, in which the weights are equal to the ratings, what give more importance to serious types of imperfection. Secondly, resulting ratings can be aggregated thanks to an equal-weighted mean to get the overall index.

A limit of this experiment was that it was not always clear why participants classified an imperfection term as 'not understood' or why they gave this specific rate to a type. In future work, to avoid this limit, we could employ a think-aloud method, in which participants express aloud all they think about, while they perform the exercises. Another limit is that we did not ask participants for rating imperfection affecting each dimension of information independently. For example, ratings may be different for spatial imperfections and temporal imperfections. These differences of impacts might be addressed thanks to the weights assigned during the ratings aggregation, what would be assessed in a future experiment.

The rating of imperfection types is probably domain- and expertise-dependent. In order to apply the proposed classification in another field of use, the evaluation method should be reproduced: from the main classification, identify imperfection types which are important for the considered users, then evaluate the weights that should be applied to each imperfection type and the relative importance of imperfect dimensions for the ratings aggregation.

In future works, a simplified version of the classification will be proposed to users: terms which have a close meaning and close rates may be grouped and least understood terms may be reformulated. Opinions of users will be collected, about this new version and compared with the exhaustive one. The quantification of imperfection types will be applied to our dataset about the impacts of historical floods and included in a map display. Our proposal of map visualization as well as the weights employed in the aggregation method will be tested with users.

**Acknowledgments** We would like to thank SNCF Company for their implication in this research project. Many thanks to the twelve railway experts who took a little of their working time to participate in the study.

## References

- Arnaud A (2009) Valorisation de l'information dédiée aux événements de territoires à risque. Une application cartographique et géovisualisation de la couronne grenobloise. Université Joseph Fourier, Grenoble
- Arnaud A, Davoine P-A (2011) Approche cartographique et géovisualisation pour la représentation de l'incertitude: Application à l'information dédiée aux risques naturels. *Rev Int Géomatique* 21:205–224. doi:[10.3166/rig.21.205-224](https://doi.org/10.3166/rig.21.205-224)
- Bodin X (2002) La représentation des incertitudes spatiales de la carte de localisation probable des avalanches
- Buttenfield BP (1993) Representing data quality. *Cartographica* 30:1–7. doi:[10.3138/232H-6766-3723-5114](https://doi.org/10.3138/232H-6766-3723-5114)
- Buttenfield BP (1988) Visualizing the quality of cartographic data. In: Third international geographical information systems symposium (GIS/LIS 88), San Antonio (USA)
- Cliburn DC, Feddema JJ, Miller JR, Slocum TA (2002) Design and evaluation of a decision support system in a water balance application. *Comput Graph* 26:931–949. doi:[10.1016/S0097-8493\(02\)00181-4](https://doi.org/10.1016/S0097-8493(02)00181-4)
- Griethe H, Schumann H (2006) The visualization of uncertain data: Methods and problems. In: Simulation and visualization, Magdeburg (Germany)

- ISO/Technical Committee 211 (2013) ISO 19157: Information géographique - Qualité des données. [http://www.iso.org/fr/home/store/catalogue\\_tc/catalogue\\_detail.htm?csnumber=32575](http://www.iso.org/fr/home/store/catalogue_tc/catalogue_detail.htm?csnumber=32575). Accessed 13 Jan 2015
- Kinkeldey C, Maceachren AM, Riveiro M, Schiewe J (2015) Evaluating the effect of visually represented geodata uncertainty on decision-making: systematic review, lessons learned, and recommendations. *Cartogr Geogr Inf Sci*. doi:10.1080/15230406.2015.1089792
- Lowell KE (1997) Outside-in, inside-out: two methods of generating spatial certainty maps. In: Second annual conference of GeoComputation, Dunedin (New Zealand)
- MacEachren AM, Robinson A, Hopper S et al (2005) Visualizing geospatial information uncertainty: what we know and what we need to know. *Cartogr Geogr Inf Sci* 32:139–160. doi:10.1559/1523040054738936
- MacEachren AM, Roth RE, O'Brien J et al (2012) Visual semiotics & uncertainty visualization: an empirical study. *IEEE Trans Vis Comput Graph* 18:2496–2505. doi:10.1109/TVCG.2012.279
- Nardo M, Saisana M, Saltelli A et al (2008) Handbook on constructing composite indicators: methodology and user guide
- O'Brien O, Cheshire J (2014) Mapping Geodemographic classification uncertainty: an exploration of visual techniques using compositing operations. In: Workshop “Visually-Supported Reasoning with Uncertainty”, GIScience 2014, Vienna (Austria)
- Olston C, Mackinlay JD (2002) Visualizing data with bounded uncertainty. *IEEE Symp Inf Vis* 2002 INFOVIS 2002 1–8. doi:10.1109/INFVIS.2002.1173145
- Pang AT (2008) Visualizing uncertainty in natural hazards. *Risk Assess Model Decis Supp* 14:261–294. doi:10.1007/978-3-540-71158-2\_12
- Plew B (2002) The nature of uncertainty in historical geographic information. *Trans GIS* 6 (4):431–456
- Pornon H (1992) Les SIG, mise en oeuvre et applications. Hermes Science Publications, Paris (France)
- Potter K, Rosen P, Johnson CR (2012) From quantification to visualization: a taxonomy of uncertainty visualization approaches. In: Uncertainty quantification in scientific computing. Springer, pp 226–249
- Saisana M, Tarantola S (2002) State-of-the-art report on current methodologies and practices for composite indicator development, Italy
- Schneiderman B (1996) The eyes have it: a task by data type taxonomy for information visualization. In: *IEEE Workshop on visual languages'96*, pp 336–343
- Secchia G, Cunty C, Chesneau É, et al (2014) Évaluer des modes de représentation cartographique de l'incertitude: Exemple d'utilisation de méthodes des sciences cognitives. In: Sageo 2014, Grenoble (France), p 5
- Skeels M, Lee B, Smith G, Robertson G (2008) Revealing uncertainty for information visualization. In: working conference on advanced visual interfaces—AVI '08. ACM Press, Napoli (Italy), pp 376–379
- Smets P (1997) Imperfect information: imprecision—uncertainty. In: Smets P, Motro A (eds) Uncertainty management in information systems. From needs to solutions. Kluwer Academic Publishers, Berlin, pp 225–254
- Snoussi M, Gensel J, Davoine P-A (2012) Extending TimeML and SpatialML languages to handle imperfect spatio-temporal information in the context of natural hazards studies. In: Gensel J, Josselin D, Vandenbroucke D (eds) Proceedings of AGILE'2012 conference. Springer, Avignon (France), pp 117–122
- Thomson J, Hetzler E, MacEachren AM et al (2005) A typology for visualizing uncertainty. In: Erbacher RF, Roberts JC, Grohn MT, Borner K (eds) SPIE, visualization and data analysis 2005. SPIE, San Jose (CA, USA), pp 146–157
- Veregin H (1989) Error modelling for the map overlay operation. In: Goodchild MF, Gopal S (eds) Accuracy of spatial databases. Taylor and Francis, pp 3–19
- Zoghalmi A, De Runz C, Akdag H, Pargny D (2012) Through a fuzzy spatiotemporal information system for handling excavation data. In: Gensel J, Josselin D, Vandenbroucke D (eds) International Agile'2012 conference. Springer, Berlin, pp 179–196

**Part IV**  
**Pedestrian and Vehicle Mobility**  
**in Smart Cities**

# Conflict in Pedestrian Networks

Jia Wang, Zena Wood and Mike Worboys

**Abstract** Encouraging pedestrian activity is increasingly recognised as beneficial for public health, the environment and the economy. As our cities become more crowded, there is a need for urban planners to take into account more explicitly pedestrian needs. The term that is now in use is that a city should be ‘walkable’. For route planning, whereas much attention has been given to shortest path, in distance or time, much less attention has been paid to flow levels and the difficulties they pose on the route. This paper considers problems posed by conflicting paths, for example cross-traffic. We use network centrality measures to make a first estimate of differing levels of conflict posed at the network nodes. We take special note of the role of collective motion in determining network usage. A small case study illustrates the method.

**Keywords** Walkability · Network centrality · Collective movement · Conflict

## 1 Introduction

Encouraging pedestrian activity is increasingly recognised as beneficial for public health, the environment and the economy. *Walkability* is the measure of how friendly an environment is to walking (City of Fort Collins 2011) and has become an important dimension in urban planning. A walkable city provides an accessible walking environment that encourages more pedestrian activity, thus providing the benefits noted in the first sentence.

---

J. Wang (✉) · Z. Wood · M. Worboys  
Old Royal Naval College, University of Greenwich, 30 Park Row,  
London SE10 9LS, UK

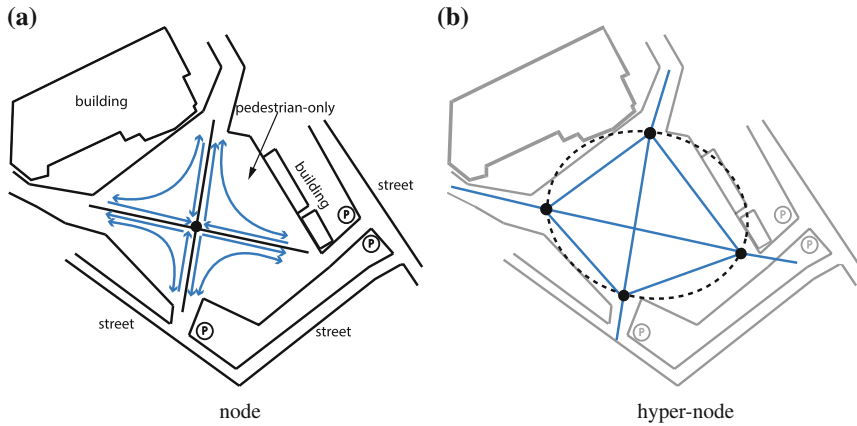
e-mail: J.Wang@greenwich.ac.uk

Z. Wood

e-mail: Z.Wood@greenwich.ac.uk

M. Worboys

e-mail: M.Worboys@greenwich.ac.uk



**Fig. 1** **a** Pedestrian system represented as a graph with conflicting motion occurring at the intersection. **b** The intersection is expanded as a hyper-node (*dotted circle*) to highlight the conflict

As part of a walkability quality measure, existing research has developed several methods that focus on crowding. Most methods are concerned with the physical properties of pedestrian flows (speed and volume) and walking areas (size). (See, for example, Gallin 2001). A certain level of crowding can make a street inviting and lively. However, there is a level at which crowding leads to congestion with conflicting motion creating an uncomfortable environment with poor walkability (Brierley 2013). When navigating an over-crowded environment, pedestrians need to constantly change their trajectories and speed to avoid conflicting motion. This can be stressful and can constrain walking (Fruin 1992; Handy 2005).

Our concern is to identify hot-spots where congestion or conflict provides obstacles to pedestrian movement. Within this paper the pedestrian system of walkways and intersections is modelled as an undirected graph, with walkways as edges and intersections as nodes. A method is presented where each node is assigned a resistance which indicates how much ‘energy’ a pedestrian might need to navigate through that node. These resistances are then used to compute the betweenness-centrality for each node to make a first estimate of differing levels of conflict posed at the network nodes. Conflict at a node would occur when paths cross and pedestrians have to navigate through potential collisions. We assume that the amount of conflict at a node is proportional to the number of crossing paths through the node. Figure 1 illustrates how a pedestrian environment could be represented as a graph. The blue paths within (a) show the different paths that a pedestrian may take through the node. The node is then expanded to a hyper-node (b) to clearly show the crossing paths and, therefore, where conflict will occur.

Analysing the common flows through the network is one way to identify hotspots where conflict is more likely to occur. A collective is a group of individuals that for some reason we wish to refer to as a single entity (Wood and Galton 2009). When considering the walkability of an area it is often more useful to consider the data at

the level of the collectives and not the level of the individuals. The evolving collective behaviour within a large group of individuals is often what an analyst requires (Andrienko and Andrienko 2007). Considering a large number of entities at the level of the collectives instead of the level of the individuals can also be computationally more efficient. Pedestrians can be grouped according to their travel purposes (Spaccapietra et al. 2008), origin and destination points (Andrienko and Andrienko 2008), or shared movement characteristics (Wood and Galton 2009; Dodge et al. 2008). This paper compares the proposed method when considering random flows versus collective motion of the network.

The paper continues in Sect. 2 by introducing related work on walkability, network centrality and collective movement. Section 3 describes the proposed method which is then applied in Sect. 4 to a network of popular pub crawls through Greenwich, London. The results of the implementation are discussed in Sect. 5. Future work is outlined in Sects. 6 and 7 concludes the whole paper.

## 2 Related Work

This section reviews related work from three areas: walkability measurement, focussing on congestion and conflicting movement, network centrality measures and collective movement.

### 2.1 *Walkability Measurement*

The term walkability originates from the transportation literature and has been used widely within urban planning, transportation and public health research to assess how environmental factors affect pedestrian walking behaviour (Frank et al. 2010; King et al. 2003; Owen et al. 2007; Saelens et al. 2003). The concept of walkability is defined as the quality of the walking environment perceived by pedestrians who live, shop and visit there (Abley 2005; Park 2008), or as “the measure of the extent to which the public realm provides for movement and other activity on foot in ways that are both efficient and enjoyable” (Transport for London 2005, p. 4). The operational definitions of walkability are provided at both macro and micro level, and are referenced to several components that can be observed and quantified (or qualified) on different spatial scales (Park 2008). On a macro-level (or neighbourhood level), the walkability measure is usually based upon the assessment of street pattern, land use diversity and housing density: A connected and accessible street pattern contributes to better walkability; increased land use diversity and housing density reduce the trip distance to amenities and increase pedestrian safety (King et al. 2003; Saelens et al. 2003). Frank et al. (2010) proposed a walkability index including four components: net residential density, retail floor area ratio, intersection density and land use mix. The walkability value is then calculated based on the values of these four components. On a micro-level (street or pedestrian level), walkability is

measured by developing and scoring multiple environmental indicators related to a local route. The Pedestrian Environmental Factor index defines and quantifies walkability by scoring the four indicators as ease of crossing, sidewalk continuity, local street characteristics and topography (1000 Friends of Oregon 1993). The pedestrian Level-of-Service (LOS) quantifies walkability by relating it to pedestrian facilities with regard to pedestrian flow (Gallin 2001). Other indices quantifying walkability on the micro-level include the Transit Friendliness Factor (Evans et al. 1997), the Walking Suitability Assessment (Emery et al. 2003) and the Irvine-Minnesota Inventory (Boarnet et al. 2006; Day et al. 2006).

Besides the above-mentioned quantitative methods, qualitative measures relating to subjective pedestrian perception have also been used to define and measure walkability. For example, Ewing et al. (2006) pointed out that using physical components only to measure walkability may not provide details relating to the walking experience in a particular environment. Ewing and Handy qualified walkability based on the ratings from a panel of urban design experts and then concluded five perceptual elements that determined walkability as imageability, visual enclosure, human scale, transparency, and complexity (Ewing et al. 2006). This approach was further developed to operationalise eight subjective perceptual qualities as imageability, enclosure, human scale, transparency, complexity, legibility, linkage and coherence in the context of commercial streets (Ewing and Handy 2009).

In general, pedestrians prefer to avoid contact with others except when overcrowding cannot be avoided. Existing studies of urban design and pedestrian behaviour proposed several approaches on defining and measuring crowding. LOS defined different levels to measure the quality of pedestrian flows based on both volumes and sidewalk or crosswalk area (Gallin 2001). Pedestrian Comfort Levels classified the level of comfort on the basis of the level of crowding experienced on the street, and the pedestrian crowding was measured in pedestrians per metre of clear footway width per minute (Transport for London 2010). Gehl and Gemzøe (2004) concluded that 13 people per meter per minute of footway was the maximum at which a comfortable level of quality can be delivered for footpath. Behavioural experiments involved personal space preferences into account and proposed the concept of minimum personal occupancy in dealing conflicting movement (Fruin 1992). Other human aspects included attaining normal walking speeds to avoid conflicts with other pedestrians (Fruin 1992). Existing research also pointed out that conflicting movements frequently happened in areas such as bus stops, tube exits, shopping centres and crossings (Transport for London 2010).

## 2.2 *Collective Movement*

When trying to improve an urban area it is important to try to satisfy as many people as possible and, therefore, you need to identify the needs of the 'collective population' (McArdle et al. 2014). Identifying areas or locations of common interest is a common goal in movement pattern analysis (Wood 2014). The locations of interest



could be origin and destination points (Andrienko and Andrienko 2008) or those that are frequently visited (McArdle et al. 2014). Analysing origin and destination points within travel behaviour can help identify the need for new facilities, such as bus stops. Groups that share similar movement patterns could also be of interest. Such information could be obtained from a spatiotemporal dataset via clustering, aggregation and similarity calculations (Wood 2014).

Focusing on general trends and collectives can be computationally more efficient especially when dealing with increasingly large datasets. Andrienko and Andrienko (2008) used data aggregation to visually analyse traffic data within Milan. Collective movement is defined as a function that relates the set of moving entities over the set of possible time moments and positions in space. The data can be viewed in two ways: *trajectory-oriented view* and *traffic-oriented view*. The former view groups the trajectories and the latter considers the possible ‘traffic situations’. The view adopted will depend on the goals of the analyst. Aggregation methods are suggested that group the trajectories according to origin and destination points, the points visited on a journey and the similarity of routes. This approach was taken further by Wang et al. (2015) in an analysis of Eulerian and Lagrangian perspectives on motion.

Many types of collective can be considered in pedestrian movement (Schadschneider et al. 2002). These phenomena often arise due to the interactions and behaviour characteristics of the pedestrians, particularly self-organisation. Schadschneider et al. (2002) used a cellular automaton (CA) to simulate these interactions to obtain the observed collectives. Four types of collective phenomenon were considered: *jamming*, *lane formation*, *oscillations* and *panic*. Jamming occurs when a blockage is encountered by pedestrians due to a lack of space or contradictory flow. Counter-flows can result in the formation of lanes and observed self-organisation. When there is a blockage and an individual manages to make their way through, it becomes easier for other individuals to follow them. This continues until someone makes their way through in the opposite direction. This repeated pattern is what the term oscillation refers to. Panics occur where the movement is counter-intuitive as a result of some situation (e.g., the wish for faster motion actually results in a slower moving crowd).

Andrienko and Andrienko (2007) referenced two types of collective behaviour that an analyst may wish to focus on: Momentary Collective Behaviour (MCB) and Dynamic Collective Behaviour (DCB). The former focuses on a set of entities at a particular moment of time; the latter focuses on multiple entities over a given temporal period. Four categories are identified that could influence the movement of entities including the activities and properties of the moving entities (e.g., how they move). Different patterns are specified that may be of interest and relevance to analysing DCB, one of which is *co-location in space* (i.e., when the paths followed by the observed entities contain at least some of the same positions). The positions visited can be further analysed according to any variation occurring in the order of the locations visited. This movement pattern could be of relevance to a network, and identifying hotspots of conflict, if each node was considered as a location of interest.

### 2.3 Network Centrality

Freeman (1977) proposed a collection of *centrality measures* that indicate degrees to which nodes have significance in a network. Let  $G$  be a graph, where  $N$  is the number of nodes and  $E$  is the number of edges. The *betweenness centrality* of a node of  $G$  gives a measure of how much the node is an intermediate point on paths in  $G$ . To be more precise, the betweenness centrality of node  $n$  is proportional to the number of shortest paths in  $G$  that pass through  $n$ . Formally:

$$B_n = \frac{1}{(N-1)(N-2)} \frac{\alpha_n}{\beta_n}, \quad (1)$$

where  $B_n$  is the betweenness centrality of node  $n$ ,  $\alpha_n$  is the number of shortest paths between any two nodes (except  $n$ ) in the graph passing through  $n$ , and  $\beta_n$  is the number of shortest paths between any two nodes (except  $n$ ) in the graph.

There is a body of research that has applied betweenness and other centrality measures to networks of urban streets. Varoudis et al. (2013) evaluated the angular betweenness measure of space syntax of urban streets implemented by two different methods, *Tasos* and *depthmapX*. The *depthmapX* method is based on the cognitive-search-agent with pedestrian walking constraints. The *Tasos* method is based on mathematical shortest path without the pedestrian walking constraints used by the *depthmapX* method. The evaluation showed that these two methods offered similar results in terms of pedestrian movement but the *Tasos* method was more computationally efficient.

Kazerani and Winter identified in (2009) the issues (e.g., no travel behaviour and temporal constraints) of using betweenness centrality in predicting traffic flow in reality and suggest some minor amendments to the classic centrality measure. The paper proposed a modified betweenness centrality method, which reflected locations of origins and destinations of recorded trips. Compared to the classic method, the proposed method was better in predicting “actual traffic counts over given time intervals” (Kazerani and Winter 2009, p. 8).

Crucitti et al. (2006a, b) studied centrality in urban streets for 18 different world cities and developed a comparative analysis of different centrality measures of urban streets. The centrality measures were based on a new approach called *multiple centrality assessment* (MCA) developed by Porta et al. (2008) for the centrality analysis in geographic systems. MCA was based on primal graphs, a set of centrality indices as well as a fully metric computation of distances. The results of Crucitti’s study indicated that the centrality measures with the four indices (closeness, betweenness, straightness and information) allowed extended visualisation and characterisation of city structures.

Based on the MCA centrality analysis, Porta’s group used the northern Italian city Bologna as study area to investigate the relationship between street centrality and densities of commercial and service activities in the city (Porta et al. 2009). The result highlighted a strong correlation between street centrality (particularly

betweenness centrality) and locations of shops and services at the neighbourhood scale. Following the same MCA method, Produit's Master thesis (Produit 2009, p. 96) reports a new GIS tool that is able to create three indexes of network density estimation of activities, network density of edges with population made from centralities and diversity of activities along the network. The centrality indexes (closeness, betweenness and straightness) were computed at both global and local level to characterise the shape of network.

While most centrality studies focused on degree, closeness, betweenness and eigenvector measures to determine who occupied critical positions in the network, a rarely answered question was about the correlation of the four centrality measures (Valente et al. 2008). Valente et al., empirically examined the correlation among the four centrality measures and found that they were strongly correlated. Their study also revealed the association of network properties such as density and reciprocity to the correlation of different centrality measures.

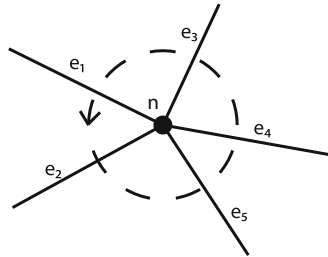
### 3 Method

There are problems with directly applying centrality measures to congestion and conflict in urban flow. Firstly, existing methods take little or no account of spatial and temporal variation in flows in the network. For example, direct application of the betweenness centrality measure to indicate those nodes which are more likely to be congested would require the assumption that flows are uniform in the network and that no one path is more used than another. Our work will generalise betweenness centrality to account for preferred paths followed by collective motion against a background of random noise in network flow.

The second problem is that no account is taken of conflict at nodes where paths cross and so pedestrians must negotiate through potential collisions. We make a simplifying assumption that the amount of conflict at a node is dependent upon the number of crossing paths through that node.

#### 3.1 *Rotation Graph and Crossing Paths*

In order to generalise previous constructions, it is necessary to take into account the embedding of graph  $G$  in Euclidean space. We assume that  $G$  is embedded in a surface and that each node is specified by a pair of coordinates. The embedding gives us information not only about the positions of nodes but also the positions and orientations of the edges. If we consider a single node  $n$  of  $G$ , then those edges that are incident with  $n$  are incident in a unique rotational cycle about  $n$ . This is shown by the example in Fig. 2, where the five edges  $e_1, e_2, \dots, e_5$  are incident with  $n$  in the specific cyclic order  $(e_1, e_2, e_5, e_4, e_3)$ . If graph  $G$  has the property that each of its nodes is equipped with a rotation cycle of edges around it, then the graph is termed a *rotation graph* (Gross et al. 2013, p. 741).



**Fig. 2** Cycle of edges  $e_1, e_2, \dots, e_5$  around node  $n$

Because the pedestrian network is modelled here as a graph embedded in Euclidean space, it is automatically a rotation graph. We can use the extra information provided by the cycles to give a better approximation of the energy required to traverse a path on the route taking into account conflicts due to crossing paths.

For example, consider again Fig. 2. In travelling through the node along edges  $e_1$   $e_5$ , say, the traveller would potentially have to cross the paths of pedestrians on routes  $e_2 - e_3$  and  $e_2 - e_4$ . (This is not taking into account direction of travel). Depending on the volume of traffic on the latter two routes, our traveller will meet more or less amounts of conflict. So, assuming we know the amounts of flow on the different paths, we can calculate the amount of resistance that each node on the path might add to the journey. We use this resistance to modify the travel distance between nodes to get a more accurate assessment of the betweenness coefficients.

In the general case, if the route through the node passes through edges  $e$  and  $f$ , then the number of crossing paths is  $i * j$  where  $i$  and  $j$  are the number of edges strictly between  $e$  and  $f$  in the cycle of edges, counting clockwise and anti-clockwise, respectively.

### 3.2 Generalising Betweenness

As has been said, the betweenness centrality measure of a graph was originally conceived as a purely topological measure, not taking account of the graph embedding. However, as it involves shortest path computations, it is easy to extend to the case where path is defined between nodes using Euclidean distances along edges. Our contribution is to take into account node resistance to a path resulting paths that cross the path of interest. In the extreme case, we can neglect edge lengths and focus entirely on node resistances. As a very simple example, consider the network shown in Fig. 3. Suppose we wish to calculate the path length  $AD$ . Then we need to take into account the crossing flows  $\phi_{FC}, \phi_{EC}, \phi_{FB}, \phi_{EB}$ . Just how much these cross-currents should contribute to the path length depends on the flow quantities and the level of disruption they course. However, once this has been calibrated, we can perform a Dijkstra-type computation for shortest paths, and from this calculate the betweenness centrality measures for nodes.

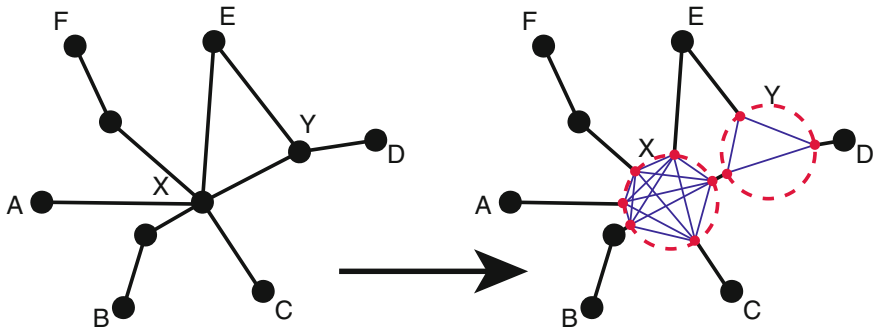


Fig. 3 Small network example

It is then interesting to compare the betweenness measures in three cases:

1. The traditional case with edge lengths calculated from geographic distances between nodes;
2. The case where edge lengths are ignored but node resistances determine path lengths. This leads to two sub-cases:
  - a. A random collection of flows is assumed through the network so all crossings are assumed to impart equal resistance
  - b. Flows through the network arise from collectives passing through it. In this case, crossings are weighted by the flows through them.

In the next section we work through a case study showing how these constructions can be applied to a specific instance.

## 4 Movement Data Analysis

Two datasets have been used to demonstrate the proposed method, both of which relate to pubs in Greenwich. Although neither dataset records the known movements of multiple individuals, we can demonstrate how the method could be used to identify conflict at nodes and show why collective movement should be considered.

### 4.1 The Case Study

A dataset containing the locations of 28 pubs within Greenwich has been used to produce the pedestrian network. The underlying road network has been used to form the network shown in Fig. 4. Two pubs are considered connected if it is possible to travel from one to the other without coming across a third. Each node, depicted by a red circle, represents a pub and each edge, depicted by a black line, represents a connection between them.

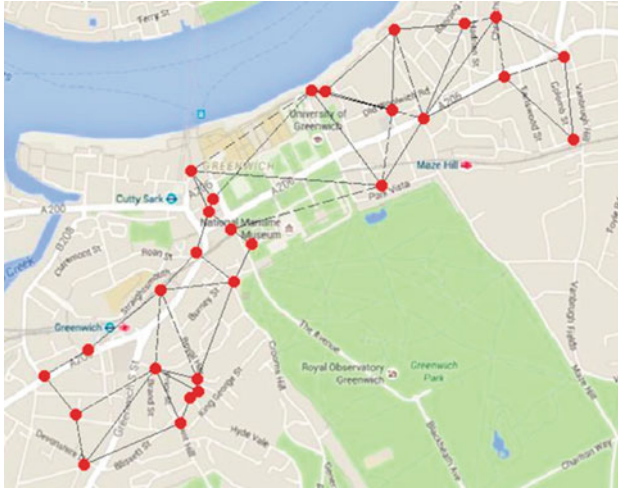


Fig. 4 The network of 28 pubs in Greenwich overlaid on a Google map

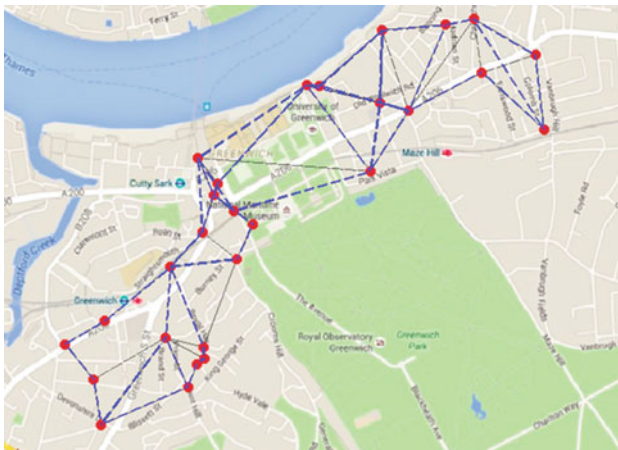


Fig. 5 The 10 pub crawls within Greenwich overlaid on a Google map

Known pub crawls within Greenwich have been identified to model the collective movement within the network (Fig. 5). A pub crawl typically involves a group of individuals visiting a collection of pubs and having a specified number of beverages in each of them. Although unrealistic, it is assumed that collectives do not spend a significant amount of time at each pub and pass quickly through them. This allows the simulation of pedestrians passing through nodes where there is conflict. Ten pub crawls have been identified to model 10 collectives moving around the network. Dijkstra's algorithm has been used to calculate the shortest path between two consecutive pubs on a particular crawl that are not joined via an edge in the network.

### 4.2 Implementation

The method has been implemented using Matlab. The program reads in two .csv files: one that stores the locations of each pub, recorded by their longitude and latitude values; and, one that records the order of pubs visited on each of the 10 pub crawls. The edges of the network are identified manually using a map of the underlying road network and stored using an adjacency matrix. If there is an edge connecting nodes  $n_i$  and  $n_j$  a 1 is placed in position  $(i, j)$  of the matrix, otherwise the value will be 0. A second adjacency matrix is computed that stores the geographic distance between each pair of connected nodes. For each node all possible paths that run through it are identified and, for each path, the number of cross flows is calculated. The cross-flow values are combined to produce a value for the node’s overall resistance.

When only considering geographic distance to calculate shortest path, each edge is weighted with the distance between the two pubs that it connects. The weighting on each edge, when only considering resistance, is the combined value of the resistance contributed by the two nodes that it connects. For example, in Fig. 2, assuming a random collection of flows so all crossings impart equal resistance, the node would have a resistance of 10. Since the node connects five edges, the node would contribute a resistance of two to each of the five edges. The remaining weight of the edges would be the relevant resistance value from the other node that it connects. When considering collective motion on the network the node resistance is recalculated. Instead of considering all possible paths through a node only the paths that are used by the collectives are considered.

A second method has been adopted to distribute the node resistance between the edges when considering collective motion. This method distributes the node resistance proportionally between connecting edges according to usage. For example, Fig. 6 shows the usage of the collectives on the edges passing through a node (node A) within the network. Ten collectives pass through the node but the edges are not used equally. The node has a resistance value of two. Proportionally distributing this value amongst the connected edges would contribute a resistance value of 0.6

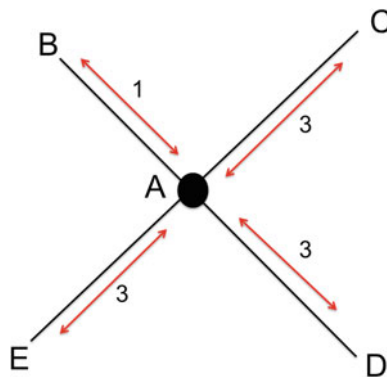


Fig. 6 An example of a node on the path of 10 collectives



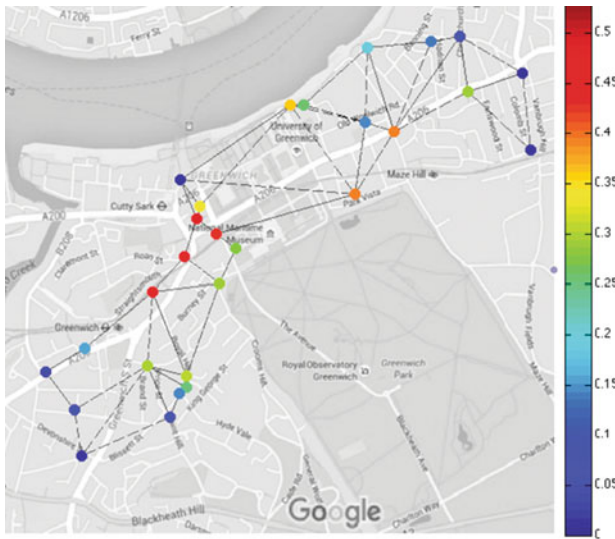
to edges  $\phi_{AC}$ ,  $\phi_{AD}$  and  $\phi_{AE}$ ;  $\phi_{AB}$  would gain a resistance value of 0.1. The remaining weighting of the edges would come from the other node that they connect. This second distribution method has been included for comparison to see the affect that it has.

Once the weights for each edge have been obtained, either using geographic distance or node resistance, the betweenness centrality of each node is calculated. For each of the 28 nodes in the network, an implementation of Dijkstra’s algorithm is used to identify the number of shortest paths between each pair of nodes, excluding the current node being considering ( $n$ ), that pass through the node. The values are normalised by dividing by the possible number of shortest paths between two nodes in the network (not considering the current node  $n$ ).

### 4.3 Results

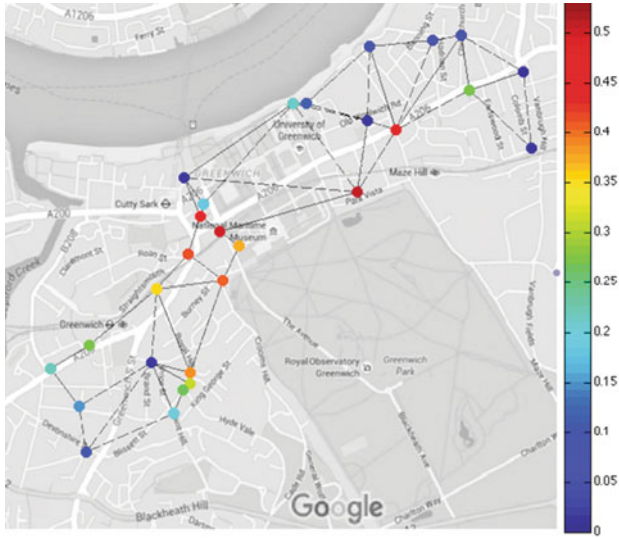
Figures 7, 8, 9 and 10 show the results of applying the proposed method to the sample dataset. In each figure a heat scale is used where red indicates the highest centrality value and blue the lowest. The same scale has been used for each figure to allow direct comparison of the results.

Figure 7 shows the betweenness centrality value for each node when only considering the geographic distance to calculate shortest path length. Figure 8 shows the betweenness centrality values when using node resistance to calculate shortest path length. A random collection of flows is assumed through the network so all

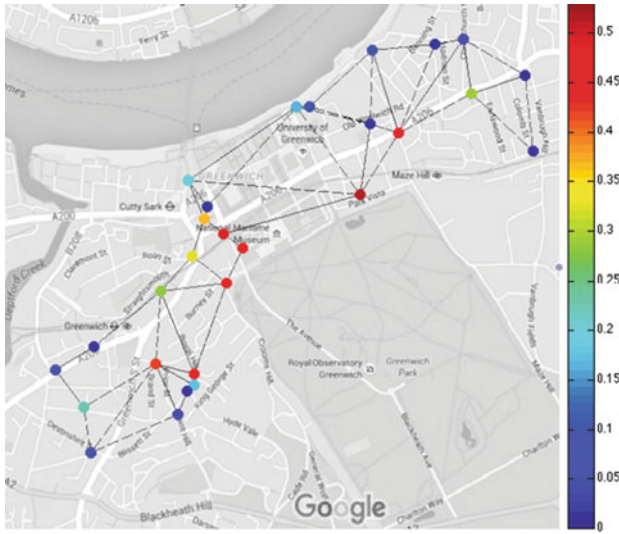


**Fig. 7** The betweenness centrality of each node shown with a heat scale (*red* indicates the highest centrality value and *blue* the lowest) when edges are weighted according to distance





**Fig. 8** The betweenness centrality of each node shown with a heat scale (*red* indicates the highest centrality value and *blue* the lowest) when edges are weighted according to node resistance assuming all crossings impart equal resistance



**Fig. 9** The betweenness centrality of each node shown with a heat scale (*red* indicates the highest centrality value and *blue* the lowest) considering collective motion with node resistance distributed equally amongst connecting edges



**Fig. 10** The betweenness centrality of each node shown with a heat scale (*red* indicates the highest centrality value and *blue* the lowest) considering collective motion with proportionally distributed node resistance

crossings impart equal resistance. Figures 9 and 10 show the betweenness centrality of each node when considering collective motion using node resistance to calculate shortest path length. Within Fig. 9 node resistance is distributed equally amongst the connecting edges; Fig. 10 shows the results of proportional distribution of node resistance.

## 5 Discussion

The application of the proposed method to a sample dataset has shown that a different set of nodes within the network is identified as key (i.e., coloured red) when considering resistance and collective motion compared with general, random motion. This section discusses the results and aspects that have been highlighted by the analysis for further consideration.

In Fig. 7, where geographic distance was used to calculate shortest path, four nodes are shown in red and thirteen are in a shade of blue. The highlighted nodes are, as expected, the ones central within the network. The majority of the blue nodes lie on the outside edges of the network. When considering node resistance with a random flow on the network, assuming all crossings impart equal resistance, (Fig. 8) four nodes are again highlighted but these are not the same as though highlighted when considering geographic distance alone. Some nodes that were previously coloured yellow (mid-to high betweenness centrality measures) in Fig. 7 are now coloured blue. The nodes shown as having the highest centrality measures within Fig. 8 can

still be considered as quite central within the network but not all outer edges have the lower centrality measures. Figures 9 and 10, where collective motion is considered, again show a different set of highlighted nodes. By allowing the key paths (i.e., the collective movement) to be considered, some of the nodes that were previously considered to be a point where conflict is likely to occur are now seen to be of little concern in Fig. 9. It is also possible to quickly identify the main hotspots where there is likely to be conflict. Where the node resistances have been distributed equally between connecting edges (Fig. 9), there are many more nodes highlighted blue (i.e., with low betweenness centrality measures). However, there are some nodes, previously considered of low importance, that are now coloured in red. The spread of red nodes across the network is larger. Figure 10 shows what happens when node resistance is spread proportionally amongst connecting edges. A smaller number of nodes are highlighted as ones where conflict is likely to occur.

The dataset that we have used is only a sample and the numbers within each collective have not been taken into consideration. Future work will need to be undertaken to extend the method to take account of the flows within the network. This information could then be used to consider a network where an event is occurring such as a carnival, protest or major sporting event. In addition to showing where walkability may be affected, emergency services and event planners could also use the results of our method to identify where they may focus their resources during the event.

The dataset does not include any temporal information. It is likely that flows within the network will vary over time. Different collectives may only exist at specific times, which would affect the identified hotspots. For example, commuters travel at peak times. The method should be extended to show the temporal variation in the betweenness values of each node. The dataset to which it is applied also would need the necessary temporal information.

## 6 Future Work

Betweenness centrality is just one of several measures that might be used as an indicator of importance of a node in a network. Application and consideration of other measures, such as ‘information centrality’, are currently being investigated. Two methods were used to distribute the node resistance around the connecting edges. Both methods highlighted different sets of key nodes but further analysis is required to see which is the most meaningful and efficient when used in a larger network.

Similar to the walkability measure described in Sect. 2, resistance in a particular walking environment can be measured in two ways. One is to use the physical components of the walking space (e.g., capacities of intersections and pathways) in addition to network flows to decide different levels of resistance. The second approach is similar to Ewing and Handy’s work (Ewing et al. 2006) in which qualifies resistance based on subjective qualities to determine resistance from the human perspective. This is an area for future work.

The graph used within the analysis is a simplification of the underlying road network. Further analysis should be carried out that takes into account more detail of the real-world network. Although based on real locations and common paths, the dataset in this paper could be considered synthetic. Work has begun in identifying a dataset that has the recorded movements of a large number of pedestrians in a mapped environment. From this the collectives will be identified using existing methods (e.g., those described in Wood (2014)).

## 7 Conclusion

This paper has presented research on issues arising from conflicting motion in networks. Although many of the methods are generalisable to any transportation system, we have focused on pedestrian networks. Conflict arises at network nodes when paths cross, and a reasonable assumption made in this paper is that the more potential for paths to cross, both in terms of the number of possible routes through the node but also in terms of the amount of flow, the more potential for conflict there arises. Collective motion has an important influence on network flows. A key methodology adopted in this work is to use network centrality and, in particular, betweenness centrality to estimate the nodes in the network with the most potential for conflict. In traditional network centrality approaches, path weights are measured by length, either spatial or temporal. A key contribution here is to measure path weight by ‘resistances’ that conflicting paths induce. This resistance is determined by flows along conflicting paths, which in turn is determined by collective motions through the network.

In our case study of routes through pubs in Greenwich, the pubs were the nodes and the network was the collection of possible paths from one pub to the next. For the sake of this experiment, we took the extreme position of only considering conflict in the weighting of paths, and so distances were neglected. Betweenness centrality of the nodes was computed in two cases:

- Flow through the network was considered to be random.
- Flow through the network was determined by collective motion induced by the pub crawls.

The results clearly demonstrated the expected results that collective motion through a network strongly influenced the resistance of nodes to paths, and hence altered which nodes became key, as measured by betweenness centrality.

**Acknowledgments** This study was supported by the University of Greenwich Faculty of Architecture, Computing and Humanities Research and Enterprise Fund through the project “Spatial Informatics for the Dynamic Smart City”.

## References

- Abley S (2005) Walkability scoping paper. <http://www.levelofservice.com/walkability-research.pdf>
- Andrienko G, Andrienko N (2007) Extracting patterns of individual movement behaviour from a massive collection of tracked positions. In: Gottfried B (ed) Workshop on behaviour modelling and interpretation, Germany, pp 1–16
- Andrienko G, Andrienko N (2008) Spatio-temporal aggregation for visual analysis of movements. In: IEEE symposium on visual analytics science and technology (VAST 2008), Columbus, Ohio, USA. IEEE Computer Society Press, pp 51–58
- Boarnet MG, Day K, Alfonso M, Forsyth A, Oakes M (2006) The Irvine-Minnesota inventory to measure built environments: reliability tests. *Am J Prev Med* 30(2):153–159
- Brierley K (2013) The effects of pedestrian delay and overcrowding on our streets and the rationale for shorter blocks and through blocks links. A report prepared for the city of Melbourne City of Fort Collins, Colorado (2011) Pedestrian Plan. <http://www.fcgov.com/transportationplanning/pedplan.php>
- Crucitti P, Latora V, Porta S (2006a) Centrality in networks of urban streets. *Chaos: Interdisc J Nonlinear Sci* 16(1):015113
- Crucitti P, Latora V, Porta S (2006b) Centrality measures in spatial networks of urban streets. *Phys Rev E* 73(3):036125
- Day K, Boarnet M, Alfonso M, Forsyth A (2006) The Irvine-Minnesota inventory to measure built environments: development. *Am J Prev Med* 30(2):144–152
- Dodge S, Weibel R, Lautenschütz AK (2008) Towards a taxonomy of movement patterns. *Inf Vis* 7:240–252
- Emery J, Crump C, Bors P (2003) Reliability and validity of two instruments designed to assess the walking and bicycling suitability of sidewalks and roads. *Am J Health Promot* 18(1):38–46
- Evans J IV, Perincherry V, Douglas G III (1997) Transit friendliness factor: approach to quantifying transit access environment in a transportation planning model. *Transp Res Rec: J Transp Res Board* 1604:32–39
- Ewing R, Handy S, Brownson RC, Clemente O, Winston E (2006) Identifying and measuring urban design qualities related to walkability. *J Phys Act Health* 3:S223
- Ewing R, Handy S (2009) Measuring the unmeasurable: urban design qualities related to walkability. *J Urban Des* 14(1):65–84
- Frank LD, Sallis JF, Saelens BE, Leary L, Cain K, Conway TL, Hess PM (2010) The development of a walkability index: application to the neighborhood quality of life study. *Br J Sports Med* 44(13):924–933
- Freeman LC (1977) A set of measures of centrality based on betweenness. *Sociometry* 35–41
- Friends of Oregon (1993). Making the land use transportation air quality connection-The pedestrian environment. <http://ntl.bts.gov/DOCS/tped.html>
- Fruin J (1992) Designing for pedestrians. Public Transportation United States
- Gallin N (2001) Quantifying pedestrian friendliness-guidelines for assessing pedestrian level of service. *Road Transp Res* 10(1):47
- Gehl J, Gemzøe L (2004) Public spaces-public life. Van Nostrand Reinhold, New York
- Gross JL, Yellen J, Zhang P (eds) (2013) Handbook of graph theory, 2nd edn. CRC Press
- Handy S (2005) Critical assessment of the literature on the relationships among transportation, land use, and physical activity. Transportation Research Board and the Institute of Medicine Committee on Physical Activity, Health, Transportation, and Land Use. Resource paper for TRB Special Report, 282
- Kazerani A, Winter S (2009) Can betweenness centrality explain traffic flow. In: Proceedings of the 12th AGILE international conference on GIS
- King WC, Brach JS, Belle S, Killingsworth R, Fenton M, Kriska AM (2003) The relationship between convenience of destinations and walking levels in older women. *Am J Health Promot* 18(1):74–82

- McArdle G, Demsar U, van der Spek S, McLoone S (2014) Classifying pedestrian movement behaviour from GPS trajectories using visualization and clustering. *Ann GIS* 20(2):85–98
- Owen N, Cerin E, Leslie E, Coffee N, Frank LD, Bauman AE, Hugo G, Saelens BE, Sallis JF (2007) Neighborhood walkability and the walking behavior of Australian adults. *Am J Prev Med* 33(5):387–395
- Park S (2008) Defining, measuring, and evaluating path walkability, and testing its impacts on transit users' mode choice and walking distance to the station. ProQuest
- Porta S, Crucitti P, Latora V (2008) Multiple centrality assessment in Parma: a network analysis of paths and open spaces. *Urban Des Int* 13(1):41–50
- Porta S, Latora V, Wang F, Strano E, Cardillo A, Scellato S, Messori R (2009) Street centrality and densities of retail and services in Bologna, Italy. *Environ Planning B: Planning Des* 36(3):450–465
- Produit T (2009) A novel GIS method to determine an urban centrality index applied to the Barcelona metropolitan area. Master dissertation (No. EPFL-STUDENT-175634)
- Saelens BE, Sallis JF, Black JB, Chen D (2003) Neighborhood-based differences in physical activity: an environment scale evaluation. *Am J Public Health* 93(9):1552–1558
- Schadschneider A, Kirchner A, Nishinari K (2002) CA approach to collective phenomena in pedestrian dynamics. In: Proceedings of the 5th international conference on cellular automata for research and industry, ACRI 2002 Geneva, Switzerland, October 9, pp 239–248
- Spaccapietra S, Parent C, Damiani L, de Macedo JA, Porto F, Vangenot C (2008) A conceptual view on trajectories. *Data Knowl Eng* 65:126–146
- Transport for London (2005). Improving walkability. <http://content.tfl.gov.uk/tfl-improving-walkability.pdf>
- Transport for London (2010). Pedestrian comfort guidance for London. <https://www.centro.org.uk/media/424876/Appendix-8.PDF>
- Valente TW, Coronges K, Lakon C, Costenbader E (2008) How correlated are network centrality measures? *Connections (Toronto, Ont.)* 28(1):16
- Varoudis T, Law S, Karimi K, Hillier B, Penn A (2013) Space syntax angular betweenness centrality revisited. In: Ninth international space syntax symposium, vol 57, pp 1–16
- Wang J, Duckham M, Worboys M (2015) A framework for models of movement in geographic space. *Int J Geogr Inf Sci*. doi:10.1080/13658816.2015.1078466
- Wood Z (2014) What can spatial collectives tell us about their environment? In: IEEE symposium on computational intelligence and data mining (CIDM), 2014. Orlando, FL, pp 329–336
- Wood Z, Galton A (2009) A taxonomy of collective phenomena. *Appl Ontology* 4(3–4):267–292

# Personalizing Walkability: A Concept for Pedestrian Needs Profiling Based on Movement Trajectories

David Jonietz

**Abstract** Recently, location-based navigation systems have evolved from purely car-centered services to incorporate route planning for other means of transportation, such as walking or cycling. In this context, a particular challenge lies in the computation of optimal routes for pedestrians, who expect a high walkability of their urban environment. In particular, route computation should explicitly incorporate such specific infrastructural needs but also acknowledge the heterogeneity of pedestrians. Thus, this research proposes a concept to create individual pedestrian user profiles based on their pre-recorded movement trajectories. This information is then used for the evaluation of the expected personalized walkability of urban areas. By exemplarily applying the method to a real-world scenario, its usefulness is demonstrated.

**Keywords** Personalization • Walkability • Movement trajectory

## 1 Motivation

Recently, location-based navigation systems have evolved from purely car-centered services to incorporate route planning for cyclists or pedestrians as well. Despite the high number of existing commercial and non-commercial applications (e.g. HERE Maps<sup>1</sup> or Google Maps<sup>2</sup>), however, the challenges are still manifold (Millonig 2006; Huang et al. 2014). Thus, for instance, existing road network data is usually tailored to motorized means of transportation, whereas the movement of pedestrians

---

<sup>1</sup><https://maps.here.com>.

<sup>2</sup><https://maps.google.com>.

---

D. Jonietz (✉)

Institute of Geography, University of Heidelberg, Heidelberg, Germany  
e-mail: david.jonietz@uni-heidelberg.de



and, to a lesser degree, cyclists is typically less network-bound and might involve short cuts through malls or public parks (Walter et al. 2006). Apart from the data itself, problems are also encountered in terms of accurate positioning, especially in indoor environments (Mautz 2009).

In the context of this study, however, the focus is put on another challenge, the computation of optimal routes for pedestrians. Thus, whereas for the majority of car drivers, it is fully satisfactory if a system provides the shortest, fastest or most fuel-efficient route, pedestrians, due to their systemic characteristics, demonstrate a more complex route choice behavior, and pose specific needs to their urban environment (Saelens et al. 2003; Millonig 2006; Agrawal et al. 2008). Since the early 1990s, the superordinate term walkability, or to be more precise micro-scale walkability, has been established to describe the degree to which the built urban environment meets the demands of pedestrians, or the “quality of walking environment perceived by the walkers” (Park 2008, p. 22). In general, there are two challenges to computing optimal routes for pedestrians: First, despite the fact that empirical research has demonstrated the richness and variability of potential environmental influences on pedestrian behavior, and also their manifestations in route choice processes, the vast majority of applications still focus on computing the shortest or fastest path, and do not take into account any notion of walkability (Millonig 2006; Gartner et al. 2011). Exceptions are Völkel and Weber (2008) or Holone et al. (2007), who enable users to annotate geographic data, and use these ratings for routing, or Huang et al. (2014), who choose a crowd sourcing approach for collecting people’s affective responses to the environment and computing according optimized routes. Second, in contrast to the recognition that pedestrians are heterogeneous in terms of their physical ability, social roles and economic constraints, in terms of analyzing their infrastructural needs, they are still treated as a homogeneous group (Buchmueller and Weidmann 2006).

Of these, the latter challenge is particularly relevant with regards to the personalization of services and applications, a topic which has received much attention in the broader context of location-based services (LBS) (Krisp and Keler 2015). Providing personalized routes to pedestrians which explicitly incorporate their individual infrastructural needs would therefore be a desirable extension of current approaches. A first step, however, implies the development of a concept to compute personalized walkability maps, which then serve as a basis for route calculation. Against this background, this study proposes to use pre-recorded movement trajectories of individual pedestrians to infer their specific infrastructural needs, and use these derived user profiles to evaluate the expected personalized walkability of urban areas. For this, using a general conceptual framework to compute personalized suitability values proposed earlier (Jonietz et al. 2013; Jonietz and Timpf 2013) as a theoretical foundation, a detailed model of the action *WALKING* is developed based on a review of empirical studies on walkability. An exemplary application of the general concept to a real-world study area is provided.

This paper is structured as follows: The following chapter provides the theoretical background, and starts with a summarizing review of the empirical literature on walkability, which will form the basis for the action model to be developed.



Then, a conceptual framework to modeling personalized suitability is briefly recapitulated. In the next chapter, our method is described which is then applied to an exemplary real-world scenario. Finally, the results are discussed and the paper is concluded.

## 2 Background

This chapter presents the background for this study, starting with a brief review of the state of empirical knowledge on walkability. In the following, a method to compute personalized suitability values from spatial data is recapitulated.

### 2.1 Empirical Research on Micro-Scale Walkability

In accordance with the results of empirical research, walkability indicators, selected attributes of the built environment from which its walkability can be inferred, exist on different scale-levels, ranging from the micro-scale of the individual street or vista to the macro-level of the entire city or region (Handy 2004; Cervero and Kockelman 1997). Until today, an extensive body of literature has developed on the topic of walkability, with the majority of studies focusing on the identification of macro- and micro-scale environmental as well as pedestrian-related determinants of the observed walking behavior.

Due to the high number of relevant studies, a review of reviews approach was chosen to identify walkability indicators. A literature search led to the identification of 51 review papers of which 18 fitted our inclusion criteria, namely that they reviewed the empirical research in a systematic rather than a narrative way, reported results for walking for transportation purposes, and included micro-scale walkability indicators. Table 1 lists the environmental variables reported as significantly related to walking outcome.

A more detailed picture of the effect of micro-scale design elements on people's perceptions is provided by a range of pedestrian surveys. These studies apply one of two strategies, thus either relate test persons' walkability ratings to objectively measured environmental criteria, or ask pedestrians what aspects they rate as important for walking. Several concrete criteria could be identified, such as the presence of urban greenery (Lynch and Rivkin 1959; Borst et al. 2008; Kaufmann et al. 2010; Adkins et al. 2012), aesthetics (Lynch and Rivkin 1959; Brown et al. 2007; Agrawal et al. 2008; Samarasekara et al. 2011), traffic safety (Brown et al. 2007; Agrawal et al. 2008), here especially a physical separation from traffic (Lynch and Rivkin 1959; Clifton and Livi 2004; Kaufmann et al. 2010; Samarasekara et al. 2011; Adkins et al. 2012), appropriate street crossing facilities (Lynch and Rivkin 1959; Agrawal et al. 2008; Borst et al. 2008; Kaufmann et al. 2010) and curb cuts (Clifton and Livi 2004), low traffic speed and volume (Sanches and Ferreira 2007;

**Table 1** Environmental variables found significant for walking outcome

Variable	Review
Aesthetics	Sugiyama et al. (2012)
	McCormack and Shiell (2011)
	Saelens and Handy (2008)
	Badland and Schofield (2005)
	Owen et al. (2004)
	Handy (2004)
Security (crime)	Handy (2005)
	Saelens and Handy (2008)
	Sugiyama et al. (2012)
	Owen et al. (2004)
	McCormack and Shiell (2011)
	Saelens and Handy (2008)
Public parks	Handy (2004)
	Handy (2005)
Presence of sidewalks	Saelens and Handy (2008)
	Sugiyama et al. (2012)
	McCormack and Shiell (2011)
	Owen et al. (2004)
	van Holle et al. (2012)
	Saelens and Handy (2008)
	Handy (2004)
	Handy (2005)
Sidewalk condition/maintenance	Sallisetal. (2012)
	Sugiyama et al. (2012)
Traffic	McCormack and Shiell (2011)
	Sugiyama et al. (2012)
	Owen et al. (2004)
	van Holle et al. (2012)
	Handy (2005)
Street lighting	McCormack and Shiell (2011)
	Sallisetal. (2012)
Street furniture	McCormack and Shiell (2011)

Borst et al. 2008; Samarasekara et al. 2011), low levels of noise or air pollution (Sanches and Ferreira 2007), safety from crime (Brown and Szalay 2007; Clifton and Livi 2004; Agrawal et al. 2008), here especially appropriate lighting conditions (Clifton and Livi 2004; Sanches and Ferreira 2007; Kaufmann et al. 2010) and a lack of signs of decay such as vacant buildings or litter (Borst et al. 2008; Kaufmann et al. 2010), the existence of sidewalks (Clifton and Livi 2004; Kaufmann et al. 2010) with appropriate width (Lynch and Rivkin 1959; Samarasekara et al. 2011), appropriate surface structure (Lynch and Rivkin 1959; Sanches and Ferreira 2007; Agrawal et al. 2008) and a lack of obstructions (Sanches and Ferreira 2007; Samarasekara et al. 2011), low rates of slopes and stairs (Borst et al. 2008), as well as the presence of benches and other street furniture (Brown and Szalay 2007; Borst et al. 2008; Kaufmann et al. 2010).

Walkability, however, is an abstract quality which is not inherent in the environment, but rather incorporates the perspective of differing types of pedestrians (Jonietz and Timpf 2013). In fact, there is a growing awareness of this issue among planners and researchers (Buchmueller and Weidmann 2006). Basbas et al. (2010) and Vukmirovic (2010), for instance, provide very detailed lists of pedestrian abilities, which include aspects related to physical, physio-motor, sensorial and cognitive abilities, and affect various aspects of the walking process.

## 2.2 Computing Personalized Suitability Values

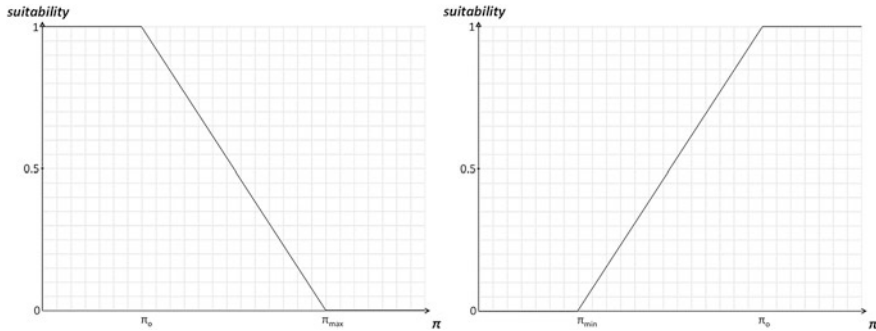
In prior work, we described how personalized spatial suitability values can be computed using a detailed user model (Jonietz et al. 2013; Jonietz and Timpf 2013). Thus, suitability is interpreted as a higher-order property  $suitability_{ij\alpha}$  of the system

$$W_{ij\alpha} = (\text{agent}_i, \text{environmental object}_j, \text{action}_\alpha)$$

In accordance with the psychological theory of affordances by Gibson (1979), and follow-up work by Warren (1995), it is proposed that the suitability depends on the correspondence of a certain capability  $cap_{ij\alpha}$  of the agent (e.g. leg length), and a disposition  $disp_{ij\alpha}$  of the environmental object (e.g. stair step height), which are relevant with regards to the action in question. According to Warren (1995), this correspondence can best be described as a ratio

$$\pi = \frac{disp_{ij\alpha}}{cap_{ij\alpha}}$$

As Fig. 1 shows, for each action, an optimal point  $\pi_O$  can be identified, which denotes the vector at which the performance of the action is possible with the least effort required by the agent. There is, however, also a critical threshold value which demarks the point  $\pi_{max}$  or  $\pi_{min}$ , above or below which the action can no longer be performed (Warren 1995). Accordingly, as seen in Fig. 1, the suitability reaches 1



**Fig. 1** Calculation of suitability from  $\pi$  (Jonietz 2016)

or 0 at these respective points, with a linear function allowing to identify values for each given  $\pi$  value between these two extremes.

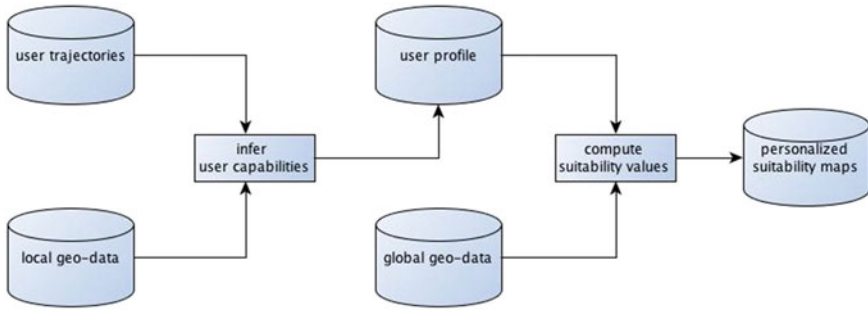
For more complex actions, which require the performance of several contributing sub-actions in order to be realized, we propose a three-level hierarchical model, thereby drawing from activity theory (Leontiev 1978; Kemke 2001):

- Pragmatic level: each action describes an intended goal
- Semantic level: each action describes a change in the state of the environment or the agent itself needed to reach the goal (i.e. a pragmatic action)
- Realization level: each action defines the agent's physical, motoric action which will lead to state changes (i.e. a semantic action)

Actions on all levels are mutually related in 1-1, 1-n or n-m-relationships. Accordingly, the  $suitability_{ija}$  for a pragmatic  $action_a$  can be computed by averaging the suitability values of all its contributing semantic actions, in the following denoted as  $sub-actions_{a'}$ , which, in turn, are averaged from their contributing realization actions, here  $sub-actions_{a''}$ . On the realization action level, the basic suitability values are computed depending on the relative closeness of the ratio  $\pi$  to the optimal point  $\pi_O$ , as described previously (Jonietz and Timpf 2013).

### 3 Method

Figure 2 shows the general framework of the proposed method. Each user's prior walking movement is tracked by recording x, y-coordinates at a predefined time interval, for instance every 5 s. The resulting points are then used in combination with detailed local geo-data of the respective area to derive user capabilities. In accordance with a hierarchical model of the action *WALKING*, these involve for instance the maximum slope the user was able to surmount, the minimum distance



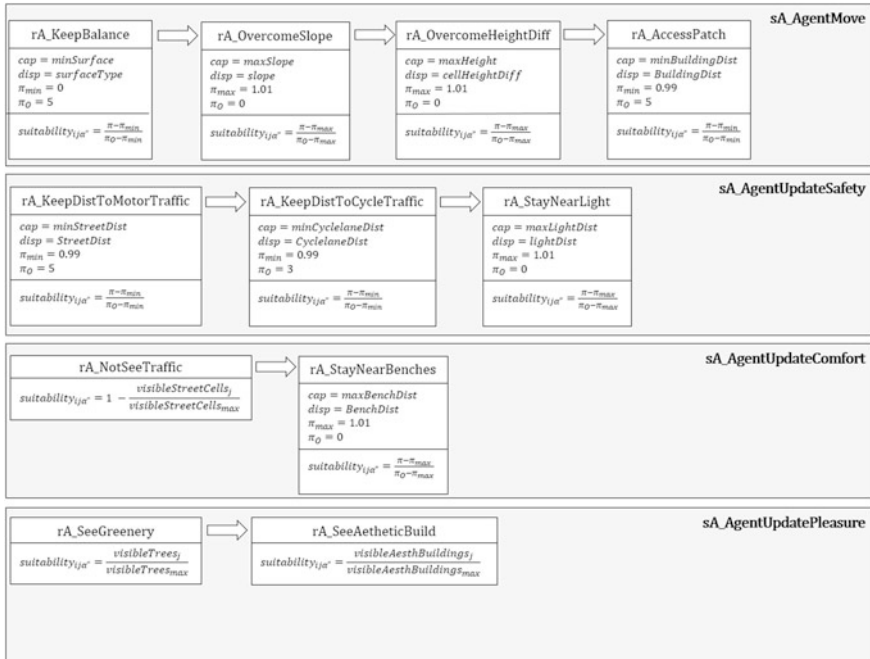
**Fig. 2** A framework to compute personalized walkability maps from movement trajectories

kept to the traffic lane, or kept visibility relations to urban greenery or aesthetical buildings. This information is stored as a user profile, and provides the basis for the calculation of personalized suitability maps. Each of these steps is further explained in the following.

### 3.1 A Subjective Model of Walkability

In accordance with our conceptual model (Jonietz et al. 2013; Jonietz and Timpf 2013), *WALKING* is understood as a pragmatic action, which can be further broken down into its contributing sub-actions. Following work by Alfonzo (2005), who describe the higher-level walking needs feasibility, accessibility, safety, comfort and pleasurability, and based on the empirical literature on walking, as reviewed in Sect. 2.1, the action *WALKING* is modeled as illustrated in Fig. 3. Thus, Alfonzo’s (2005) walking needs can be understood as agent states which in combination represent the goal state that the pragmatic  $action_{\alpha} PA\_WALK$  describes. Thus, apart from feasibility, which is omitted here since it does not refer to the built environment, the remaining needs are directly translated into semantic  $sub-actions_{\alpha}$ .

The need for accessibility describes the requirement of pedestrians to be able to reach destinations, both with regards to the overall trip distance as well as potential barriers for walking (Alfonzo 2005). In our action model, we translate this need to the semantic  $sub-action_{\alpha} SA\_AGENTMOVE$ . After the definition of a semantic action, its contributing realization  $sub-actions_{\alpha}$  need to be identified. In the case of  $SA\_AGENTMOVE$ , these include  $RA\_ACCESSPATCH$ , which denotes the agent’s ability to physically fit on the unobstructed sidewalk space,  $RA\_KEEPBALANCE$ , which represents the agent avoiding falls due to insufficient surface quality, and  $RA\_OVERCOME_SLOPE$  and  $RA\_OVERCOMEHEIGHTDIFFERENCE$ , which capture how movement can be hindered by slopes or height differences in the form of curb stones or stair steps, respectively.



**Fig. 3** Action model for  $pA\_WALK$

Apart from a change in location, however,  $pA\_WALK$  affects additional agent states, and accordingly, consists of other semantic  $sub\text{-actions}_\alpha$  as well, such as safety (Alfonzo 2005). With the semantic action  $sa\_AGENTUPDATE\text{SAFETY}$ , therefore, a pedestrian's mental state of feeling safe is described, which is realized by the  $sub\text{-actions}_\alpha$   $ra\_KEEPDISTANCETOMOTORTRAFFIC$  and  $ra\_KEEPDISTANCETOCYCLETRAFFIC$ , which refer to keeping safety buffer distances to the traffic and cycle lanes, and  $ra\_STAYNEARLIGHT$ , the need to remain within the area illuminated by street lights.

Furthermore, some measures of walking comfort and pleurability have been found significant for walking (Alfonzo 2005). Accordingly, the need for comfortable walking conditions translates to a semantic action  $sa\_AGENTUPDATE\text{COMFORT}$ , which is realized by  $ra\_NOTSEETRAFFIC$  and  $ra\_STAYNEARBENCHES$ . The first realization action describes the pedestrian trying to keep either at a certain distance from streets, or to have an object, such as a building, in between him- or herself and the road in order to be shielded from negative sensory inputs. The second sub-action focuses on the possibility to rest during a walking trip.

The aspect of pleurability is captured in  $sa\_AGENTUPDATE\text{PLEASURE}$ . In order to achieve this state, and evoke mental restoration, a pedestrian should maintain a visibility relation with aesthetically valuable environmental settings (Hidalgo et al. 2006; Lindal and Hartig 2013). Although the operationalization of aesthetics is certainly not a trivial task, studies have found the presence and visibility of urban

greenery as well as historically valuable buildings to be significant for the perception of aesthetical urban areas (Lynch and Rivkin 1959; Borst et al. 2008; Kaufmann et al. 2010; Adkins et al. 2012; Hidalgo et al. 2006). Therefore, in our model, *SA\_AGENTUPDATEPLEASURE* is realized by the realization actions *RA\_SEEGREENERY* and *RA\_SEEAESTHETICALBUILDING*.

After the identification of the contributing sub-actions, it is necessary to define relevant agent and environment-related properties which can then be related to each other to receive  $\pi$ -values (Jonietz and Timpf 2013). The list of environmental properties is based on the empirical literature on walkability, as described in Sect. 2.1, and involves the slope, the surface quality (classified into 5 types based on its evenness), the walking distance to the nearest bench, the Euclidean distances to the closest streetlight, the nearest building or other obstacle, the nearest street or cycle lane, the number of visible trees and aesthetical buildings, and the height difference of a location in contrast to the surrounding area (e.g. in case of stair steps or curb stones).

Regarding the definition of pedestrian-related capabilities, it is deliberately avoided to develop an overly detailed pedestrian model, separately incorporating, for instance, the range of factors listed by Vukmirovic (2010) or Basbas et al. (2010). In contrast, due to issues of feasibility and practicality, we choose to develop aggregated capabilities. Thus, to give an example, the ability to walk up an inclined surface depends on a bundle of pedestrian properties, such as strength, stamina and balance. If the influences of all factors were to be modeled separately, a detailed biometrical model would be needed to determine the exact dependency relationships between the particular capability and the environmental disposition. Further, the process of calculating suitability values would become overly complex, and require very detailed input data. Instead, the use of a less detailed, abstract property, such as the maximum slope which a pedestrian is able to overcome, aggregates the entirety of potentially influential pedestrian factors into one value. The full list of pedestrian capabilities can be seen in Fig. 3.

As a further step, the critical threshold values  $\pi_O$  and  $\pi_{min}$  or  $\pi_{max}$  must be determined and a suitability mapping function defined. In the case of  $\pi_{min}$ , its value is always set to 0.99, which in our model denotes the smallest unit below 1, and has the effect that with a  $\pi$  of 1, the suitability is still above 0. Accordingly, in the case of  $\pi_{max}$ , its value is set to 1.01. To return to the example of walking up an inclined surface, this means that if its slope equals the maximum slope which can be overcome by an agent, the suitability will not be 0, since the action is still possible, but with a low suitability value. Regarding the value chosen for  $\pi_O$ , in the case of  $\pi_{max}$ , it is set to 0, but the range of values is more diverse if there is a  $\pi_{min}$ . For this, values can be derived from the empirical literature, established guidelines or expert opinion. Thus, for instance, in the case of *RA\_ACCESSPATCH*, the optimal point  $\pi_O$  is set to 5, which means that if there is five times more space available than a pedestrian needs at a minimum, the suitability will reach its highest value 1.

Finally, a function needs to be defined which maps between  $\pi$ - and suitability values (Jonietz and Timpf 2013). Without going too much into detail here, we generally assume a linear relationship, and cut the allowed value ranges at

$0 \leq \text{suitability} \leq 1$ . The respective formulas are also listed in Fig. 3. A special case is posed by *RA\_NOTSEETRAFFIC*, *RA\_SEEGREENERY* and *RA\_SEEAESTHETICALBUILDING*. In contrast to the other realization actions, their related pedestrian characteristic is unlikely to be related to a capability, but rather to the specific preferences of a particular user. Therefore, we simply normalize the environmental attribute by division by the maximum value found in the data.

### 3.2 Inferring a User Profile from Trajectory Data

The action model described in the previous chapter provides the conceptual foundation for the identification of relevant user capabilities and environmental dispositions with regards to the high-level action *PA\_WALK*. The idea is to use actual user movements, derived via GPS-based tracking, to infer from priorly visited locations and their infrastructural characteristics the capabilities of a user. Thus, as a first step, a user is tracked in order to receive sequential x, y-coordinate pairs. Then, a detailed semantically rich environmental model is developed based on spatial data. Apart from environmental attributes such as slope or surface structure, more abstract information is also pre-calculated and embedded within the environmental model, including visibility and distance relations with streets, trees and aesthetically pleasing buildings.

The pre-processing of the spatial data for the environmental model requires several steps, which are conducted using ESRI's ArcGIS (ESRI 2012). First, it is necessary to determine the geometry of the walkable area. This involves several steps: First, all building footprints, as derived from Open Street Map<sup>3</sup> (OSM), are cut from a polygon of the whole study area. Then, smaller permanent obstacles such as lanterns or fixed trash cans need to be identified and deleted from the walkable polygon, which in our case required a field audit. Finally, the polygon is segmented based on the classification into sidewalk, street, and cycle lane, and locations with drastic height differences such as curb stones and stair steps are mapped and measured. In the next steps, the environmental characteristics of relevance for *WALKING* can be computed using ArcGIS geo-processing functionality. Thus, the slope is calculated based on a digital elevation model with 1 m cell size. Other indicators which are computed for each cell include the walking distance to the closest bench as obtained from OSM, as well as the Euclidean distances to the nearest unobstructed street light and the nearest obstacle, such as a building wall. In all of the above cases, the resulting values are stored in separate raster files with 40 cm cell size, the typical minimal footprint of a pedestrian (Chen et al. 2009). In addition, for each cell, the number of visible trees and historical buildings is computed and stored as raster files, as well as the distance to nearest unblocked street or cycle lane, hereby taking into account the positive shielding effect of buildings or other obstacles on perceived

---

<sup>3</sup><https://www.openstreetmap.org>.



traffic disturbances. Finally, the relative amount of visible road area for each point can be identified by counting the number of visible street cells (also 40 cm cell size) and normalizing the value to range from 0 to 1.

The environmental model being created, each user's movement trajectory is mapped and for each point, the respective attributes of the underlying raster cells are extracted and further analyzed. In particular, and in accordance with the model illustrated in Fig. 3, we extract the following movement characteristics: maximum distance kept to benches (*maxBenchDist*), minimum distance kept to buildings (*minBuildingDist*), maximum number of visible aesthetical buildings (*maxVisibleAesthBuildings*), maximum distance kept to street lights (*maxLightDist*), minimum distance kept to traffic (*minStreetDist*) and cycle lanes (*minCyclelaneDist*), maximum height difference of visited cell relative to its neighboring cells (*maxHeight*), maximum slope (*maxSlope*), the normalized maximum number of visible street cells (*maxVisibleStreet*), the minimum surface quality type (*minSurface*), and the maximum number of visible trees (*maxVisibleTrees*). These values give an account of the individual pedestrian's physical capabilities as well as preferences with regards to walking, and represent the user profile.

### 3.3 Computing Personalized Walkability Maps for Routing

In the next step, the user capabilities which were extracted from local geo-data, as described in the previous chapter, are used to compute personalized suitability maps. Thus, they provide the input for the agent-side of the action model of *PA\_WALK*, as illustrated in Fig. 3. The maximum slope found to have been overcome by the user based on the trajectory data, for instance, determines the capability value which will be set into a  $\pi$ -ratio with the slope of each cell of the study area. Therefore, the higher the capability of a user, the less sensitive the suitability calculation is with regards to high gradients. In the special case of *RA\_NOTSEETRAFFIC*, *RA\_SEEGREENERY* and *RA\_SEEAESTHETICALBUILDING*, which, according to our action model presented in Sect. 3.1, require no distinct user capability, our approach involves the use of the maximum values found when analyzing the user's movement trajectory to infer its perceived importance. Thus, for instance, we assume that the higher the maximum number of simultaneously visible trees, the higher the user's perceived importance of this particular sub-action *RA\_SEEGREENERY*. When evaluating the expected suitability of urban areas, therefore, a higher number of visible trees should be needed to receive higher suitability values in comparison to the calculation for a user with lower expectations.

In this manner, for each realization action involved in *PA\_WALK*, separate suitability values are computed as a raster file, which are then combined and averaged to receive an overall suitability map, which could in a next step be used for routing purposes, which, however, is omitted in this study.

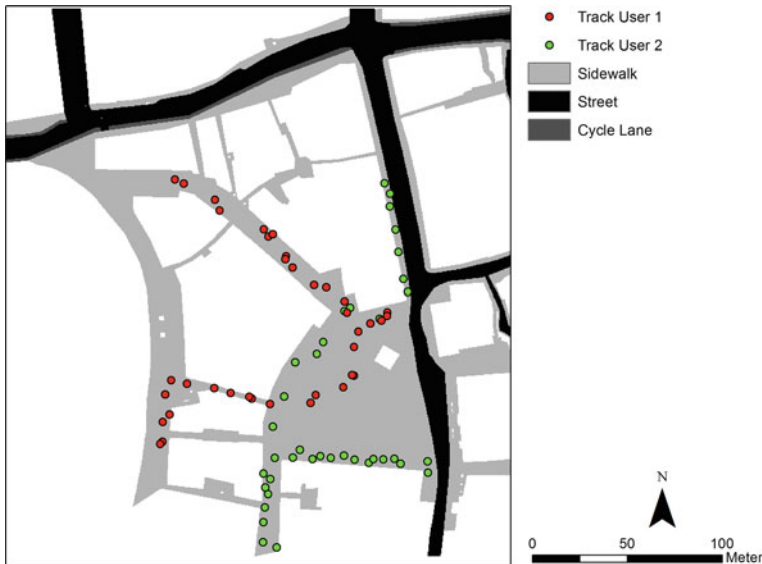
## 4 An Exemplary Application of the Method

In order to demonstrate the general functionality of our method, we applied it to the exemplary scenario of two hypothetical pedestrians moving in an inner city area. Figure 4 shows the extent of the study area. It is set in the city center of Augsburg, Germany, and encompasses part of the pedestrian precinct with the primary shopping area and the main city square in the south, as well as sidewalks along streets in the north and east. Also visible in Fig. 4 are exemplary tracks for the two hypothetical users, which, however, were in fact recorded by the authors of this paper using a GPS-enabled tracking device.

According to our framework, the first step involved the computation of separate raster files for the relevant environmental criteria, as described in Sect. 3.2. Then, for each of the resulting raster files, each point of the user trajectories was enriched with the values of the underlying raster cell. These numbers were used to compute the user profiles. Table 2 shows the resulting values for the two pedestrians.

As one can see, there are various differences, such as User 1 requiring more space for walking (*minBuildingDist*), staying further from street lights (*maxLightDist*) and streets in general (*minStreetDist*), being able to surmount higher slopes (*maxSlope*), or picking routes with a higher visibility of trees (*maxVisibleTrees*) in comparison to User 2.

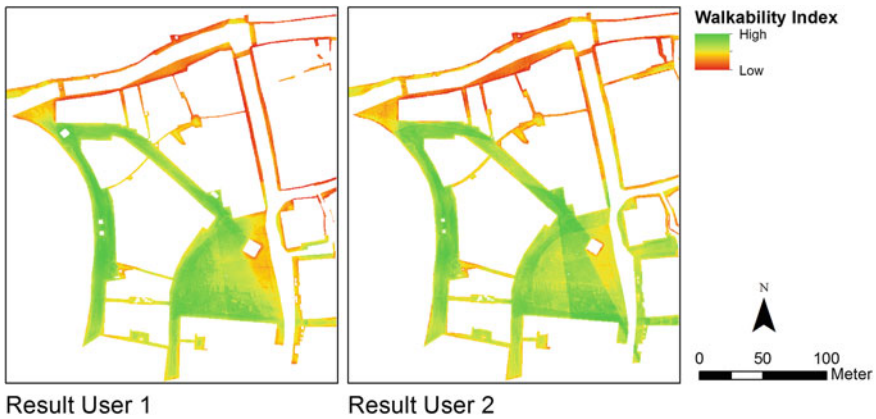
In the next step, these values were used as input for the formulas involved in computing suitability values for the whole study area. The resulting suitability maps mirror the differences found when analyzing the users' movement behavior, and can



**Fig. 4** Study area and hypothetical user trajectories

**Table 2** User profiles derived from trajectories

	User 1	User 2
maxBenchDist	81.2196	79.9245
minBuildingDist	0.8	0.4
maxVisibleAesthBuildings	7	8
maxLightDist	27.4446	17.8403
minCyclelaneDist	77.3068	71.7501
minStreetDist	12	0
maxHeight	0	0
maxSlope	11	8
maxVisibleStreet	0.1450	0.3836
minSurface	3	4
maxVisibleTrees	5	2



**Fig. 5** Resulting suitability maps for both users

be seen in Fig. 5. Thus, for instance, the fact that there is a lower surface quality in the center of the main city square in the south-east of the study area is clearly visible due to lower suitability values in the resulting map for User 2, who previously preferred higher-quality surface structures. User 1, in contrast, receives lower values in the direct proximity to streets, which is due to the generally higher buffer distance characterizing his or her prior movements. Also noticeable is the triangle shaped area of relatively higher suitability values at the main city square which can be seen in the map for User 2. This refers to an area with a high visibility of trees. Since, however, User 2 has priorly picked routes with less visible trees, it is assumed that, compared to User 1, he or she places lower importance to being able to see urban greenery. Therefore, in direct comparison, a lower number of visible trees is necessary to receive high suitability values.

## 5 Discussion and Conclusion

This study proposed to personalize walkability calculations by inferring the specific infrastructural needs of individual pedestrians from their pre-recorded movement trajectories. Based on a detailed model of the action *WALKING*, which was developed on the basis of a conceptual model of personalized suitability and a review of the empirical literature on walkability, user profiles were created and used to evaluate the expected personalized walkability of urban areas. An exemplary application demonstrated the general functionality of our approach by delivering plausible results.

Our work contributes to the topics of personalization and user profiling for LBS and other spatial recommender systems by providing a method for the personalized calculation of suitability values. So far, although suitability mapping is a standard application area of GIS, there has been no approach to incorporate the subjectivity and individuality of this process. Since our method delivers suitability values which are tailored to the individual user, it is of particular relevance for such systems. Further, our strategy of inferring pedestrian profiles from their movement trajectories and analyzing them with regards to walkability is novel and useful for pedestrian navigation applications. In addition, to our knowledge, this study represents the first approach to develop a general model of subjective walkability.

There are, however, still several shortcomings of our method. Thus, the requirements concerning the level of detail with regards to the spatial data are very high. To obtain such data, for instance related to the surface structure or the height of curb stones, it will in most cases be necessary to manually measure them at high financial costs, although micro-mapping is also conducted in OSM. Another crucial point is certainly the positional inaccuracy of the tracking mechanism. Due to the cell size of 40 cm, which, however, determines the values which are recorded and stored in the user profiles, an extremely high level of accuracy is needed for truly accurate results, which current positioning systems are only partly able to deliver. Also, a very dense temporal resolution of the recorded points is necessary to avoid missing critical walking situations, such as the pedestrian moving up a curb stone. The additional use of pedestrian dead reckoning (PDR) techniques might be of use for addressing this particular issue (e.g. Randell et al. 2003). Further, our system has currently no learning capabilities in the sense of updating an existing user profile with new data. Finally, although being based on the empirical literature, the model of *WALKING* is certainly drastically simplified, for instance omitting the process of road crossing, other aspects which are not directly attributed to the built environment such as the presence of other pedestrians and crowding, or the direction of movement for visibility-related aspects as well as slope.

For future work, we plan to test the method with actual users, and implement a learning and updating as well as a routing functionality.

## References

- Adkins A, Dill J, Luhr G, Neal M (2012) Unpacking walkability: testing the influence of urban design features on perceptions of walking environment attractiveness. *J Urban Des* 17 (4):499–510
- Agrawal AW, Schlossberg M, Irvin K (2008) How far, by which route and why? A spatial analysis of pedestrian preference. *J Urban Des* 13(1):81–98
- Alfonzo MA (2005) To walk or not to walk? The hierarchy of walking needs. *Environ Behav* 37 (6):808–836
- Badland H, Schofield G (2005) Transport, urban design, and physical activity: an evidence-based update. *Transp Res Part D* 10:177–196
- Basbas S, Konstantinidou C, Ribas DJM (2010) Factors and mechanisms which determine the outcome of strategic decisions with regard to walking. In: COST 358 Pedestrian Quality Needs Final Report, Part B, pp 127–156
- Borst HC, Miedema HM, de Vries SI, Graham JM, van Dongen JE (2008) Relationships between street characteristics and perceived attractiveness for walking reported by elderly people. *J Environ Psychol* 28(4):353–361
- Brown BB, Szalay C (2007) Walkable enroute perceptions and physical features converging evidence for en route walking experiences. *Environ Behav* 39:34–61
- Buchmueller S, Weidmann U (2006) Parameters of pedestrians, pedestrian traffic and walking facilities. IVT-Report Nr. 132, Institute for Transport Planning and Systems, ETH Zurich
- Cervero R, Kockelmann K (1997) Travel demand and the 3Ds: density, diversity, and design. *Transp Res D* 2(3):199–219
- Chen M, Bärwolff G, Schwandt H (2009) Automation model with variable cell size for the simulation of pedestrian flow. In: Proceedings of the 7th international conference on information and management sciences 2008, Urumtschi, pp 727–736
- Clifton KJ, Livi AD (2004) Gender differences in walking behavior, attitudes about walking, and perceptions of the environment in three Maryland communities. In: Transportation research board conference proceedings 35, conference on research on women's issues in transportation, Chicago, pp 79–88
- ESRI (2012) ArcGIS Desktop: Release 10.2. Environmental Systems Research Institute, Redlands, CA
- Gartner G, Huang H, Millonig A, Schmidt M, Ortig F (2011) Human-centered mobile pedestrian navigation systems. *Mitteilungen der Österreichischen Geographischen Gesellschaft* 153:237–250
- Gibson J (1979) The ecological approach to visual perception. Houghton Mifflin Company, Boston
- Handy SL (2004) Critical assessment of the literature on the relationship among transportation, land use, and physical activity. Report, Transportation Research Board, Washington, DC
- Handy S (2005) Does the built environment influence physical activity? TRB special report 282. Transportation Research Board, Washington DC
- Hidalgo CM, Berto R, Galindo MP, Getrevi A (2006) Identifying attractive and unattractive urban places: categories, restorativeness and aesthetic attributes. *Medio Ambiente y Comportamiento Humano* 7(2):115–133
- Holone H, Misund SE, Ramachandran B (2007) Users are doing it for themselves: pedestrian navigation with user generated content. In: NGMAST 2007, New York, USA
- Huang H, Klettner S, Schmidt M, Gartner G, Leitinger S, Wagner A, Steinmann R (2014) AffectRoute—considering people's affective responses to environments for enhancing route-planning services. *IJGIS* 28(12):2456–2473
- Jonietz D, Timpf S (2013) An affordance-based simulation framework for assessing spatial suitability. In: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D,

- Vardi MY, Weikum G, Tenbrink T, Stell J, Galton A, Wood Z (eds) *Lecture notes in computer science*. Springer International Publishing, Cham, pp 169–184
- Jonietz D, Schuster W, Timpf S (2013) Modelling the suitability of urban networks for pedestrians: an affordance-based framework. In: Vandembroucke D, Bucher B, Crompvoets J (eds) *Geographic information science at the heart of Europe*. Lecture notes in geoinformation and cartography. Springer, Heidelberg, pp 369–382
- Jonietz D (2016) *From space to place—A computational model of functional place*. Doctoral thesis, University of Augsburg, Germany
- Kaufmann C, Papaioannou P, Blaszczyk M, Marques Almeida D de (2010) Preconditions and how they are perceived. In: COST 358 *Pedestrian quality needs final report, Part B*, pp 15–48
- Kemke C (2001) *About the ontology of actions*. Technical Report MCCS-01-328, Computing Research Laboratory, New Mexico State University
- Krisp J, Keler A (2015) Car navigation: computing rourest hat avoid complicated crossings. *IJGIS* 29(11):1988–2000
- Leontiev AN (1978) *Activity, consciousness, and personality*. Prentice-Hall, New Jersey
- Lindal P, Hartig T (2013) Architectural variation, building height, and the restorative quality of urban residential streetscapes. *J Environ Psychol* 33:26–36
- Lynch K, Rivkin M (1959) A walk around the block. *Landscape* 8:24–34
- Mautz R (2009) Overview of current indoor positioning systems. *Geodezija ir Kartografija* 35 (1):18–22
- McCormack GR, Shiell A (2011) In search of causality: a systematic review of the relationship between the built environment and physical activity among adults. *Int J Behav Nutr Phys Act* 8 (125):1–11
- Millonig A (2006) Routennetze für mobile Fußgänger-Navigationsanwendungen: ein neuer Ansatz für die Optimierung auf Basis von quantitativen Bewegungsdaten. In: CORP 2006, Vienna, Austria
- Owen N, Humpel N, Leslie E, Baumann A, Sallis JF (2004) Understanding environmental influences on walking. review and research agenda. *Am J Prev Med* 27(1):67–76
- Park S (2008) *Defining, measuring, and evaluating path walkability, and testing its impacts on transit users' mode choice and walking distance to the station*. Dissertation, University of California, Berkeley
- Randell C, Djallil C, Muller H (2003) Personal position measurement using dead reckoning. In: *Proceedings of the seventh international symposium on wearable computers*. IEEE Computer Society, pp 166–173
- Saelens BE, Handy S (2008) Built environment correlates of walking: a review. *Med Sci Sports Exerc* 40(7 suppl):550–566
- Saelens BE, Sallis JF, Frank LD (2003) Environmental correlates of walking and cycling: findings from the transportation, urban design, and planning literatures. *Ann Behav Med* 25(2):80–91
- Sallis JF, Floyd MF, Rodriguez DA, Saelens BE (2012) Role of built environments in physical activity, obesity, and cardiovascular disease. *Circulation* 125(5):729–737
- Samarasekera GN, Fukahori K, Kubota Y (2011) Environmental correlates that provide walkability cues for tourists: an analysis based on walking decision narrations. *Environ Behav* 43(4):501–524
- Sanches SP, Ferreira MAG (2007) How the elderly perceive the quality of sidewalks. In: *ITE 2000 annual meeting and exhibit*, Nashville
- Sugiyama T, Neuhaus M, Cole R, Giles-Corti B, Owen N (2012) Destination and route attributes associated with adults' walking: a review. *Med Sci Sports Exerc* 44(7):1275–1286
- van Holle V, Deforche B, van Cauwenberg J, Goubert L, Maes L, van de Weghe N, Bourdeaudhuij I de (2012) Relationship between the physical environment and different domains of physical activity in European adults: a systematic review. *BMC Public Health* 12
- Völkel T, Weber G (2008) *RouteCheckr: personalized multicriteria routing for mobility-impaired pedestrians*. In: *ASSETS'08*, Halifax, Canada

- Vukmirovic M (2010) Functional abilities of humans and identification of specific groups of pedestrians. In: Walk 21 XI international conference on walking and liveable communities, The Hague, Netherlands, 16–19 November 2010
- Walter V, Kada M, Chen H (2006) Shortest path analyses in raster maps for pedestrian navigation in location based systems. In: Commission IV, WG IV/6
- Warren WH (1995) Constructing an econiche. In: Flach J, Hancock P, Caird J, Vicente K (eds) The ecology of human-machine systems. Erlbaum, Hillsdale, pp 210–237

# Learning On-Street Parking Maps from Position Information of Parked Vehicles

Fabian Bock, Jiaqi Liu and Monika Sester

**Abstract** Many drives in crowded cities end with a challenging parking search, and visitors often do not know which streets allow on-street parking. Therefore, we present a learning-based approach to automatically generate on-street parking maps from parked vehicle positions detected by sensing vehicles. Multiple sets of features are proposed to describe the occupancy of every small road segment and its surroundings at different time instances. The usage of k-means algorithm as unsupervised learning and random forests as supervised learning are compared by applying these feature sets. The proposed approach is evaluated with repeated LiDAR measurements on more than five kilometers of potential parking space length. Our approaches, while keeping the model more generic, reveal slightly better results than an approach from literature. In particular, the unsupervised approach does not need a training data set and is free of any area specific parameter choice.

**Keywords** Map generation · Parking management · Crowd-sensing · Machine learning

## 1 Motivation

The search for a parking space is often very time-consuming and costly for drivers in crowded cities. Van Ommeren et al. (2012) estimate based on a nation-wide survey that 30 % of trips in the Netherlands (excluding residential and employer-provided

---

F. Bock (✉) · M. Sester  
Institute of Cartography and Geoinformatics, Leibniz University, Hannover, Germany  
e-mail: fabian.bock@ikg.uni-hannover.de

M. Sester  
e-mail: monika.sester@ikg.uni-hannover.de

J. Liu  
Electrical and Computer Engineering Department, Rice University,  
Houston, Texas, USA  
e-mail: jiaqi.liu@rice.edu



parking) end with parking search. Considered from a traffic management view, the amount of traffic due to parking search is measured by several studies to be between 8 and 74 % of the total traffic in congested areas (Shoup 2006). This traffic also leads to a huge amount of CO<sub>2</sub> pollution. Shoup (2007) estimates that a small district in Los Angeles with less than 500 on-street parking spaces causes more than 700 tons of CO<sub>2</sub> per year by 950,000 miles of parking search.

Parking search is not only fostered by missing information about parking availability. Visitors also need information about the locations of parking facilities. While traffic signs often exist that guide to large off-street parking facilities, only a few cities provide central information about on-street parking opportunities. Even if maps of on-street parking spaces exist, they need to be updated continuously as parking regulations change from day to day. Such parking maps could be used inside the car in navigation systems that visualize parking opportunities close to the destination. Also looking further into the future, automated valet parking (e.g. Furgale et al. 2013) needs the latest knowledge about parking space locations.

To provide up-to-date information about on-street parking spaces for a large number of cities, automated methods are very beneficial. Once set up, they come with lower costs and effort compared to manual recording. Also, they can easily be extended to additional areas. Modern vehicles, equipped with sensors that detect parked vehicles or empty parking spots (e.g. Park et al. 2008), can be used as data sources. Aggregation of this information can be used to estimate the parking regulations automatically. However, this task is not trivial due to challenges arising from human behaviors. While some legal parking spaces might be used only infrequently, vehicles are parked more often in other spots where parking is not allowed. As such, it is necessary to learn typical characteristics of valid parking spaces.

To the best of our knowledge, there is very few literature about the automated generation of on-street parking maps. Both Ge et al. (2013) and Coric and Gruteser (2013) use parking information from the ParkNet project (Mathur et al. 2010) where ultrasonic sensors are mounted on passengers' side of the vehicle to record data on specified tracks. Ge et al. (2013) count the vehicles parked at a certain position in different time windows and use an absolute threshold to decide on the position's legality. Coric and Gruteser (2013) developed an algorithm that decides the legality based on a weighted occupancy average. This value is compared to a threshold followed by a post-processing to smooth the results. Both approaches have the disadvantage that they depend on manually defined parameters and thresholds which may vary a lot for different investigation areas. Furthermore, weak assumptions are presumed for the length of parking prohibitions and for data importance at different occupancy levels.

Our approach turns the identification of legal parking areas into a machine learning problem. This way, we reduce or even avoid the need for manual parameter optimization. Furthermore, we use LiDAR sensors instead of ultrasonic sensors which allow the explicit identification of vehicles. In addition to the occupancy measure used in Coric and Gruteser (2013), we propose several additional features. These features include more details of the sensed parking information in space and time. For example, the average occupancy of a spot is compared to the average

occupancies in its vicinity. With the calculation of these features for different length scales, we avoid further assumptions on the parking characteristics.

The features are used to compare a supervised (classification) and an unsupervised (clustering) approach. The random forest algorithm (Breiman 2001) is chosen for classification and the k-means algorithm for clustering as both are established and robust machine learning methods. For the clustering approach, an additional step is necessary, since clustering only assigns data points to groups (clusters). It does not provide information on which cluster corresponds to legal parking spaces. However, a simple decision is possible if the number of clusters is set to two (legal/illegal): the cluster with the higher average occupancy rate of parked vehicles corresponds to the legal parking spaces. This unsupervised approach has the strong benefit that there is no need for a training step with elaborate training data.

Summarized, the contributions of this paper are

- (1) the description and evaluation of multiple features for distinguishing between legal parking spaces and no-parking zones,
- (2) the comparison of a method from the literature with the classification method which uses the new features, and
- (3) the proposal and evaluation of an unsupervised approach that shows a similar performance to the supervised approaches.

The rest of the paper is structured as follows: in Sect. 2, we place our work in relation to further literature. Section 3 contains the description of the methodology including the data preprocessing, the feature definition, and the learning part. In Sect. 4, we present the evaluation of the results of proposed clustering and classification methods and analyze the relevance of the different features. Finally, a conclusion is given in Sect. 5 as well as an outlook on future work.

## 2 Related Work

For the observation of parking spaces, there are several approaches for both static and mobile sensors. While static sensors allow for a continuous observation of specific parking spots with high accuracy, they usually come with high costs and little flexibility (used e.g. in SFMTA 2014). Therefore, mobile sensors are favored in many situations. Modern vehicles are often equipped with cameras and ultrasonic sensors that can be used for the detection of parked vehicles or gaps in parking lanes. Ultrasonic sensors are already widely spread for automatic parking assistance systems in series vehicles (Bengler et al. 2014). They also provide information about parking gaps during driving (Mathur et al. 2010; Park et al. 2008). An approach for sensor fusion of ultrasonic sensors and cameras is proposed by Choi et al. (2014). For more precise measurements, laser scanning systems can be used. Combined with a high-precision global navigation satellite system (GNSS) unit, they are able to provide precise positions of parked vehicles with a high detection quality (Thornton et al. 2014; Bock et al. 2015). Furthermore, smart phones can also serve as a sensor when

they are carried with the driver. Stenneth et al. (2012) detect the moving patterns of different travel modes of the smart phone owner and concludes parking events for travel mode transitions from driving to walking.

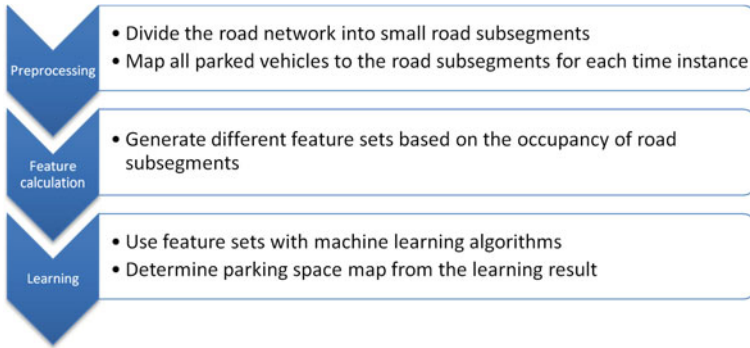
The generation of maps is present in many domains. While maps were created manually in recent times, more and more maps are generated automatically nowadays. In robotics, a main task is the perception of the environment and the generation of obstacle maps both for localization and collision avoidance (e.g. Thrun 2002). In traffic context, a main focus lies on the automatic generation of road networks from GPS trajectories and from further in-vehicle sensors to keep navigation maps up-to-date (e.g. Davies et al. 2006). Compared to that research field, research about parking map generation is very rare. Ge et al. (2013) and Coric and Gruteser (2013) generate on-street parking maps using data from ultrasonic sensors of the ParkNet project (Mathur et al. 2010). Ge et al. (2013) simply count the number of vehicles parked at each position. If the number reaches a certain threshold, this position is assumed to allow parking.

A more sophisticated solution is proposed by Coric and Gruteser (2013). They first divide the road into segments of one meter. Then they calculate a weighted average of occupancy for these road segments. The weights depend on the general occupancy level of the road. They assume that more occupied roads provide more information about parking space legality. While this assumption might be true in many situations, this does not hold for situations with high parking demand. Then, illegal parking grows considerably with increasing occupancy level (White 2007) which disturbs the algorithm. The weighted average is compared to a fixed threshold. In the post-processing, they smooth the result to get rid of small areas of legal or illegal parking up to a fixed length threshold. While it is reasonable that a parking spot shorter than a vehicle length is implausible, parking prohibition also exists for shorter distances in our investigation area. Both thresholds are manually defined and no calibration method is proposed. This weighted occupancy rate thresholding approach (called WORT in the following) is applied to our data set (Sect. 4) to compare the results with our proposed methods.

## 3 Methodology

### 3.1 Overview

The generation steps for on-street parking maps consist of the assignment of parked vehicle positions to small road segments, the aggregation of information from multiple time instances to features, and the classification based on several features (see Fig. 1). The positions of parked vehicles at different times (described in Sect. 3.2) and a road network from OpenStreetMap are used as inputs. The pre-processing step (Sect. 3.3) contains the separation of the road network into small road segments (called road subsegments in the following) and the assignment of the parked vehicle



**Fig. 1** Overview of the steps for the generation of the on-street parking map from parked vehicle positions

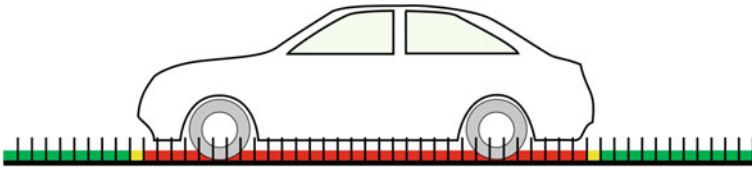
information to them. This occupancy information of all road subsegments is used to calculate various features based on the occupancy of the road subsegments themselves and the occupancy in the neighborhood. Finally, the decision whether parking is legal is made in the learning step (Sect. 3.4). Here, both supervised and unsupervised approaches are described. This procedure results in a set of road subsegments with parking legality information which can be fused to a complete parking map.

### 3.2 Description of Required Data

For the generation of on-street parking maps, the main input is the position of parked vehicles at different time instances (e.g. at different times of the day or various days). This information can be generated with several sensors as described in Sect. 2. Our approach is assumed to work properly for all of these sensor types. While laser scanners, cameras, and ultrasonic sensors provide information about the position and the extent of the vehicle directly, GPS trajectories from smart phones or in-vehicle recordings only provide a single position. In the latter case, a typical length for a vehicle can be assumed. Furthermore, each vehicle detection needs to be annotated with a timestamp.

### 3.3 Data Preprocessing

The road network for the region of interest is obtained from OpenStreetMap. It is processed to obtain road segments with nodes at every intersection. Each road segment is then split into small subsegments to learn the parking legality individually for each of them. We chose a length of 10 cm, but a coarser partition should also



**Fig. 2** Projection of a detected vehicle on the road subsegments (small strokes). The color represents the occupancy: the occupancy value is 1 for completely occupied road subsegments (*red*), between 0 and 1 for partially occupied road subsegments (*yellow*), and 0 for not occupied road subsegments (*green*)

be feasible as long as small parking prohibition areas can be represented. The road subsegments are connected to a graph to determine neighborhood relations. Then, the extent of the detected vehicles is projected on the road subsegments (see Fig. 2). In addition, a flag is set whether the vehicle is parked on the left or right side of the road in digitalization direction.

### 3.4 Feature Set Definitions

We use eight feature sets for learning. They contain both raw and aggregated data based on the road subsegment itself and its neighboring road subsegments. An overview is given in Table 1. Many of the feature sets contain a distance parameter for the relevant neighborhood. As we want to keep the model generic and avoid parameter optimization, we extend the feature sets for multiple generic values. The distance parameter of the features is set to 0.5 m, 1 m, 3 m, 5 m, 10 m, 20 m, and 40 m to cover the effects in both the short and distant neighborhood.

**Table 1** Overview of all feature sets used in our evaluations

Feature set number	Name	Size of feature vector
1	Raw occupancy	#(measurement drives)
2	Occupancy rate	1
3	Weighted occupancy rate	1
4	Raw neighbor occupancy	#(measurement drives) * #(distance values)
5	Average neighbor occupancy	#(measurement drives) * #(distance values)
6	Gaussian average neighbor occupancy	#(measurement drives) * #(distance values)
7	Segment saturation	#(measurement drives)
8	Road subsegment attractiveness	#(distance values)

- FS1: Raw occupancy** This is basically the input data by itself. At each road subsegment, we take the occupancies for each time instance and for each road side as features. Therefore, this feature set contains a feature column for each time instance.
- FS2: Occupancy rate** This feature set, by its name, is the average occupancy of the road subsegment over all time instances.
- FS3: Weighted occupancy rate** Weighted occupancy rate is the concept suggested in Coric and Gruteser (2013). It contains the average occupancy over all time instances for each road subsegment weighted by the average occupancy of the full road segment at each time instance.
- FS4: Raw neighbor occupancy** After having all of the low and high level information about the road subsegment, we include information about its neighbors. This is done by traversing the road subsegment graph, and calculating the average occupancy of all the neighbors at a certain distance for each time instance. Note that there are usually two neighbors of the same distance because there are two directions (to the left and right) where the neighbors can be. In some special cases like at intersections, the number of neighbors, that are at the same distance to the current subsegment, can be even larger than two. To cope with these situations where the number of neighbors can vary, we define that, for each distance, we only have one feature for each time instance which is the average occupancy of all the possible neighbors at that distance for that particular time instance.
- FS5: Average neighbor occupancy** For each time instance, we take the average occupancy of all the neighbors within a predefined range. The identification of neighbors is calculated by traversing the neighborhood graph like in the previous feature set.
- FS6: Gaussian average neighbor occupancy** The weighted average occupancy of neighbors is similar to the ordinary average occupancy of neighbors. The only difference is that when calculating the average, we apply a Gaussian function over the neighbors' occupancy rates based on their distance to obtain the weighted average occupancy of neighbors. The width of the Gaussian function is chosen such that its value is 10 % of the maximal value at the distance limit. This weighting is applied assuming that closer road subsegments give a stronger hint on the parking legality, but the occupancy of distant road subsegments still provide some valuable information.
- FS7: Segment saturation** The segment saturation describes the occupancy level of a complete road segment. The number of occupied road subsegments of a road segment at one time instance is divided by the maximal number of occupied road subsegments of all time instances on the specific road segment. It is assumed that the maximum value represents the fully occupied road segment. If the value of this feature is low, parking demand is low at this time instance while the parking demand is high if this value is high.
- FS8: Road subsegment attractiveness** The road subsegment attractiveness represents the occupancy rate of the specific road subsegment compared to the road subsegments in the neighborhood. The occupancy rate is divided by the maximum occupancy rate of the neighbors within a certain range. A low value means

that this road subsegment is less attractive than others in the neighborhood. This is a clue that parking might be not allowed there. If this value is high, this road subsegment is comparably attractive as the most occupied subsegment. Therefore, it is more likely that parking is allowed there.

### ***3.5 Learning the Parking Legality of Road Subsegments***

The decision whether parking at a given road subsegment is legal is a typical classification problem. Each road subsegment belongs to the classes “legal” or “illegal”. Therefore we investigate the use of a classification algorithm. However, classification belongs to the group of supervised learning algorithms. This means a training data set with labeled data is always needed before application to an unknown area. Clustering methods as a subset of unsupervised learning algorithms do not have this requirement. They group objects to clusters based on their similarity. Some of the clustering algorithms provide the possibility to define the number of clusters (two in our case). For the assignment to the correct class, however, a manual or automatic post-processing step is necessary. Nevertheless, it has the strong advantage that it can be applied to different areas without the need to generate representative training data. In the following, both approaches are described.

#### **3.5.1 Unsupervised Learning: K-Means**

We used the k-means algorithm (implementation from MATLAB) for unsupervised learning as a basic and established clustering algorithm. This algorithm iteratively improves the assignment of the observations to the clusters based on their distance to the cluster centers. The decision boundaries of the clusters are hyperplanes in the middle between the cluster centers. The k-means algorithm has the advantages of being fast and allowing users to define the number of clusters. The latter is important in our problem setting since we have the two classes “legal” and “illegal”, but we do not know which cluster represents the legal parking spaces. The idea in this paper is the assumption that legal parking spaces have a higher average occupancy rate than parking prohibitions. That means our algorithm assigns the cluster with higher average occupancy rate to the legal parking spaces.

#### **3.5.2 Supervised Learning: Random Forests**

For supervised learning, we used a random forest classifier (Breiman 2001, implementation from MATLAB). Since the training data is assigned to the two classes “legal” or “illegal”, the classifier directly estimates the classes for the test data and we do not need to guess the class assignment. The random forest algorithm is based on the generation of a large set of different decision trees. The diversity of these trees

results from the random choice of a subset of training data and a random choice of features for each decision step. The benefits of random forests are that they are fast, quite intuitive to interpret, and robust against overfitting.

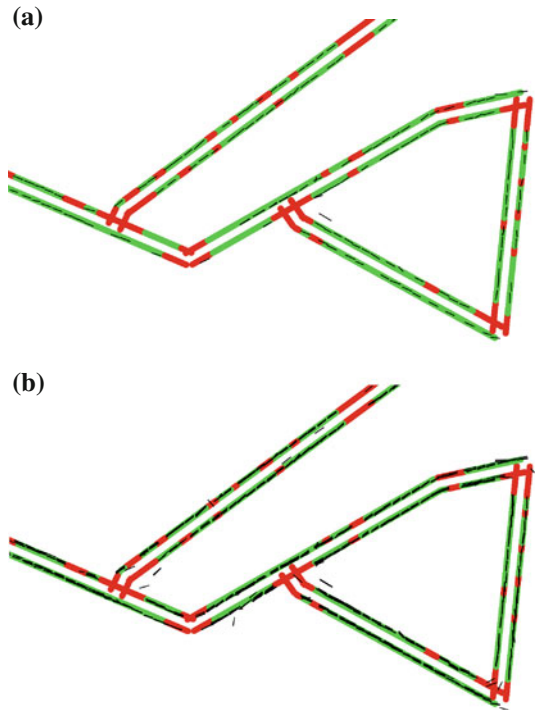
## 4 Evaluation

### 4.1 Evaluation Approach

#### 4.1.1 Test Scenario

A mobile mapping system equipped with a light detection and ranging (LiDAR) sensor is used to record the streets of a test track nine times during the course of a day. In an offline procedure, the positions of parked vehicles are extracted from sensor data. Two examples of the extracted data are shown in Fig. 3. Precision and recall of the detection were both higher than 95 %. The test track has an effective length of more than 2.5 km in a large city. This means a length of more than 5 km of potential parking space is evaluated, since both sides of the road are observed. As

**Fig. 3** Example for input data of **a** one measurement drive and **b** all nine measurement drives. The *black lines* represent the extent of the parked vehicles, *red* (illegal) and *green* (legal) are the ground truth classes of the road subsegments





the detection procedure cannot distinguish between parking and stopping vehicles, the test track is chosen to cover single lane roads to reduce the impact of stopping vehicles at intersections. It possesses parking spaces on a total length of 3.1 km for about 500 vehicles parallel to the road. Road subsegments for areas with parking perpendicular to the road as well as parking areas for special groups like taxi parking spaces are excluded from the evaluation.

#### 4.1.2 Ground Truth Recording

The ground truth of the parking space map is obtained by a combination of approaches. For most of the roads, we used a handheld differential GPS device to record the starting and ending positions of the legal parking spaces. The standard deviation of the GPS device measurements ranges from a few centimeters to multiple meters in our measurement. For the streets with low GPS accuracy due to limited sky view, we used Google satellite images for a first estimation, as well as 3D point clouds from our laser scan data for a precise measurement of the ground truth. Measurement accuracy and precision of the laser scanning itself is 10 and 5 mm, respectively. The positioning unit of the mobile mapping system has an accuracy of 20 cm in horizontal directions for urban scenarios. Boundaries of the parking area like curb stones or traffic signs can be clearly identified from the laser scan point clouds.

#### 4.1.3 Uncertainty in Ground Truth

In addition to the measurement inaccuracy of the equipment, the start and end of a parking space often cannot be identified precisely. For example, the curb at the end of a parking lane is often not perpendicular to the road. In order to account for both uncertainties, we do not evaluate the borders of the legal and illegal parking spaces in the output. More precisely, 0.5 m to each side of the borders between the legal and illegal parking spaces is not evaluated and not counted in the quality measures.

#### 4.1.4 Quality Measures

There are four basic types of quality measures, namely the count of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Each road subsegment is assigned to one of these counts. Based on these counts, we calculate the false positive rate  $FPR = \frac{FP}{TN+FP}$  and the true positive rate  $TPR = \frac{TP}{TP+FN}$ . Also, we can calculate the overall accuracy by the formula:

$$acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

### 4.1.5 Cross Validation

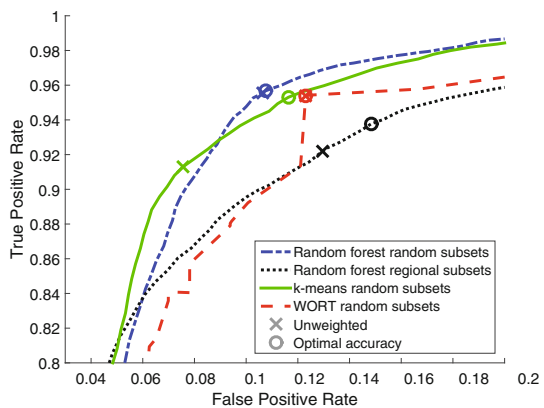
For better evaluation of the limited evaluation data set, a three-fold cross validation is applied to the supervised learning and the weighted occupancy rate thresholding (WORT) with smoothing (Coric and Gruteser 2013). The k-means approach is also evaluated for different subsets of the data set to investigate the impact of the buffer size. To this end, each road segment is assigned to one of three similar sized sets. To investigate the generalization of the models, we generated two kinds of cross validation sets. The first cross validation sets are based on the geographical location of the road segments (called regional road subsets in the following). For the second, the road segments are assigned randomly in such a way that the length of all three sets is about the same. Note that cross validation with a random split of subsegments would lead to an improper evaluation since adjacent subsegments have very similar neighborhood features (features 4-6).

## 4.2 Results

### 4.2.1 Overview of Results

A comparison of results for our supervised and unsupervised learning approaches as well as for the weighted occupancy rate thresholding (WORT) approach (Coric and Gruteser 2013) is shown in Fig. 4 and Table 2. The table contains both the accuracy without variation of cost weights and optimal accuracy values of each approach for different choices of the cross validation sets. Since WORT does not suggest a method to choose proper threshold values, we applied a brute-force search for the best parameters with cross validation and cost function  $c = FP + \alpha \cdot FN$  ( $\alpha$  is a weighting parameter). The quality measures in Table 2 reveal the best results for random forest with random road subsets, but also similar results for nearly all other approaches

**Fig. 4** Receiver Operating Characteristic (ROC) curve for comparison of random forest, k-means, and WORT with smoothing



**Table 2** Results for all methods with different choice of data subsets, parameters, and feature sets

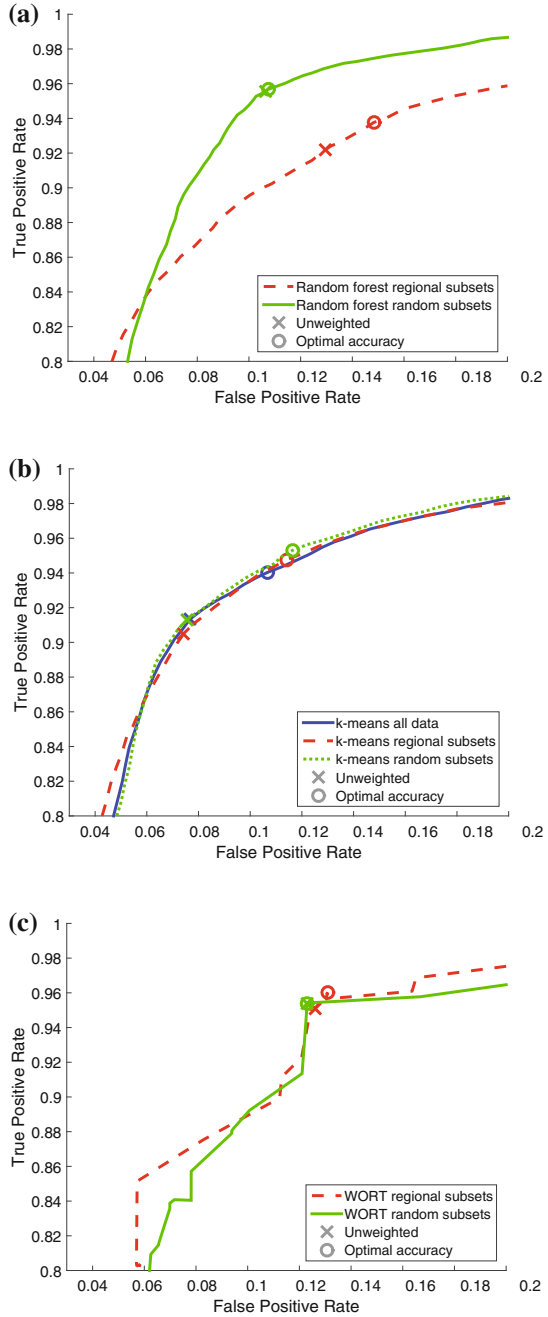
Method	Choice of data subsets, parameters, and feature sets	Accuracy (unweighted) (%)	Optimal accuracy (%)
Random forest	Regional road subsets	90.1	90.3
	Random road subsets	93.0	93.1
k-means	Regional road subsets	91.3	92.2
	Random road subsets	91.8	92.5
	Full data set clustered	91.8	92.1
Weighted occupancy rate thresholding (WORT) with smoothing (Coric and Gruteser 2013)	Suggested parameters	85.5	–
	Optimized (regional road subsets)	92.0	92.3
	Optimized (random road subsets)	92.3	92.3

with values larger than 90 %. Only if the suggested parameters of Coric and Gruteser (2013) are used with their approach, the result is significantly worse. We assume that their observation area has very different parking characteristics than ours. Figure 4 compares the Receiver Operating Characteristic (ROC) curves. For random forests, this curve is generated changing the threshold for the estimated class probabilities. Shifting the separation plane between the cluster centers is used for k-means. For WORT, the relative weight of FP and FN is varied with parameter  $\alpha$  in the brute-force search. The plot in Fig. 4 shows that k-means and random forest with random subsets have the best results for nearly the complete curve. Clearly worse are the curves for WORT and random forest with regional subsets. The clear difference between the two random forest results is discussed in Sect. 4.2.2. A qualitative comparison is presented in Sect. 4.2.3. Finally, we discuss details about the feature importance in Sect. 4.2.4 and the necessary number of measurement drives in Sect. 4.2.5.

#### 4.2.2 Impact of Subset Choice for Cross Validation

The results show clear differences in the supervised learning approach for different subsets in the cross validation. If the roads are divided into three sets according to their geographical locations, the ROC curve is clearly worse than a random split of roads into three sets as shown in Fig. 5a. The area under the curve is 0.953 compared to 0.966 for the random split of roads. Such clear differences between different subsets in the cross validation do not exist for k-means and WORT (see Fig. 5b, c). We assume that this effect is caused by different parking characteristics for the different regions of our evaluation data subsets. If the roads are chosen randomly for the

**Fig. 5** Comparison of different subsets for cross validation with **a** random forest, **b** k-means, and **c** WORT with smoothing. Only for the random forest calculations, a clear impact of different subset choice is visible



subsets, all subsets have about the same characteristics and are therefore more representative for the other subsets. Since the random forest classifier is able to learn finer differences than k-means with its linear separation hyperplane in the feature space, the results depend more on the representative choice of training data than the other methods.

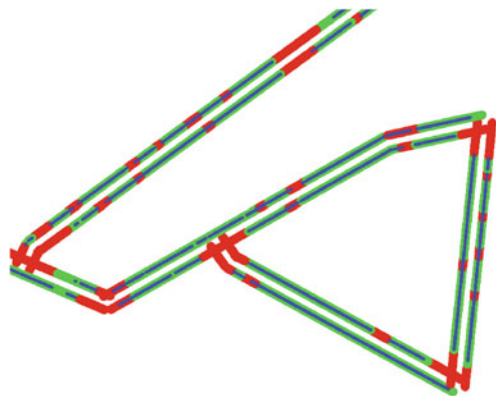
### 4.2.3 Qualitative Evaluation of Methods

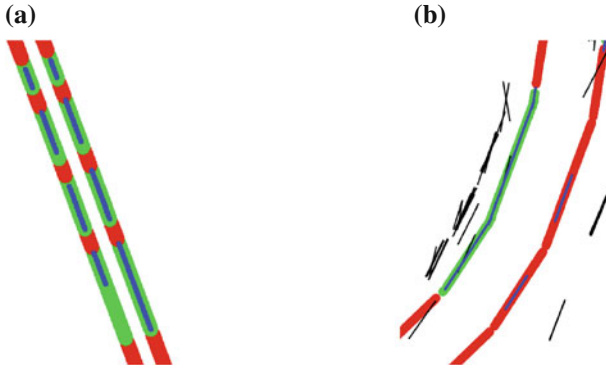
Qualitative comparison of the results reveals several similarities. All methods provide a reliable decision for parking legality in most situations. In particular, long parking lanes and highly occupied parking spaces are well identified. However, false positives mainly occur in small areas of parking prohibition like in front of garage entrances if some vehicles are parked there during observation time (e.g. right road in Fig. 6). False negatives are less frequent. They appear at rarely parked places like at the end of parking lanes (e.g. Fig. 7a). Also, at few places, the detection method systematically misses parked vehicles leading to locations without parked vehicles at any time instance. The algorithms differ in determining the beginning of illegal zones. The random forest approach often still classifies a few meters as legal in the illegal zone. In situations with only one vehicle parked illegally for a few hours, k-means and random forest interpret these situations correctly while the WORT approach marks these spots as legal (see Fig. 7b).

### 4.2.4 Evaluation of Features

To compare the relevance of the feature sets, we evaluate their impact both for unsupervised and supervised learning. For unsupervised learning with k-means clustering, we compare the results for runs with all and with a subset of feature sets (see Fig. 8). If only feature sets 1–3 are used, i.e. the neighborhood features are ignored,

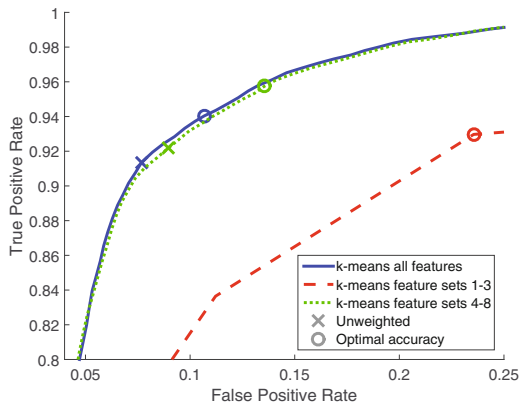
**Fig. 6** Example for the resulting parking map. *Blue* is the estimated parking space, *red* (illegal) and *green* (legal) show the ground truth





**Fig. 7** Examples for wrong results (estimated parking space in *blue*, ground truth for legal/illegal parking in *green/red*): **a** shows false negative results at the end of a parking zone. **b** Visualizes two false positive parking spaces for WORT with smoothing. The *black lines* represent the raw vehicle detections. At the two wrong parking spaces, the same vehicle was detected two and four times, respectively

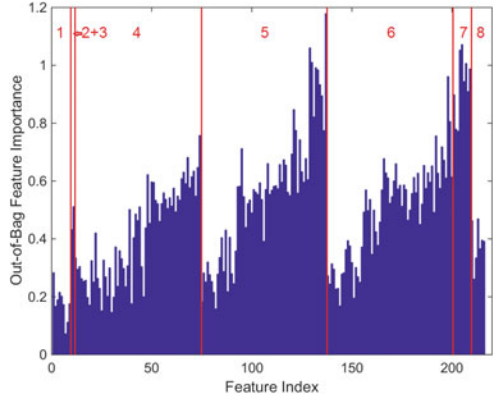
**Fig. 8** ROC curve for different feature sets with k-means



the resulting ROC curve is considerably worse than the curve for all features. The main reason for this result stems from marked parking spaces where the gaps between the parked vehicles are always at the same positions and cars rarely cover that space. If only the neighborhood features (feature sets 4–8) are used, the result is very similar to the usage of all features. Most differences are only at the end of parking lanes, where this result is less accurate than using all features.

For supervised learning, the random forest method has the advantage of already providing the feature importance already after the training step because it leaves out a part of the training data for each tree. A plot of the feature importance is given in Fig. 9. The most important feature sets are the average neighbor occupancy (5), weighted neighbor occupancy (6), and segment saturation (7). For the feature sets 4–6, we see an increasing trend for increasing feature index. In these cases, the maximal distance of the neighborhood is chosen increasingly. This means that the first

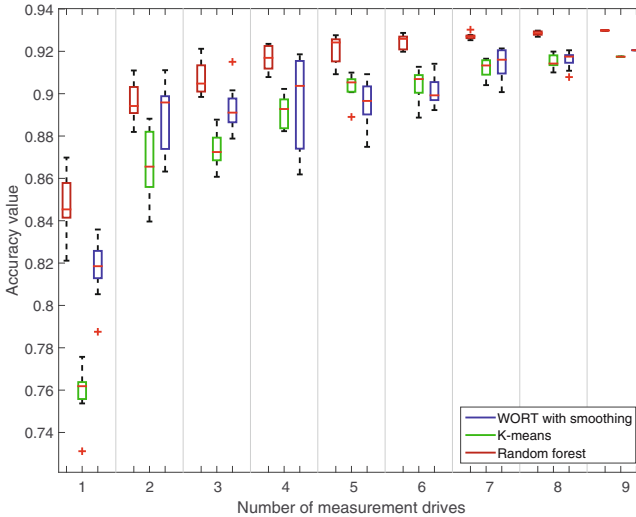
**Fig. 9** Feature importance from random forest training. The intervals of the feature index correspond to the feature sets separated and named in red



values are for very short distances (in this case 0.5 m) and the last values are for long distances of 40 m. So, the neighborhood features are more important for farther distances, but still relevant for shorter distances.

#### 4.2.5 Evaluation of Required Number of Measurement Drives

To investigate the influence of the number of measurement drives on the parking map result, we compare the presented methods for every number of measurement drives using nine random drive subsets. The result is shown in a boxplot in Fig. 10. The



**Fig. 10** Boxplot showing the accuracy of different methods for different numbers of measurement drives. Note that for nine drives, there is no variation for different permutations since all drives are used in the calculation

plot reveals mostly increasing values with increasing number of measurement drives for all methods. For a low number of measurement drives, the random forest shows significantly better results than the other two methods. If only one measurement drive is used, k-means is also clearly worse than WORT with smoothing. For seven or more measurement drives, only small improvements can be observed for all approaches.

## 5 Conclusion

This paper presents a novel approach for the generation of on-street parking maps from parked vehicle positions using supervised and unsupervised learning methods. We propose multiple feature sets to describe the occupancy characteristics of small road subsegments and their surroundings. Furthermore, we compare our methods to an implementation of the method from Coric and Gruteser (2013). Parked vehicle detections from repeated LiDAR measurements are used to evaluate the methods on more than 5 km of potential parking space. We have shown that both of our approaches show slightly better results than the method from the literature while keeping the model more generic. Most interestingly, the main advantage of our unsupervised approach is the total avoidance of parameter choice and optimization while still providing results comparable to the supervised learning. Also, it is very robust against the variation of parking characteristics for different areas. The random forest method also provides reliable results in general and the best results for low numbers of measurement drives. However, it reveals a clear dependence on the representative choice of the training data set.

All approaches show weaknesses for untypical input data. If a legal parking space is never occupied in the data set, it can hardly be identified. The same holds for parking prohibition areas which are occupied by parked vehicles most of the time. The latter often leads to wrong results at garage entrances. Since the evaluation is based on data from only one day, the data set is biased for situations where parking spaces are occupied for a long time by the same vehicle.

In the future, we plan to investigate more elaborate clustering algorithms like EM clustering to further improve the result for the unsupervised approach. Furthermore, an extension of the approaches for more parking classes like special parking legislation (e.g. parking for handicapped people) is an interesting challenge. Since the occupancy characteristics are less distinct in this case, the supervised approach is assumed to be superior over the other described approaches. Finally, it would be very interesting to evaluate our approach for low-cost automotive ultrasonic sensor data. Since more and more sensor-equipped vehicles are able to communicate their data to a server, our methods hold high potential to provide up-to-date and inexpensive on-street parking maps in an industrial scale.



**Acknowledgments** 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 city's future road traffic, through cooperative approaches. This support is gratefully acknowledged.

## References

- Bengler K, Dietmayer K, Farber B, Maurer M, Stiller C, Winner H (2014) Three decades of driver assistance systems: review and future perspectives. *IEEE Intell Transp Syst Mag* 6(4):6–22
- Bock F, Eggert D, Sester M (2015) On-street parking statistics using lidar mobile mapping. In: 2015 IEEE 18th international conference on intelligent transportation systems. IEEE, pp 2812–2818
- Breiman L (2001) Random forests. *Machine learning*, pp 5–32
- Choi J, Chang E, Yoon D, Ryu S, Jung H, Suhr J (2014) Sensor fusion-based parking assist system. Technical Representative, SAE Technical Paper
- Coric V, Gruteser M (2013) Crowdsensing maps of on-street parking spaces. In: 2013 IEEE International conference on distributed computing in sensor systems. IEEE, pp 115–122
- Davies J, Beresford A, Hopper A (2006) Scalable, distributed, real-time map generation. *IEEE Pervasive Comput* 5(4):47–54
- Furgale P, Schwesinger U, Ruffli M, Derendarz W, Grimmert H, Muhlfellner P, Wonneberger S, Timpner J, Rottmann S, Li B, Schmidt B, Nguyen TN, Cardarelli E, Cattani S, Bruning S, Horstmann S, Stellmacher M, Mielenz H, Koser K, Beermann M, Hane C, Heng L, Lee GH, Fraundorfer F, Iser R, Triebel R, Posner I, Newman P, Wolf L, Pollefeys M, Brosig S, Effertz J, Pradalier C, Siegwart R (2013) Toward automated driving in cities using close-to-market sensors: an overview of the v-charge project. 2013 IEEE intelligent vehicles symposium (IV). IEEE, Iv, pp 809–816
- Ge Y, Xue W, Shu Z (2013) Improved system for parknet mobile network. Proceedings of the FISITA 2012 world automotive congress SE—11. Lecture notes in electrical engineering, vol 200. Springer, Berlin, pp 131–145
- Mathur S, Jin T, Kasturirangan N, Chandrashekharan J, Xue W, Gruteser M, Trappe W (2010) Parknet: drive-by sensing of road-side parking statistics. In: Proceedings of the 8th international conference on mobile systems, applications, and services—MobiSys '10
- Park Wj, Kim Bs, Seo De, Kim Ds, Lee Kh (2008) Parking space detection using ultrasonic sensor in parking assistance system. In: 2008 IEEE intelligent vehicles symposium, pp 1039–1044
- SFMTA (2014) SFpark Putting Theory Into Practice
- Shoup DC (2006) Cruising for parking. *Transp Policy* 13(6):479–486
- Shoup DC (2007) Cruising for parking. *Access* 30:16–22
- Stenneth L, Wolfson O, Xu B, Yu PS (2012) Phonepark: street parking using mobile phones. In: 2012 IEEE 13th international conference on mobile data management, pp 278–279
- Thornton DA, Redmill K, Coifman B (2014) Automated parking surveys from a LIDAR equipped vehicle. Transportation research part C: emerging technologies 39:23–35
- Thrun S (2002) Robotic mapping: a survey. In: Nebel B, Lakemeyer G (eds) Morgan kaufmann, exploring artificial intelligence in the new millenium
- Van Ommeren JN, Wentink D, Rietveld P (2012) Empirical evidence on cruising for parking. Transportation research part A: policy and practice 46(1):123–130
- White P (2007) No vacancy: park slope's parking problem and how to fix it. Transaltorg

# Visualizing Location Uncertainty on Mobile Devices: Assessing Users' Perception and Preferences

Champika Manel Ranasinghe and Christian Kray

**Abstract** Location information is rarely perfectly accurate: usually it is subject to variations, errors and uncertainty, which may affect the quality of location-based services such as Pedestrian Navigation Systems (PNS). Visualizing location uncertainty is one option to address this issue. However, it is unclear how users interpret these visualizations. This paper investigated whether different types and styles of visualizing location uncertainty on mobile devices have an impact on user's perceptions, and which options they prefer. We also proposed a new visualization (cloud shape) to represent location uncertainty and compared it to the two existing shapes (circle and colored street segments—CSS). Results indicate that the design and the style of a visualization influence users' understanding of where they are located. More importantly, results indicated that the cloud could be a better option to visualize the uncertainty if the uncertainty of location information is very high. These findings can be used by the designers to modify/change the shape and style of the visualization to provide users with a more accurate picture of the quality of location information.

**Keywords** Visualizing uncertainty · Localization · Pedestrian navigation · Tracking · Mobile human computer interaction

## 1 Introduction

With the exponential growth of smartphones equipped with Global Navigation Satellite Systems (GNSS) sensing capability, GNSS-based pedestrian navigation has become ubiquitous. Pedestrian Navigation Systems (PNS) that are available with

---

C.M. Ranasinghe (✉) · C. Kray  
Situated Computing and Interaction Lab, Institute for Geoinformatics,  
University of Muenster, Muenster, Germany  
e-mail: champika.manel@uni-muenster.de

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

© Springer International Publishing Switzerland 2016  
T. Sarjakoski et al. (eds.), *Geospatial Data in a Changing World*,  
Lecture Notes in Geoinformation and Cartography,  
DOI 10.1007/978-3-319-33783-8\_18

modern mobile phones are now widely used to find the current location and to navigate to desired locations. However, the accuracy of the position calculated using the information obtained from GNSS is variable due to many technical, environmental and contextual factors (Lachapelle 2007). For example, GNSS is subject to reflections and intermediate unavailability in urban areas due to the tall buildings or overpasses. The quality of the navigation support that users receive is subject to such variability as well, and it might have a major impact on the user experience and the system reliability. For example, if users notice that the current location displayed on screen is not their actual location, they might lose trust in the system and become reluctant to rely on it in the future.

Benford et al. (2006) argue that uncertainty is a fundamental characteristic of location information rather than an exception, and they suggest dealing with uncertainty instead of trying to treat it as a bug. They propose five ways to deal with uncertainty: removing it, hiding it, managing it, revealing it and exploiting it. A lot of research has focused on removing or minimizing uncertainty by improving the overall accuracy, for example by enhancing sensor technology (use of a better antenna or GNSS chip) or by fusing different sensors such as WiFi, cellular phone network and accelerometers (Arikawa et al. 2007; Djuknic and Richton 2001; Kouroggi et al. 2006). However, it is unlikely that a universal solution capable of removing all the inherent uncertainty under all circumstances will ever emerge. Therefore, there is a need for alternative ways of dealing with uncertainty.

Although 'revealing location uncertainty' has not received much attention, it has been shown that the awareness of location uncertainty is beneficial to users (Aksenov et al. 2012; Dearman et al. 2007). In line with this, designers now pay more attention to convey the uncertainty in location information in order to improve the user experience and the trust in the system. Consequently, most current systems incorporate a simple visualization to convey location uncertainty. This option enables users to perceive the uncertainty associated with their current location. The simple visualization that current systems use usually consists of a dot that represents the last known position or the approximate position in the middle and a circle around it to represent the location error or the uncertainty. This circle may have a border, and it is often colored in a semi-transparent shade. The size of the circle grows with the location error.

Different systems have used different variations of this representation. For example, Dearman et al. (2007) used a circle with a middle dot and a border, Burigat and Chittaro (2011) used a middle dot without a border, Baus et al. (2002) used a border without a middle dot, and Lemelson et al. (2008) uses neither a middle dot nor a border. However, there is no clear understanding or underlying theory that provides a sound basis for choosing amongst these visualization options. Furthermore, there is an open question whether these visualization options have an impact on users' understanding of the uncertainty associated with their current location. In addition, research investigating shapes other than the circle to represent location uncertainty is scarce (Lemelson et al. 2008; Burigat and Chittaro 2011). Exploring these aspects may help in designing more meaningful and effective location uncertainty visualization techniques for mobile phones.

The goal of this paper is to address this knowledge gap. In particular, we report on our investigation of the following two questions: (1) Does the use of a middle dot and/or a border in a visualization have an impact on the users' understanding of the uncertainty associated with their current location? (2) Can a cloud-like shape relay location uncertainty as well or better than standard visualizations, and what options do users prefer? The rest of the paper is organized as follows: Sect. 2 presents related work on location sensing and uncertainty visualization. Section 3 discusses and classifies existing uncertainty visualization techniques and highlights two aspects that require further research. Section 4 presents the details of the user study we conducted. A discussion of our findings is presented in Sect. 5. Section 6 contains the concluding remarks and highlights future work.

## 2 Related Work

While the idea of visualizing uncertainty in location information is not new, it has received little attention until recently. Early systems such as GUIDE (Cheverst et al. 2000) did not even use GPS: they relied on a 'Bar of Connectivity' to indicate the reception of location information and a 'Location Status Window' that shows the last known location and the time in minutes that has elapsed since it was measured. REAL (Baus et al. 2002) used a dot to represent the current GPS location whose size increased with decreasing quality of location information. LOL@ (Pospischil et al. 2002) displayed location accuracy as a tool tip text attached to the current position symbol. They also proposed the use of a disc shape whose radius corresponds to the current accuracy.

Recent research has focused on studying visualization of location information in more detail. Dearman et al. (2007) compared four different ways of presenting the location error: only the predicted location, a fixed size circle of 95 % confidence, a variable sized circle of N % confidence (where the user has to select N), and a variable sized circle whose size is determined by the optimal confidence calculated by the system. In addition to the ring (circle border), they labelled the middle of the ring as 'Me'. This is similar to the use of a middle dot. Their study showed that the users do benefit from error visualizations compared to non-visualization. In addition, the optimal error visualization was shown to positively and strongly influence the search strategies of the users. (In the optimal error condition, the size of the circle changes according to the current location error). If the error was small, the circle was small as well whereas if the error was large, the circle was large. Therefore, users could limit their search space according to the true error. In contrast, in all the other three visualizations, users always had to search in an area of constant size.

Lemelson et al. (2008) investigated four different visualization techniques for depicting the position as well as the position error: (i) displaying the associated error as an area, around the current predicted position (for example as a circular area, in this case they neither use a middle dot nor a border); (ii) textually displaying the position and the error; (iii) displaying the current position on the map and showing the

error graphically (stylized traffic lights, analog round gauge, analog bar gauge) next to the map; and (iv) displaying the cumulative error distribution plot. Their results revealed that the users preferred the visualization of the positional error together with the position in the map itself (technique (i)).

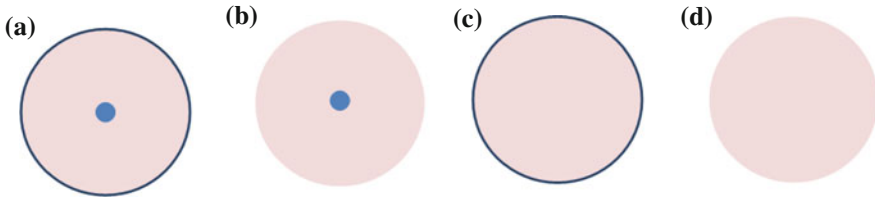
Burigat and Chittaro (2011) evaluated three different uncertainty visualization techniques: just the last known position (basic visualization), circle of variable size whose size dynamically changed based on the location error, and colored street segments. For the last option, those street segments that the user might be in were dynamically colored based on the error. In the first representation (circle), they used a middle dot without border. The study revealed that the street coloring visualization was perceived to be more beneficial by the users than the basic visualization. In addition, the overall workload and the mental demand for the street coloring visualization were found to be lower than those for the basic visualization. Furthermore, both the circle and the street coloring visualizations required a lower effort compared to the basic visualization. However, there was no significant difference between the three visualizations in the accuracy with which users assessed their position.

Aksenov et al. (2011) also used a circular shape (with a middle dot and a border) to visualize uncertainty in location information. In addition to the uncertainty, they visualize the reasons for such uncertainty using four different versions of the circle (changing size, pulsating middle dot, changing the middle dot to a cross, removing the border). Results of the study provide evidence for users perceiving visually more demanding ways of displaying location uncertainty as beneficial as they relay useful information about what led to the uncertainty.

From this short review of related work, we can conclude that the main technique to visualize location uncertainty is the use of a circle whose size increases with increasing location error. There is a  $K\%$  confidence that the user's current position is anywhere inside the circle. In practice, it is the case that in many off-the-shelf commercial systems  $K$  is a constant number. For example, in Android-based systems,  $K$  is  $68\%$  (AOSP 2015). In addition to the circle, the street coloring visualization proposed by Burigat and Chittaro (2011) constitutes an interesting alternative that was perceived as beneficial by users. This approach colors and shows only the street segments in a circular area determined by the location error. In the next section, we will look at different visualization options in more detail.

### 3 Categorizing Options to Visualize Location Uncertainty

While 'Coloured Street Segments' (CSS) (Burigat and Chittaro 2011) appear to be a promising option, the circular visualization is by far the most widely used one. For the latter one, four fundamental visualization options are available (as depicted in Fig. 1). The circle can be shown (1) with border and middle dot (e.g. Dearman et al. 2007; Aksenov et al. 2011); (2) with middle dot but no border (e.g. Burigat and Chittaro 2011); (3) with border but no middle dot; and (4) without border and no middle



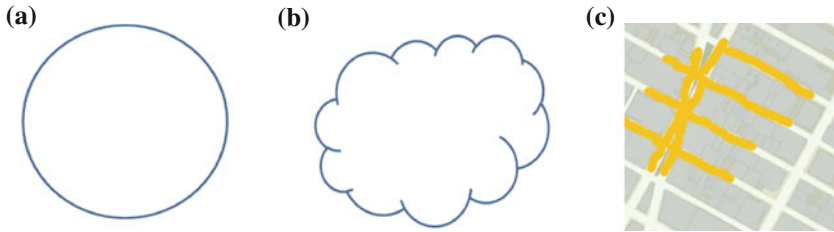
**Fig. 1** Key visualization options. **a** Border, dot (BD). **b** No border, dot (ND). **c** Border, no dot (BN). **d** No border, no middle dot (NN).

dot (e.g. Baus et al. 2002; Lemelson et al. 2008). These four visualizations can be classified into two classes: Middle Dot representations and Border representations.

The middle dot is used to indicate the approximate position, or the last known position. The size of the circle usually represents the level of uncertainty: the larger the circle, the higher the uncertainty. The level of uncertainty is determined either by the location error value (Burigat and Chittaro 2011) or by the time that has been elapsed since the last update (Aksenov et al. 2011). Previous work does not provide a rationale for the decision of whether a border is present or not: some studies/systems used a border while some did not.

When analysing existing uncertainty visualization techniques, we identified two main areas that require further research. Firstly, it is unclear what users actually understand when they see these different visualizations, and what mental model users construct in seeing these visualizations. Further research on this aspect could help to design more effective and more meaningful error visualization techniques. In addition, a deeper understanding of what people make of the visualizations could be useful in developing systems that adapt to varying quality of location information. Secondly, there has been little research on the feasibility of using shapes other than the circle for location uncertainty visualization. There may be alternative shapes that better represent the uncertainty and convey it more clearly to users so that they could apply more effective search strategies. Therefore, further research on this aspect could also contribute to designing more user-friendly error visualization techniques that improve user experience.

One possible alternative shape could be a cloud (see Fig. 2b). Compared to the crisp and homogeneous shape of a circle, the shape of a cloud is more vague and in sketching, it is often used to indicate something that is uncertain or not so clearly defined. Therefore, a cloud might also be well-suited to convey a feeling of uncertainty with respect to the location. However, the circle has a constant distance (radius) from its center to the boundary in all directions, which makes it very easy to draw a representative circle with a single location error value. In contrast, the distance from the middle to the boundary of a cloud shape is variable in different directions. Therefore, an approach is needed similar to what Burigat and Chittaro (2011) used to create CSS, where an underlying circular shape can be used to draw a representative CSS. The following sections will present a user study aimed at assessing the viability of a cloud shape for representing location vagueness. A second key



**Fig. 2** The three basic shapes considered in the study: *Circle*, *Cloud*, *Colored Street Segments (CSS)*

goal is to evaluate whether different visualization options do have an impact on users' understanding of where they think they are.

## 4 Comparison Study

For the comparison study, we considered three shapes: Circle, Cloud and CSS, depicted in Fig. 2. In addition, we selected four different visualization options: Border and middle dot (BD), no border but middle dot (ND), border but no dot (BN), and no border no middle dot (NN). These options represent the two classes of visualizations, Middle Dot and Border as shown in Fig. 1. In the study, we wanted to assess users' preferences with respect to the three shapes (Circle, Cloud, CSS) and with respect to different visualization options (BD, BN, ND, NN) of a given shape. In addition, we wanted to find out if the different shapes and visualization options have an impact on what users think where they are likely to be, and on the perceived accuracy. The study thus tries to answer the following two research questions:

**RQ1:** Is there a difference in preference between the three shapes (Circle, Cloud and CSS), and is there a difference in preference between different visualization options (BD, ND, BN, NN)? For example: Do people like the cloud shape better than circle? Do people prefer the middle dot representation over the border representation?

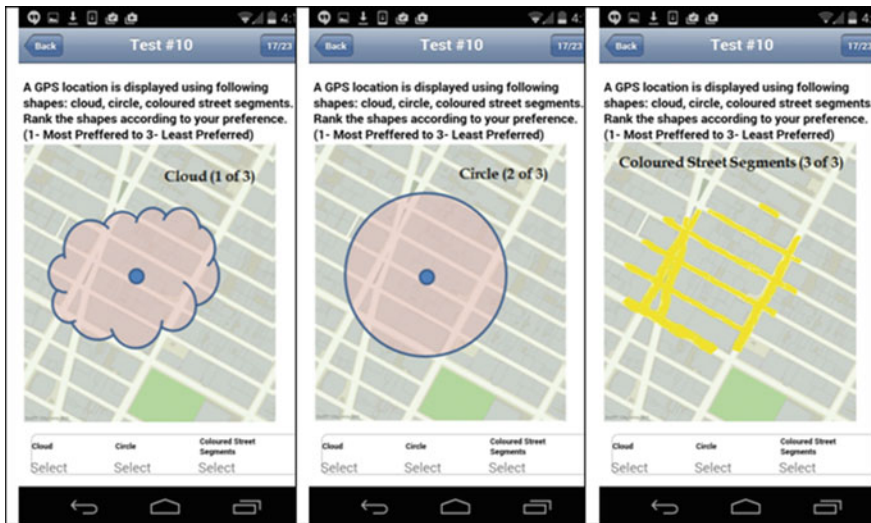
**RQ2:** Is there an impact of different shapes and different visualization options on people's understanding of where they are likely to be located? For example: Is there a difference in what people understand when they see a cloud instead of a circle? Is there a difference when the middle dot or a border is present?

### 4.1 Design

The study was designed as a within-subject lab-based study. We used a mobile phone to display different types of shapes and visualization options on an unlabeled map

along with questions. The questions were aimed at answering the two research questions. As an initial step towards addressing RQ1, we compared the users' preferences for the three shapes, Circle, Cloud and CSS. Users were presented with the three shapes as an animated slide show on a single screen, and they were asked to rank them according to their preference. The maps were extracted from Google (2011) and then processed to remove street names to avoid any bias. The three shapes were shown one at a time over the same map as a repeating slide show. A Likert scale at the bottom was continuously visible to enable participants to indicate their preferences. The question was repeated with the four different visualization options, BD, ND, BN and NN. However, the CSS option was presented without these visualization options because the initial design proposed by Burigat and Chittaro (2011) did not contain a border or a dot in the CSS. Consequently, there were four questions of this nature. This type of questions were classified as 'Type 1 Questions' (cf. Fig. 3 for an example).

To investigate preferences for different visualization options for a given shape, the four visualization options (BD, BN, ND and NN) were presented as an animated slideshow on a single screen. Users were asked to rank the visualization options according to their preference (1-Most preferred, 4-Least preferred). There were two questions of this nature, one for the circle and one for the cloud. This type of questions were classified as 'Type 2 Questions'. Figure 4 shows an example for a question

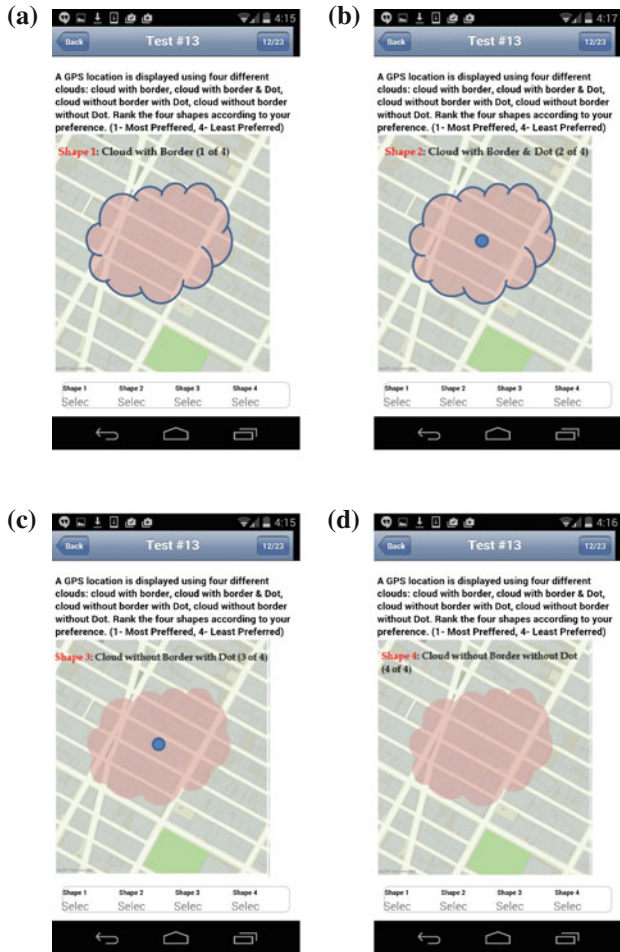


**Fig. 3** Comparing the three shapes: Example map with the three shapes (*Cloud*, *Circle* and *CSS*) with visualization option BD; the three shapes were shown as a repeating slideshow until participants answered the ranking question shown at the *bottom* of the screen



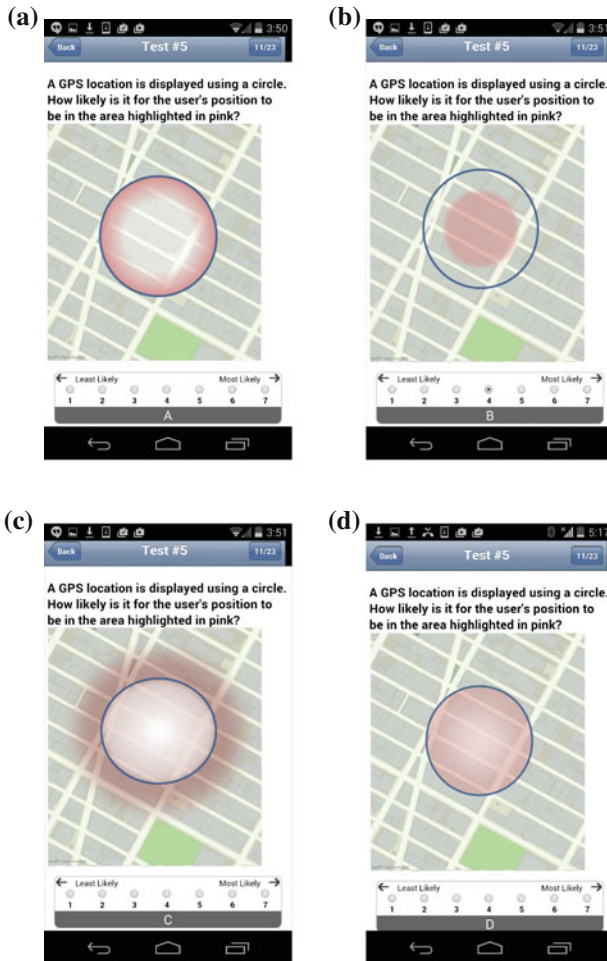
of this type: the four screenshots (one for each of the four shapes resp. four visualization options) were displayed as a repeating slide show.

In order to investigate RQ2, four different visualizations (BD, ND, BN, NN) of a given shape (either circle or cloud) were presented to the user. Within each visualization, there was one of four different regions highlighted. These regions are ‘Inside Border’ (IB), ‘Center Inside’ (CI), ‘Outside Border’ (OB) and ‘All Inside’ (AI). Users were asked to mark the likelihood of being in each region on a Likert scale of 1 to 7 (1-Least likely, 7-most likely). The question was repeated for the two shapes, circle



**Fig. 4** The four visualization options (BN, BD, ND, NN) for the *Cloud* shape (shown as a slide show). **a** Visualization = BN. **b** Visualization = BD. **c** Visualization = ND. **d** Visualization = NN

and cloud, and for all the four different visualization options, BD, ND, BN and NN. Consequently, there were eight questions of this nature. Within a given question, the highlighted regions were displayed in random order. Figure 5 depicts a question of this nature where the shape is a circle and the visualization option is BN. This type of questions was classified as ‘Type 3 Questions’.



**Fig. 5** Likelihood of being in a particular region for a given visualization of a given shape. In this example, shape is *circle*, visualization is BN, and the order of regions is IB, CI, OB, AI. (The Likert scale at the *bottom* is used to mark the likelihood). **a** Region = IB. **b** Region = CI. **c** Region = OB. **d** Region = AI.

Furthermore, users were presented with two different sizes (Small and Large) of a given shape (Circle, Cloud or CSS) and were then asked to select the most accurate and the least accurate representation from the two according to the perceived accuracy. The question was repeated for all the three shapes (Circle, Cloud, CSS) and for all the four different visualizations (BD, BN, ND, NN). Therefore, there were nine different questions of this nature. This type of questions were classified as 'Type 4 Questions'.

## 4.2 *Participants*

We recruited 35 users, 27 male and eight female, to participate in the study. However, the data gathered from three of the male users had to be discarded due to incompleteness thus reducing the effective number of participants to 32 (24 male and eight female). The age of these 32 participants ranged from 19 to 42 averaging at 23. Thirty of them owned a smartphone while the remaining two did not. Fifteen participants used their smartphone regularly to navigate in unfamiliar environments, and 11 of them did so sometimes. Four people stated that they rarely used their smartphone for this purpose. There was only one person who never used the phone to navigate. Eleven rated their English proficiency as 'Very Good', 13 as 'Good' and the remaining eight as 'Moderate'. (The study was conducted in English in an international environment.)

## 4.3 *Materials*

The study was carried out on a Google Nexus 5 Android mobile phone with 4.95-inch touchscreen. The screen resolution was  $1080 \times 1920$  pixels. Prior to the test, the following data was gathered from the participants through a web-based system: (1) demographic data: nationality, age, gender, English proficiency. (Names of the participants were not recorded); and (2) technology familiarity data: data about whether they own a smartphone and how often they use the smartphone to navigate in an unfamiliar environment. The user was then directed to the user study, which consisted of 23 questions in total. There were four types of questions: Type 1, Type 2, Type 3 and Type 4 as described in the design section. These types are summarized in Table 1.

The questions were randomly and dynamically generated for each participant using a web-based application developed using PHP, HTML, Apache and MySQL. For each type of question, the question was displayed at the top of the screen and the Likert scale or the ranking scale was displayed at the bottom of the screen. The visualization itself covered the middle portion of the screen, which took up about 80% of the total screen size. Users marked their preferences or ranks on the respective

**Table 1** Types of questions

Type	Description	No. of questions
1	3 shapes (Cloud, Circle, CSS) were displayed as an animation on the screen. At the bottom of the screen, there was a preference scale to rank the shapes according to the user's preference (See Fig. 3)	4
2	4 different visualizations (BD, BN, ND, NN) of a given shape (either circle or cloud) were displayed randomly as an animation on one screen. At the bottom of the screen, there was a preference scale to rank the shapes according to the user's preference (See Fig. 4)	2
3	4 different regions (AI, CI, IB, OB) of the same visualization (either of BD, BN, ND or NN) of a given shape (either circle or cloud) were displayed to the user. At the bottom, there was a Likert scale to indicate the likelihood of being in a given region on a scale of 1 to 7 (See Fig. 5)	8
4	Two different sizes of the same visualization (either BD, BN, ND or NN) of a given shape (either Circle, Cloud or CSS) were displayed as an animation on one screen. At the bottom, there was a scale to rank the shapes according to the perceived accuracy	9
	Total number of questions per participant (excluding demographic questions)	23

scales by touch. Participants were allowed to go back if they wanted to change their answers to previous questions. The answers were sent to an online database at the end of the test.

#### 4.4 Procedure

The study was conducted in a lab with only the experimenter being present besides the participant. After a brief explanation about the study users were asked to take an example test to familiarize themselves with the system. Once they were ready to answer the questions, they were allowed to take the test on the phone. The application first prompted them to enter personal data (demographic details and familiarity data). After they had done this, a summary of their personal data was displayed for them to confirm the correctness. Once they had confirmed this, they were presented with the 23 questions in randomized order. Visualizations for each question were also generated randomly. After participants had completed all the 23 questions, a completion screen was displayed. To motivate participants to participate in the study as well as per the standard practice, a small payment was given in return for their time spent on the study.

## 4.5 Results

The results were analysed with respect to the preferred shape (circle, cloud, CSS), to the preferred visualization option for a given shape (BD, BN, ND, NN), to the region where users thought that they were in when they saw a particular visualization, and with respect to the perceived accuracy when users saw the visualizations of different sizes (small and large). The following sections report on these results in more detail.

### 4.5.1 Preferred Shape: Circle, Cloud or CSS

This section presents the results of the analysis of the preferences recorded for the three shapes, circle, cloud and CSS (Type 1 questions).

Friedman's test revealed that there is a significant difference ( $\chi = 137.518, p = 0.000 < \alpha = 0.01$ ) between the scores obtained for different shapes (Cloud, Circle, CSS). A post-hoc analysis using the Wilcoxon Signed Rank test also confirmed that the circle was more preferred than the cloud ( $Z = -8.403, p = 0.000$ ) and the cloud was more preferred than the CSS ( $Z = -9.149, p = 0.000 < 0.05$ ). The mean rank obtained for each shape is, Circle = 1.21, Cloud = 2.14 and CSS = 2.65. The lower the rank, the more preferred the shape. Therefore, the order of preference for the shapes in general is: Circle > Cloud > CSS.

### 4.5.2 Preferred Visualization Option: BD, BN, ND or NN

The data was also analysed to find out whether there is a difference in preference between the four visualization options, BD, BN, ND and NN (Type 2 questions). Friedman's test revealed that the preferences obtained for different visualizations are significantly different within each shape, circle ( $\chi = 40.245, p = 0.000 < \alpha = 0.01$ ) and cloud ( $\chi = 45.123, p = 0.000 < \alpha = 0.01$ ). Follow up analysis for the circle using Wilcoxon Signed Rank test revealed that BD was preferred over all other visualizations (ND ( $Z = -1.963, p = 0.050$ ), BN ( $Z = -4.284, p = 0.000$ ), NN ( $Z = -4.518, p = 0.000$ )); ND was more preferred than BN ( $Z = -2.937, p = 0.003$ ) and there was no significant difference between BN and NN ( $Z = -1.646, p = 0.100$ ). Wilcoxon Signed Rank test for the cloud revealed that there was no significant difference between BD and ND ( $Z = -0.762, p = 0.446$ ); ND was more preferred than BN ( $Z = -3.383, p = 0.001$ ) and BD was more preferred than BN ( $Z = -4.502, p = 0.000$ ). However, there was no significant difference between BN and NN ( $Z = -1.855, p = 0.064$ ).

These results generally indicate that inclusion of the dot is preferred over no dot and furthermore, a dot with a border is preferred over a dot without a border.

### 4.5.3 Users' Understanding Regarding Where They Are

This section presents an analysis of the data gathered from Type 3 questions. There are two main objectives to this analysis. First, we wanted to know whether there is a difference in users' understanding with respect to where they are when they see a particular visualization, depending on the two shapes (circle, cloud). The four regions considered were, AI, CI, IB and OB (see Fig. 5). Second, we were interested to find out whether there is an impact of different visualizations (with-dot vs. without-dot, with-border vs. without-border) on users' understanding of where they are located with respect to these four regions.

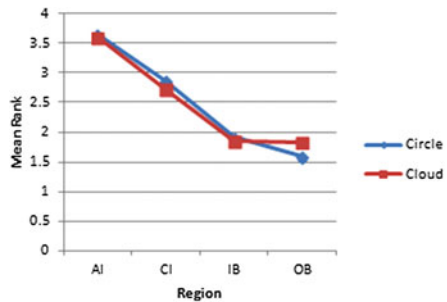
**Circle versus Cloud:** Friedman's test revealed that the scores obtained for the four regions are significantly different within each shape (circle ( $\chi = 51.115, p = 0.000 < \alpha = 0.01$ ), cloud ( $\chi = 42.555, p = 0.000 < \alpha = 0.01$ )). The mean ranks obtained for the regions for each shape are listed in Table 2 and are illustrated in Fig. 6.

This implies that for circle the order in which users think they are most likely to be in is: AI > CI > IB > OB. However, for the cloud, this is: AI > CI > either IB or OB, because the ranks for IB and OB are roughly the same. This implies that, when they see a cloud they do not see much difference in being in the IB and in the OB region. This is an indication to say that cloud is a better shape to represent uncertainty when the location uncertainty is very high. However, we need further analysis to support this.

**Table 2** Circle versus cloud: mean ranks obtained for the four regions

Mean rank		
Region	Circle	Cloud
AI	3.64	3.59
CI	2.86	2.73
IB	1.91	1.84
OB	1.59	1.83

**Fig. 6** Circle versus cloud: mean ranks obtained for the four regions (the higher the rank, the more the user thinks that he or she is likely to be in that region)



In order to further analyse whether there is an impact of using a middle dot and a border on each shape, we conducted a Friedman’s test for categories (different visualization options: BD, BN, ND, NN). The mean, standard deviation and mean rank for each category are summarized in Table 3. The summary statistics shown in the figure imply that there are differences in ranks for different visualizations.

**With-dot versus without-dot (within shape):** We applied Wilcoxon Signed Rank test to investigate whether there is difference in scores between the with-dot and without-dot representation within each shape. For the circle there was no significant difference in the scores obtained between the with-dot representation and the without-dot representation for the regions, CI, IB and OB. However, the scores obtained between the AI region with dot and the AI region without dot were significantly different ( $Z = -2.353, p = 0.019$ ), where the circle with dot obtained higher scores than the circle without dot. Similar results were obtained for the cloud as well (No difference between with-dot vs. without dot for CI, IB and OB regions, but a significant difference between the cloud with dot and cloud without dot for the AI region ( $Z = -2.531, p = 0.011$ )), where cloud with dot received higher scores than the cloud without dot.

This implies that the users associated a higher likelihood of being in the AI region in the with dot representation compared to the without dot representation. We observed this for both the circle and the cloud shape. This implies that, use of a middle dot with either shape will increase the likelihood that the users think that they are in the AI region. Consequently, the designers can use the middle dot to indicate high quality location information (to indicate that the location error/ uncertainty is small).

**With-border versus without-border (within shape):** We applied Wilcoxon Signed Rank test to investigate whether there is a difference in scores obtained between the with-border representation and without-border representation within each shape.

**Table 3** Circle versus cloud: summary statistics for different visualizations, BD, ND, BN, NN

Shape	Region	With border						Without border					
		With dot			Without dot			With dot			Without dot		
		Mean	Std. dev	Mean rank	Mean	Std. dev.	Mean rank	Mean	Std. dev.	Mean rank	Mean	Std. dev.	Mean rank
Circle	AI	5.91	3.50	3.50	5.50	3.45	3.45	5.97	3.48	3.48	5.91	3.50	3.50
	CI	4.97	2.81	2.81	4.59	2.95	2.95	4.91	2.75	2.75	4.97	2.81	2.81
	IB	3.78	1.95	1.95	3.50	2.08	2.08	3.53	1.86	1.86	3.78	1.95	1.95
	OB	3.28	1.73	1.73	2.94	1.52	1.52	3.47	1.91	1.91	3.28	1.73	1.73
Cloud	AI	5.50	3.44	3.44	5.06	3.38	3.38	5.84	3.41	3.41	5.34	3.30	3.30
	CI	4.53	2.59	2.59	4.31	2.75	2.75	4.88	2.80	2.80	4.53	2.83	2.83
	IB	3.50	2.20	2.20	3.16	1.94	1.94	3.25	1.75	1.75	3.44	1.88	1.88
	OB	3.12	1.77	1.77	3.19	1.94	1.94	3.91	2.05	2.05	3.66	2.00	2.00

For the circle, there was a significant difference between the scores obtained for the IB and OB in with border representation ( $Z = -2.145, p = 0.032$ ). However, there was no significant difference between IB and OB in without border representation for the circle ( $Z = -0.834, p = 0.404$ ). For the cloud, there was no significant difference between the scores obtained for the IB and OB for both with border representation ( $Z = -1.035, p = 0.301$ ) and without border representation ( $Z = -1.1, p = 0.271$ ). In addition, OB region in the cloud without border representation scored significantly higher than the OB region in the cloud with border representation ( $Z = -2.862, p = 0.004$ ). These results imply that users do not see much difference in the likelihood of being in the IB and OB regions, when a border is not present (irrespective of the shape). More importantly, if the shape used is a cloud, users always did not see a significant difference between IB and OB (irrespective of whether a border is present or not). In addition, for the cloud, users associated a high likelihood of being in the OB region in the without border representation compared to the with border representation. In line with this, designers can use either without-border representation (cloud/circle) or a cloud (with/without border) to indicate high uncertainty in location information. This also implies that the cloud without border is more suitable to visualize higher levels of uncertainty.

#### 4.5.4 Size of Shape and Perceived Accuracy

This section presents the analysis of the data gathered from Type 4 questions. The objective of this analysis is to find out what users understand (with respect to the accuracy of the displayed position) when they see two different sizes (small and large) of the same shape. In general, Wilcoxon Signed Rank test revealed that there is a significant difference in the perceived accuracy between the two sizes, small and large ( $Z = -4.576, p = 0.000$ ). Users perceived that the small size represents higher location accuracy than the large size. Friedman's test revealed that there is no significant difference between the different visualization categories (with dot, without dot, with border, without border) in selecting the most accurate shape ( $\chi = 54.112, p = 0.000 < \alpha = 0.01$ ). This indicates that people generally associate smaller shapes with lower location uncertainty, regardless of which shape is used to visualize it.

## 5 Discussion

In the following, we discuss the results reported in the previous section in the light of the two research questions we had formulated. In addition, we provide our interpretation of the outcomes we recorded, and suggest ways in which our results can be used when designing visualizations for location uncertainty.

**Shape preference (RQ1):** The order of preference for the shapes is Circle > Cloud > CSS. A potential explanation for this ranking is the high degree of familiarity with



the circular visualization: most people have used a navigation system on their phones, which almost all use the circle shape. The unfamiliar cloud shape was also received well—it was ranked as the second best option.

**Preferred visualization options (RQ1):** Users generally preferred BD representations over all other representations. There was no difference between BN and NN. In addition, with-dot representations (either BD or ND) were preferred over border only representations (BN). This implies that, the use of a border was received positively only when it was used in combination with a dot. For the dot, the question remains whether it was preferred because people associate it with higher likelihood of being in the AI region. Generally speaking, our results do not provide clear evidence whether preferences for specific shapes and/or visualization options result purely from their visual properties, from the assumed accuracy of the displayed location or a combination of both. This aspect will have to be investigated further by follow-up studies.

**Perceived accuracy with respect to the size of the shape (RQ2):** Our results clearly support the assumptions that users associate the size of the shape with the level of accuracy (small shape: more accurate, large shape: less accurate). Furthermore, middle dot or the border had no impact on this decision. Consequently, the size of a shape seems to be the only factor that people use to form an understanding of the accuracy of the sensed location. Designers are thus free to use the other options (dot, border, shape) to convey other aspects of location uncertainty.

**What users think where they are (RQ2):** Overall, participants thought they were most likely somewhere inside the AI region. This was true for both Circle and Cloud. In addition, they thought that the second most likely region is CI (for both Circle and Cloud). They ranked the IB and OB regions as least likely regions.

It is evident that the border and the shape has an impact on the scores obtained for IB and OB regions. For the circle, if a border is present, users thought that it is more likely to be in the IB region than in the OB region. However, if a border is not present, they did not see a significant difference between the IB region and the OB region for the Circle. For the cloud, users never saw a difference between IB and OB regions irrespective of whether a border is present or not. In addition, users associated a high likelihood of being in the OB region in the cloud without border representation compared to cloud with border representation. This implies that, in general, while users thought it is most likely to be in the AI region and then in the CI region, they did not distinguish between areas just inside and just outside of the shape if a border is not present or if the shape used is a cloud. This observation could be used by a system to convey more uncertainty: if the uncertainty associated with the positional information is very high, use a shape without a border or use a cloud to represent the location uncertainty. In addition, the observations about the cloud indicate that the cloud can be a better shape to represent higher uncertainty (for example, if the uncertainty is very high, a cloud without a border could be the most appropriate visualization to convey the uncertainty).

Furthermore, it is evident that the middle dot did have an impact on the score obtained for the AI region: the AI region with dot received higher scores than the AI region without dot. But for all the other regions, the presence of a middle dot

did not have an impact on the user's ratings. Consequently, a system could use this observation to only show a middle dot when the GPS data indicates a very high likelihood of the user being in the AI region.

Generally speaking, we can observe several differences in preferences and perceived likelihood depending on the shape, its size and the presence of various graphical elements. While our study was certainly not exhaustive in terms of shapes, sizes and possible graphical variations, it nevertheless provides some initial evidence that these aspects have an impact on where people think they are located. Consequently, application developers can now take these insights into account when selecting a specific uncertainty visualisation in order to ensure that what people think where they are corresponds more closely to what the location sensor data predicts.

## 6 Conclusions and Future Work

The study presented in this paper investigated whether different types and styles of visualizing location uncertainty have an impact on what users perceptions, and which options they prefer. We also proposed a new visualization (cloud shape) to represent uncertainty and compared it to the two existing shapes (circle and colored street segments—CSS). Our comparison considered four visualization options: boundary and dot (BD), dot without boundary (ND), boundary without dot (BN) and neither boundary nor dot (NN). Our results indicate that the circle was preferred over all other shapes.

From the four visualization options, with-dot representations (BD, ND) were preferred over the without-dot representation (BN, NN). Furthermore, for the circle BD was preferred over all other visualizations whereas for the cloud, there is no difference between BD and ND.

With respect to the users' understanding of their current location, they thought it to be most likely to be anywhere inside the shape (AI region). If a middle dot was present, this likelihood increased even further. The perceived likelihood generally decreased from the center to the outside. However, users did not see any difference between the area just inside and just outside of the shape if a border was not present or if the shape used is a cloud. Furthermore, if a cloud without a border was used, users perceived that there is a high tendency to be in the area just outside compared to the area just inside the shape.

Our findings thus indicate that the design and the style of a visualization influence users' understanding of where they are located. Therefore, designers could modify/change the shape and style of the visualisation to provide users with a more accurate picture of the quality of location information. For example, designers can: use a middle dot when the quality of the location data is very high, remove the border when the quality of the location data is low, use a cloud without a border if the quality of location data is very low. However, these need to be confirmed by a deployment study.

The reason for the higher preference for the circle might be the participants' familiarity with mobile navigation applications: 94 % of the participants owned a smartphone, and 94 % of those who owned a smartphone used it to navigate in unfamiliar environments. The most popular navigation apps for those smartphones use a circle to visualize location uncertainty. In order to assess whether the circle is truly 'better' than other shapes, it would be useful to expose people to the alternatives for longer periods of time, i.e. via an extended deployment study.

In future, we are planning to carry out a field study to also assess the interaction between people's immediate physical environment and the visualization of location uncertainty shown on their mobile phone. This could also be a first step towards assessing the effectiveness of the different shapes and provide opportunities for participants to get used to the unfamiliar shapes. Additionally, we are currently preparing a repetition of our study in a different country to facilitate a cross-cultural comparison.

## References

- Aksenov P, Luyten K, Coninx K (2011) A unified scalable model of user localisation with uncertainty awareness for large-scale pervasive environments. In: Proceedings of next generation mobile applications, services and technologies, NGMAST 2011. Cardiff, UK
- Aksenov P, Luyten K, Coninx K (2012) O brother, where art thou located?: raising awareness of variability in location tracking for users of location-based pervasive applications. *J Location Based Serv* 6(4):211–233
- Android Open Source Project (2015) Location. <http://developer.android.com/reference/android/location/Location.html>. Accessed 30 Jan 2015
- Arikawa M, Konomi SI, Ohnishi K (2007) Navitime: supporting pedestrian navigation in the real world. *IEEE Pervasive Comput* 6(3):21–29
- Baus J, Krüger A, Wahlster W (2002) A resource-adaptive mobile navigation system. In: Proceedings of the 7th international conference on Intelligent user interfaces. ACM, pp 15–22
- Benford S, Crabtree A, Flinham M, Drozd A, Anastasi R, Paxton M, Tandavanitj N, Adams M, Row-Farr J (2006) Can you see me now? *ACM Trans Comput-Human Interact (TOCHI)* 13(1):100–133
- Burigat S, Chittaro L (2011) Pedestrian navigation with degraded gps signal: investigating the effects of visualizing position uncertainty. In: Proceedings of the 13th international conference on human computer interaction with mobile devices and services, MobileHCI 2011. ACM, pp 221–230
- Cheverst K, Davies N, Friday A, Efstratiou C (2000) Developing a context-aware electronic tourist guide: some issues and experiences. In: Proceedings of CHI 2000. Netherlands, pp 17–24
- Dearman D, Varshavsky A, De Lara E, Truong KN (2007) An exploration of location error estimation. Lecture notes in computer science. Springer, Heidelberg
- Djuknic GM, Richton RE (2001) Geolocation and assisted GPS. *Computer* 34(2):123–125
- Google (2011) Google maps for mobile. <http://www.google.com/mobile/maps>. Accessed 14 Feb 2011
- Kouroggi M, Sakata N, Okuma T, Kurata T (2006) Indoor/outdoor pedestrian navigation with an embedded gps/rfid/self-contained sensor system. In: Advances in artificial reality and tele-existence. Springer, pp 1310–1321
- Lachapelle G (2007) Pedestrian navigation with high sensitivity GPS receivers and MEMS. *Pers Ubiquitous Comput* 11(6):481–488

- Lemelson H, King T, Effelsberg W (2008) A study on user acceptance of error visualization techniques. In: Proceedings of the 5th annual international conference on mobile and ubiquitous systems: computing, networking, and services. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Dublin, Ireland, pp 53–58
- Pospischil G, Umlauf M, Michlmayr E (2002) Designing lol@, a mobile tourist guide for umts. In: Human computer interaction with mobile devices. Springer, pp 140–154

**Part V**  
**Information Retrieval, Modelling and**  
**Analysis**

# Feature-Aware Surface Interpolation of Rooftops Using Low-Density Lidar Data for Photovoltaic Applications

René Buffat

**Abstract** Digital surface models (DSM) are used to estimate the solar irradiation on rooftops. Estimates are more accurate when the precise geometrical characteristics of roofs are well represented in the DSM. The existing DSM covering Switzerland has a low accuracy for buildings. It was derived from a low density Lidar dataset with an average point density of 0.5 points per square meter. In this paper, we present a method to interpolate a DSM from point cloud data focusing on geometric modelling of rooftops. The method uses a combination of robustly fitted planes to local point clouds and inverse distance weighting interpolation. It was applied to roughly 3 million buildings, and compared to a reference DSM from a high density point cloud, which revealed a significant reduction of error compared to the existing DSM.

**Keywords** Lidar · Digital surface model · Interpolation · RANSAC

## 1 Introduction

Electricity generation from photovoltaic (PV) systems is expected to be the renewable energy source with the highest expansion potential (Bundesamt für Energie BFE 2013, p. 21). Typically PV installations are mounted on roofs. Not every roof is suitable for PV installation. Before a decision is made to install PV panels, the potential energy need to be known to guarantee the economic feasibility. The PV potential on building roofs can be estimated with GIS using digital surface models (DSM) (Hofierka and Kaňuk 2009; Brito et al. 2012; Nguyen and Pearce 2012). These methods calculate the solar radiation for each cell of the DSM based on shading and solar radiation models for tilted surfaces. Slope and orientation of tilted surfaces are important as they influence the solar irradiation (Reindl et al. 1990). Thus the quality of the DSM influences the PV potential estimates.

---

R. Buffat (✉)

Institute of Cartography and Geoinformation, ETH Zürich,  
Stefano-Franscini-Platz 5, 8093 Zürich, Switzerland  
e-mail: rbuffat@ethz.ch

Currently one DSM covering Switzerland exists. It has a spatial resolution of 2 m. The documentation of the dataset specifies an elevation accuracy for buildings with a standard deviation of 1.5 m (Swiss Federal Office of Topography 2005). This renders the dataset less suitable for PV potential estimates as building geometries are only roughly modelled. The DSM was created using Lidar data and inverse distance weighting (IDW) based interpolation with a search radius of 2 m (Roberto 2015). It is derived from a Lidar dataset with a low average point density of 0.5 points per square meter (Swiss Federal Office of Topography 2005).

The interpolation method used does not distinguish between buildings and other areas. However, buildings are special since rooftops consist typically (but not exclusively) of planes. Thus an interpolation method aware of this property can improve the interpolation accuracy. Subsets of point clouds belonging to buildings can be determined directly from the point cloud (Ortner et al. 2007) or alternatively, if available, with the help of building footprints. Even low density point clouds provide enough evidence to fit planes to large and uniform roof segments. However, the geometry of small structures may not be accurately detectable due to not enough available data points. Thus the interpolation needs to adaptively decide if a plane model is appropriate for a geospatial feature or not.

The goal of this work is the creation of a DSM using low accuracy input data with an improved representation of buildings and a higher resolution and accuracy as the input cloud. The developed interpolation method detects uniform roof segments by robust plane fitting to local point clouds combined with a region growing roof segment detection algorithm. Inverse distance weighting is used when no planes are detected. The newly created DSM is validated against a reference DSM. A significant improvement compared to the existing DSM is achieved. The results demonstrate that infusion of a-priori knowledge about geospatial features can significantly improve the quality of modelled geospatial data.

## 2 Related Work

The simplest method to generate a DSM from point cloud data is to use the maximum or mean value of the points within a cell of the DSM. With this method no value can be determined for cells without data points. Thus the resolution of the DSM need to be chosen based on the available point density.

In Nguyen and Pearce (2012), a triangulated irregular network using the Delaunay triangulation was used to derive a DSM to estimate the PV potential on rooftops. Different interpolation methods, including nearest neighbour, inverse distance weighting, triangulation with linear interpolation, minimum curvature, kriging and radial basis functions were applied to urban point cloud datasets with different densities from 25 to 0.25 points per square meter in Gonçalves (2006). The estimation error increases for all methods with datasets having a point density below 4 points per square meter. Influence of noise in the point cloud was not studied. No previous

work was found in the literature to specifically improve the interpolation of building roofs in DSMs from low density point cloud data.

A related problem is the construction of geometric 3D building models from point clouds. DSMs can then be derived from the 3D building models. A review of recent work is given in Haala and Kada (2010), Musialski et al. (2013). In general 3D building modelling methods focus on high density point clouds. While no one expects anything better to result from low density data other than rough and generalized roof shapes, the expectations are increasing along with the quality of the input data (Haala and Kada 2010, p. 571).

In Vosselman et al. (2001) 3D building models were derived with point clouds having 5–6, respectively 1.25–1.5 points per square meter. The methods reconstructs 3D building models from detected planar roof segments and segmented ground planes. Planar segments are detected by a 3D Hough transformation. The 3D Hough transformation detects points within the same plane. However it is stated, that the Hough transformation may find spurious planes in the case of many roof segments (Vosselman et al. 2001). Planes are fitted to points belonging to the same planar segment using least-squares. The results presented in the paper show that with the low point density dataset not fully recognized, small structures are completely ignored.

According to Haala and Kada (2010) fitting parametric building shapes into point clouds is suitable for low density point clouds. The method was applied to a point cloud with approximately 1 point per square meter in Brenner and Haala (1998). The method requires manual interaction when the 3D model cannot be automatically detected. Structures not present in the building shapes cannot be modelled.

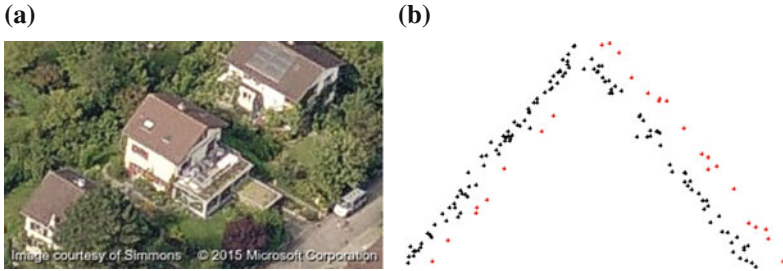
Random sample consensus (RANSAC) (Fischler and Bolles 1981) is an established algorithm in computer vision to fit a mathematical model to noisy data and is implemented in various libraries such as Rusu and Cousins (2011). In Schnabel et al. (2007), Möser et al. (2009), RANSAC is applied to detect features in the point cloud. The detected features are then used to create 3D models of buildings.

No previous research focused on the interpolation of rooftops to create DSMs from low density point clouds. The focus of the research is on the creation of 3D building models. A 3D model requires that every segment of a building needs to be detected and geometrically modelled. This requires high density point clouds. In our approach we acknowledge that low point density does not allow to model every roof segment geometrically. This is not required when creating DSMs because if it is not possible to detect a feature other interpolation techniques like inverse distance weighting can be used. Similar to methods for 3D building modelling, features need to be detected in the point cloud. In previous work 3D Hough transformation and RANSAC are used to detect planar roof segments. We chose a RANSAC based approach due to the robustness of the algorithm against outlier (Fischler and Bolles 1981). However the low point density of our input point cloud made it necessary to develop a new region growing plane detection algorithm.

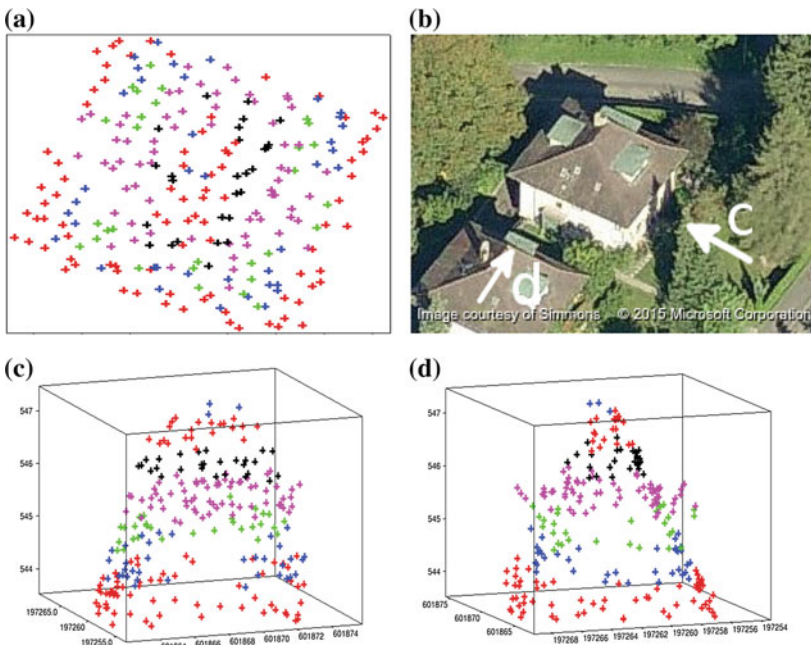


### 3 Data

The Federal Office of Cartography swisstopo created a Lidar dataset covering Switzerland below 2000 m elevation with flights in the period between 2001 and 2008. On average the dataset has a point density of 0.5 points per square meter (Swiss Federal Office of Topography 2005). The elevation accuracy of the laser used is 30 cm (Swiss Federal Office of Topography 2005). Our analysis of the data showed



**Fig. 1** Example of low accuracy input Lidar data. **a** 45° Aerial photo of building. **b** Front view of building roof point cloud, *red points* seem to be shifted compared to *black points*



**Fig. 2** Example of a building were iteratively application of RANSAC is likely to detect *horizontal planes* instead of roof planes due to low point density and noise. Orientation of (a) as in (b). Orientation of (c) and (d) as indicated in (b). **a** Point cloud (top). **b** 45° aerial photo. **c** Point cloud (side). **d** Point cloud (other side)

that the dataset contains a high ratio of noise. Examples of the quality of the point clouds are shown in Figs. 1b and 2c, d. Figure 1 shows a case where a subset of the points seem to have a constant offset compared to the rest of the points. In Fig. 2d the east and west facing roof surfaces of the building are clearly visible in the point cloud while the south and north facing roof surfaces of the same building in Fig. 2c are much more noisy.

Building footprints are available from different datasets, including cadastral survey, the cartographic SwissTLM dataset (Schmassmann and Bovier 2010) and Openstreetmap (Fan et al. 2014). We created a unified dataset for Switzerland with 88.91 % of the building footprints from the cadastral survey, 6.86 % from Openstreetmap and 4.22 % from SwissTLM.

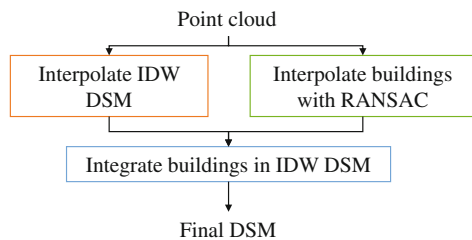
In 2014 a DSM was created for the canton of Zurich using Lidar data of the same year. The point cloud used has an average point density of 8 points per square meter (Amt für Raumentwicklung 2014). Both the DSM as well as the point cloud are publicly available. This dataset is used as reference DSM for the validation of our method.

### 4 Method

An overview of the implemented method is shown in Fig. 3. The focus of this work is to create a DSM with an improved representation of building roofs for PV applications. Areas outside of buildings are less important. Thus an initial DSM is created using IDW after outlier removal. Buildings are then interpolated separately by noise tolerant fitting of planes. This is done individually for each building. The interpolated buildings are then integrated in the initial DSM. The focus of this section is the interpolation of buildings. More details about the implementation of the whole process can be found in Sect. 6.

The assumption of the interpolation method is the presence of planes in the point cloud. Thus planes need to be detected in the point cloud. RANSAC is a suitable algorithm for this purpose due to its robustness to outliers (Fischler and Bolles 1981). With RANSAC a subset of the input data is randomly selected from the data and used to fit a model. Typically the least possible number of points of a model is selected randomly. For a plane model the minimal number of points is three. All points within

Fig. 3 Process to create DSM



a user given error tolerance are then counted. Typically the error tolerance is specified as the maximal allowed distance between a point and the model. Points within the error tolerance are called “inliner”. This procedure is repeated for a fixed number of times. The configuration with the highest number of inliners is returned. Alternatively a fixed number or ratio of inliners can be used as break condition. Appropriate break conditions need to be selected depending on the available data and models (Fischler and Bolles 1981).

In Schnabel et al. (2007), Möser et al. (2009), RANSAC is used to iteratively detect features. If a feature is detected, all inliner data points are removed from the point cloud. This is repeated iteratively. Noise present in the input data prevents the selection of a small error tolerance. The low point density as well as the shape of building roofs can lead to situations where a high number of points are at the base of a building roof. An example building is shown in Fig. 2. For such cases RANSAC selects with high probability a flat plane at the base of the roof. Iteratively applying RANSAC will lead then to a stack of flat planes. An interpolated building roof will then have a step pyramid like shape.

We eliminated this problem by applying RANSAC only to local point clouds. Pseudocode of the algorithm is shown in Algorithm 1. The algorithm can be divided into three phases. In the first phase, the grid of the resulting DSM is overlaid over the point cloud. For each grid cell center, a local point cloud is created. The point density of a building point cloud need not be uniform. Different buildings can have different point densities. Therefore the search radius of each point cloud is chosen adaptively such that at least 9 points are included.

In the second phase, for each point cloud the best plane is determined using RANSAC. A plane is found if more than 70 % of the points of a local point cloud are inliners. The initial error tolerance of the plane fitting is set to a maximal distance to the plane of 0.2 m. If no planes could be detected, the maximal distance is iteratively increased to 0.6 m. Detected planes are tested in the neighborhood with a region growing algorithm. In the final phase, for each cell, the plane is selected that is (a) valid for the cell and (b) valid for the most cells in total. This plane is then used to calculate the elevation value at the center of the cell. Roof segments are homogeneously interpolated since condition b asserts that only local dominant planes are used for interpolation.

Not all parts of a roof fit a plane concept. When no plane model fits the local point cloud, the elevation value is interpolated using inverse distance weighting. This allows a best effort representation of small features not representable by planes.

**Algorithm 1** Local Resample

---

```

1: procedure RESAMPLE
2:   for all cells do
3:     cloud[cell] = createLocalPointCloud(cell)
4:   for all cells do
5:     plane = findPlane(cloud[cell])
6:     if foundPlane then
7:       applyNeighborhood(plane, cell)
8:   for all cells do
9:     plane = bestPlane(planes[cell])
10:    if planeExists then
11:      elevation[cell] = calcElevation(plane, cell)
12:    else
13:      elevation[cell] = inverseDistanceWeighting(cell)
14:  return elevation

```

---

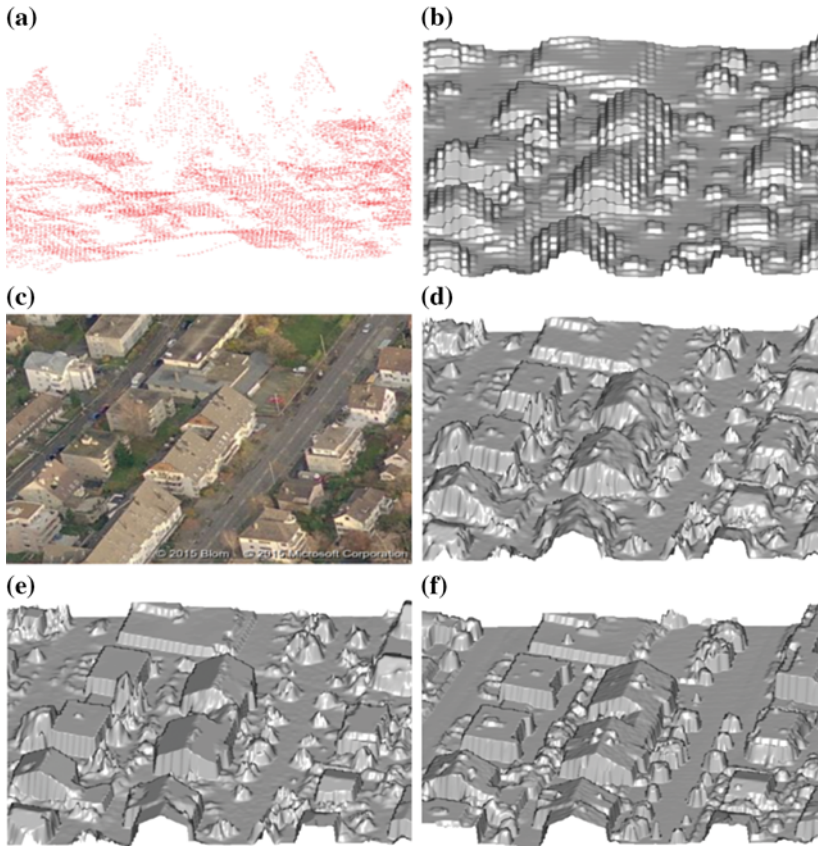
## 5 Results

Figure 4 shows a comparison of the existing DSM (Fig. 4b), the intermediate (IDW based) DSM (Fig. 4d), as well as the final DSM (Fig. 4e). The same extent of the reference DSM is shown in Fig. 4f.

Compared to the DSM interpolated using inverse distance weighting the new DSM contains homogenous roof segments. A stair like pattern can be observed in the roof surfaces of the reference DSM. Compared to the existing DSM, buildings are more detailed and are similar to the buildings of the reference DSM. Small structures like chimneys were smoothed out in the new DSM. Rough patches are sometimes present at the ridges. Most likely no planes could be detected for point clouds located at the ridges. Thus the interpolation for these cells is based on inverse distance weighting. In the top right corner, the reference DSM shows a new building not present in the aerial photo or the input cloud (Fig. 4a).

We used the city of Zurich shown in Fig. 5 to validate our method against the reference DSM. The region used for the validation covers an area of 210 km<sup>2</sup> and contains 73 440 building footprints.

For all cells inside a building footprint the absolute difference to the reference DSM were calculated both for the new DSM as well the existing DSM. Then for each building percentiles of the absolute differences were computed. Figure 6 shows the 30, 50 and 80 % percentiles of all buildings sorted in ascending order. In the existing DSM, 50 000 buildings have 50 % of the cells with an absolute distance of more than 1 m to the reference DSM. In comparison the new DSM has an absolute distance



**Fig. 4** Comparison of input data, interpolated DSMs as well as reference DSM. **a** Raw point cloud. **b** Existing DSM. **c** 45° aerial photo. **d** IDW. **e** New DSM. **f** Reference DSM

for this percentile of only 20 cm. With the new DSM around 15 000 buildings have 95 % of the cells with an absolute difference to the reference DSM of less than 50 cm compared to over 2 m with the existing DSM.

In Fig. 7 it can be seen that the absolute differences are reduced for all percentiles from the existing DSM (Fig. 7b) to the new DSM (Fig. 7a). For the new DSM, the absolute distances increase only slowly until the 70 % percentiles. Both datasets have a distinct increase of absolute distances for percentiles above 70 %.



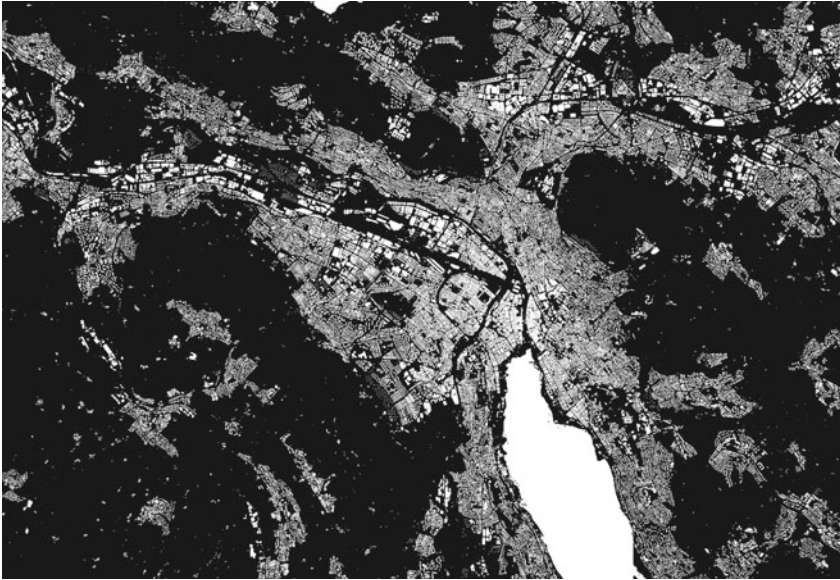


Fig. 5 Region (city of Zurich) used for validation. Buildings are colored white

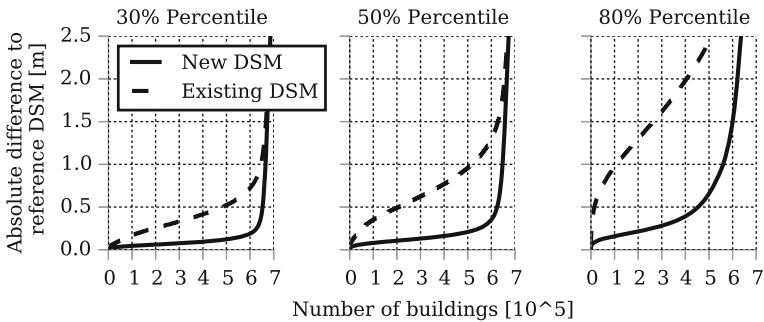
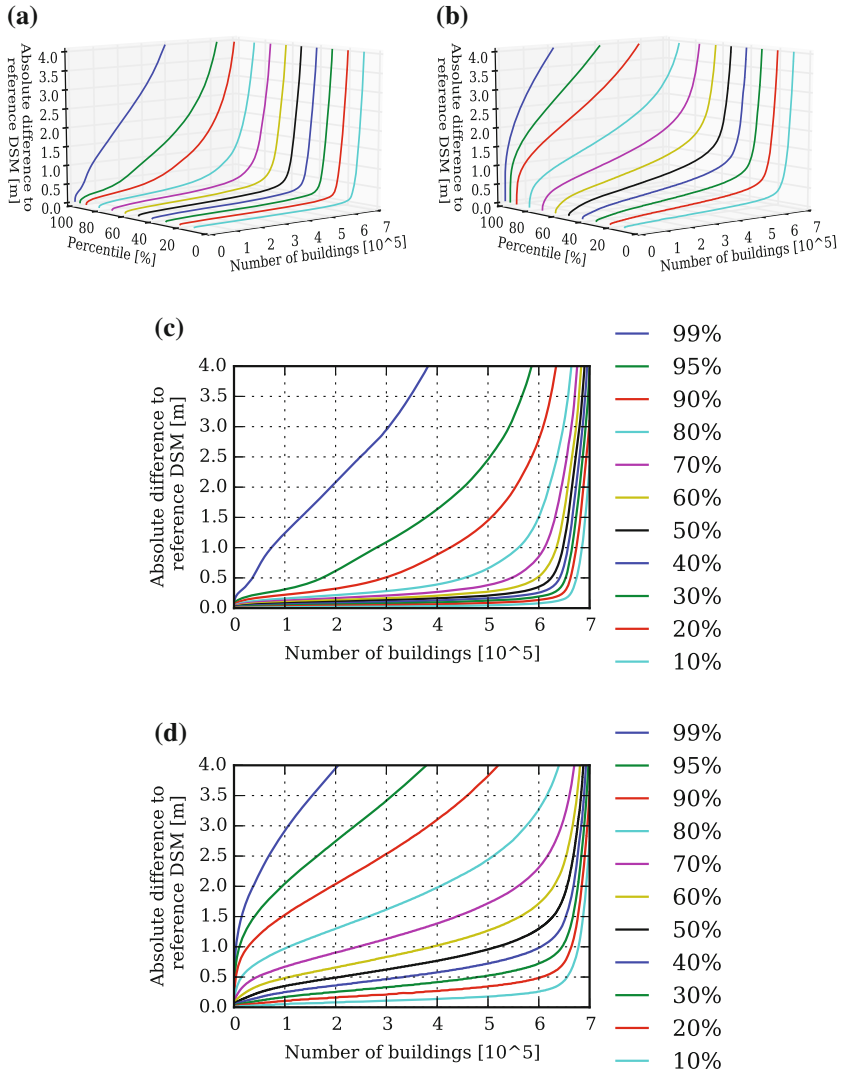


Fig. 6 Comparison of 30, 50 and 80 % percentiles of absolute differences between cells of new DSM and existing DSM to reference DSM. Absolute differences are sorted in ascending order

## 6 Implementation

The different steps of the workflow as shown in Fig. 3 were implemented as separate programs. Most of the programs are designed to operate only on one instance of e.g. one input point cloud file or one building dataset. They are designed that multiple programs operating on different instances can run in parallel without interference. Critical steps are implemented using the programming language C++. Python programs were responsible to execute the C++ programs with the proper parameters.



**Fig. 7** Percentiles of absolute distances of cells within building footprints of new DSM (a) (c) and existing DSM (b) (d) to reference DSM sorted in ascending order

Python can call external programs with the subprocess module. The Python programs use a queue in combination with as many worker processes as available CPUs to parallelize the processing.

The input data consist of the raw point cloud and the building footprint dataset. The point cloud is available in the form of nearly 3200 compressed text files with a size of roughly 300 GB. The dataset divides Switzerland in even spaced, rectangular regions. Each file contains all points of the point cloud within one such region. Each

line of the input files consists of one point with its x, y and z values. The points are in no particular order.

As a first step, the point cloud is divided into smaller datasets. For each building a dataset containing only the points within the extent of the building footprint is created. This allows later a simple parallelization of the building interpolation step using RANSAC. The building datasets are created by a C++ program that expects the path to an input point cloud file and to the building footprints. A spatial index is used to efficiently determine if a point belongs to a building extent or not. This is achieved by first creating an R-tree (Guttman 1984) of all building extents within the region of the input file. All points of the input file are then processed and in case they belong to a building written to the corresponding building data file. The C++ library Boost (BOOST 2015) was used for the spatial index and reading of compressed files. It should be noted that buildings can overlap multiple input point cloud files. This leads to multiple data files created for the same building. In a separate step after the creation of the datasets multiple datasets belonging to the same building are merged.

In a similar fashion the process to generate the IDW DSM is implemented with C++ and Python. A C++ program creates first a spatial index for all points of one raw point cloud input file. In the next step the points in the neighbourhood of each centroid of the cells of the output raster are determined with the spatial index. These points are used for the interpolation using inverse distance weighting. As our focus is on improving the DSM for buildings we did not use data points of neighbouring input files for the interpolation of border regions. The program stores the interpolated raster in ASCII format on the hard disk.

To resample the buildings, a program was created that expects a building point cloud dataset as input. The program uses the RANSAC implementation of PCL (Rusu and Cousins 2011) to detect planes. As output it writes the interpolated raster to the file system. This process was again executed in parallel using Python to create sub processes calling the C++ program with the appropriate input parameters.

Finally, a Python program reads an IDW raster file using the raster processing library rasterio. The program goes iteratively through all buildings touching the region covered by the IDW DSM raster file. All cells touching a building footprint are replaced with the interpolated cells of the building datasets. At the end, the final raster file is written to disk with rasterio.

## 7 Discussion

For the new as well as the existing DSM the absolute differences to the reference DSM increase exponentially after around 60 000 buildings. The reference DSM was created 12 years after the creation of the point cloud used to create the new DSM. The building footprints used for the interpolation and validation represent more closely the existing buildings at the time the reference DSM was created. We explain this increase with buildings that are either newly created, modified or replaced.



A possible reason for the increase of the absolute distances for large percentiles are small structures smoothed out with the interpolation. Additionally badly modelled sharp height discontinuities between parts of the buildings can contribute to errors of this magnitude.

The algorithm presented is simple to implement. Despite the low density input point cloud and the noise the presented method managed to achieve a median absolute difference of less than 50 cm for 85 % of buildings, compared to 1.47 m with the existing DSM. We did not integrate advanced concepts for roof reconstruction present in 3D building modeling literature. This might reduce errors even further.

It was possible to process roughly 3 million buildings with the algorithm within 3 weeks on a virtual machine with 32 cores (Intel Xeon CPU E5-2650 CPU). From the input point cloud for each building datasets were created containing only the points in proximity of the building. Each building was then interpolated individually. This allowed to parallelize most of the processing.

In recent years more and more regions are covered by Lidar data. Due to the improvement of Lidar technology new datasets tend to have significantly higher point densities compared to our input data. However point density is still an important factor of the production costs (Balsa-Barreiro and Lerma 2014). A large number of existing Lidar datasets do not have the density of todays technology. For example large parts of the US are covered with Lidar data having an average ground spacing between Lidar postings of 1–2 m (USGS 2011). Thus interpolation methods to enhance building roofs in DSMs have a huge potential to enhance existing datasets for PV potential estimation.

As seen in Fig. 4, despite the much larger resolution of the original point cloud, the roof surfaces of the reference DSM are not smooth and a stair like pattern is



**Fig. 8** Comparison of solar radiation estimates using the same method with the new DSM as well as reference DSM. **a** New DSM. **b** Reference DSM

present. This influences PV potential estimates. Figure 8 shows a comparison of the estimated yearly solar radiation using the reference DSM, respectively the newly created DSM. The same solar radiation estimation method is used for both DSMs. Higher level of solar radiation are indicated by brighter colours. In Fig. 8b it can be seen that the stair like pattern leads to uneven estimates of planar roof segments with the reference DSM. In comparison the solar potential of the same roof segments are more evenly estimated with the new DSM (Fig. 8a). We need yet to quantify the influence of this effect on solar potential estimates. In case the effect has a significant influence on the estimates an adapted version of our method can be used to create DSMs from high density point clouds with a smooth representation of building roofs.

## 8 Conclusion

The existing DSM of Switzerland does not model buildings accurately. The Lidar point cloud used to create the DSM has a low density and contains noise. Building footprints were used to distinguish between buildings and other areas. An interpolation method was developed specifically for roofs. The method fits planes robustly to local point clouds. As fallback inverse distance weighting is used. A comparison with a reference DSM revealed a significant improvement compared to the existing DSM due to the utilization of a-priori knowledge about geospatial features, specifically the presence of planes in building roofs.

## References

- Artuso R (2015) (Swiss Federal Office of Topography) private communication
- Balsa-Barreiro J, Lerma JL (2014) Empirical study of variation in lidar point density over different land covers. *Int J Remote Sens* 35(9):3372–3383
- BOOST (2015) Boost C++ Libraries. Version 1.58.0. <http://www.boost.org>
- Brenner C, Haala N (1998) Rapid acquisition of virtual reality city models from multiple data sources. *Int Arch Photogrammetry Remote Sens* 32:323–330
- Brito MC, Gomes N, Santos T, Tenedrio JA (2012) Photovoltaic potential in a lisbon suburb using lidar data. *Sol Energ* 86 (1):283–288. ISSN 0038-092X
- Bundesamt für Energie BFE. Energieperspektiven 2050, Zusammenfassung, Oct 2013
- Fan H, Zipf A, Fu Q, Neis P (2014) Quality assessment for building footprints data on open-streetmap. *Int J Geogr Inf Sci* 28(4):700–719. ISSN 1365-8816. doi:10.1080/13658816.2013.867495. <http://dx.doi.org/10.1080/13658816.2013.867495>
- Fischler MA, Bolles RC (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun ACM* 24(6):381–395. ISSN 0001-0782
- Gonçalves G (2006) Analysis of interpolation errors in urban digital surface models created from lidar data. In: Proceedings of the 7th international symposium on spatial accuracy assessment in natural resources and environment sciences

- Guttman A (1984) R-trees: a dynamic index structure for spatial searching. In: Proceedings of the 1984 ACM SIGMOD international conference on management of data, SIGMOD'84. ACM, New York, USA, pp 47–57. ISBN 0-89791-128-8. doi:[10.1145/602259.602266](https://doi.org/10.1145/602259.602266). <http://doi.acm.org/10.1145/602259.602266>
- Haala N, Kada M (2010) An update on automatic 3d building reconstruction. *ISPRS J Photogrammetry Remote Sens* 65(6):570–580. ISSN 0924-2716. ISPRS Centenary Celebration Issue
- Hofierka J, Káuk J (2009) Assessment of photovoltaic potential in urban areas using open-source solar radiation tools. *Renew Energ* 34(10):2206–2214. ISSN 0960-1481
- Kanton Zurich Amt fr Raumentwicklung, Baudirektion. Projekt luftaufnahmen42 technische spezifikationen auszug aus dem pflichtenheft, Nov 2014
- Möser S, Wahl R, Klein R (2009) Out-of-core topologically constrained simplification for city modeling from digital surface models. *Int Arch Photogrammetry Remote Sens Spat Inf Sci*, XXXVIII-5/W1. ISSN 1682-1777
- Musialski P, Wonka P, Aliaga DG, Wimmer M, Gool LV, Purgathofer W (2013) A survey of urban reconstruction. In: *Computer graphics forum*, vol 32. Wiley Online Library, pp 146–177
- Nguyen HT, Pearce JM (2012) Incorporating shading losses in solar photovoltaic potential assessment at the municipal scale. *Sol Energ* 86(5):1245–1260. ISSN 0038-092X
- Ortner M, Descombes X, Zerubia J (2007) Building outline extraction from digital elevation models using marked point processes. *Int J Comput Vis* 72(2):107–132. ISSN 0920-5691
- Reindl DT, Beckman WA, Duffie JA (1990) Evaluation of hotmpurly tilted surface radiation models. *Sol Energ* 45(1):9–17. ISSN 0038-092X
- Rusu RB, Cousins S (2011) 3d is here: point cloud library (pcl). In: 2011 IEEE international conference on robotics and automation (ICRA), pp 1–4
- Schmassmann E, Bovier R (2010) Topografisches landschaftsmodell tlm: swisstlm3d. *Mensuration Photogrammetrie Genie Rural (alternate title)* 108(9):407
- Schnabel R, Wahl R, Klein R (2007) Efficient ransac for point-cloud shape detection. In: *Computer graphics forum*, vol 26. Wiley Online Library, pp 214–226
- Swiss Federal Office of Topography (2005) Swisstopo DOM
- USGS (2011) Publicly available lidar point cloud data
- Vosselman G, Dijkman S et al (2001) 3d building model reconstruction from point clouds and ground plans. *Int Arch Photogrammetry Remote Sens Spat Inf Sci* 34 (3/W4):37–44

# Critical Situation Monitoring at Large Scale Events from Airborne Video Based Crowd Dynamics Analysis

Alexander Almer, Roland Perko, Helmut Schrom-Feiertag,  
Thomas Schnabel and Lucas Paletta

**Abstract** Comprehensive monitoring of movement behaviour and raising dynamics in crowds allow an early detection and prediction of critical situations that may arise at large-scale events. This work presents a video based airborne monitoring system enabling the automated analysis of crowd dynamics and to derive potentially critical situations. The results can be used to prevent critical situations by supporting security staff to control the crowd dynamics early enough. This approach enables preventing upraise of panic behaviour by automated early identification of hazard zones and offering a reliable basis for early intervention by security forces. This approach allows the surveillance and analysis of large scale monitored areas of interest and raising specific alarms at the management and control system in case of potentially critical situations. The integrated modules extend classical mission management by providing essential decision support possibilities for assessing the situation and managing security and emergency crews on site within short time frames.

---

A. Almer (✉) · R. Perko · T. Schnabel · L. Paletta

Joanneum Research Forschungsgesellschaft mbH, Steyrergasse 17, 8010 Graz, Austria  
e-mail: alexander.almer@joanneum.at

R. Perko  
e-mail: roland.perko@joanneum.at

T. Schnabel  
e-mail: thomas.schnabel@joanneum.at

L. Paletta  
e-mail: lucas.paletta@joanneum.at

H. Schrom-Feiertag  
AIT Austrian Institute of Technology GmbH, Giefinggasse 2, 1210 Vienna, Austria  
e-mail: helmut.schrom-feiertag@ait.ac.at

**Keywords** Airborne event monitoring • Automated situation awareness • Video based crowd dynamics analysis • Crowd management • Decision support

## 1 Introduction

Dramatic examples such as the disasters at Hillsborough football stadium (96 dead and 730 injured persons, 15 April 1989), the Love Parade in Duisburg (21 dead and 541 injured people, 24 July 2010) and recent past the dramatic crowd collapse during the annual Hajj pilgrimage in Mina, Mecca (2,070 pilgrims were suffocated or crushed, 24 September 2015), have shown that safety and security can only be ensured by controlling the movement dynamics of crowds. Critical situations can only be managed properly if they are detected at an early stage, allowing sufficient time to take appropriate countermeasures.

To date, only few studies and empirical data are available that identify dynamic crowd phenomena and derive relevant criteria documenting the behaviour of crowds in critical situations (Helbing et al. 2007). A key aim is therefore to identify the critical parameters and values, to classify situations and to investigate the reliability of criteria based on empirical data. The end users have defined three basic requirements for quickly assessing critical situations and taking appropriate countermeasures:

- Early detection of critical situations before they develop into dangerous situations.
- Clear identification of all intervention options, including information on the locations of all available security and emergency staff as well as available, blocked and open transport routes etc.
- Monitoring the effectiveness of the measures taken to enable appropriate crowd control; ensuring near real-time flow of information including all relevant persons and available information.

Terrestrial video systems are already being widely used for monitoring pedestrian flows. The combined deployment of terrestrial and airborne (airplane and helicopter based, as well as remotely piloted or unmanned aerial vehicle (UAV) based) video systems offers the opportunity to substantially increase the efficiency in detecting critical situations, making fast decisions in due time for coordinated interventions, controlling the measures taken and assessing their effectiveness. The key requirement for joint targeted actions in crisis situations is comprehensive and objective situation awareness on the basis of a situation map. For this purpose, all information required for decision-making and crowd control must be made available to the involved stakeholders. The “Donauinsel Festival 2013”—a highly crowded annual popular music event in the center of Vienna—provided a good opportunity to analyse the processes involved at large-scale events in cooperation with the Federal Police. A plane operator—Diamond Aircraft Industries (DAI)—supported the safety and security management of

the event with a DA 42 MPP<sup>1</sup> that has been equipped with a HDTV sensor as part of a related project—PUKIN (Periodical monitoring of critical infrastructure) which was a research project funded under the Austrian national security research program (see [www.kiras.at](http://www.kiras.at)) provided by the Federal Ministry of Transport, Innovation and Technology (BMVIT). A total of 2.9 million visitors were registered during the 3 days of the festival. The flood of data generated by the event and the time-critical decision-making processes involved clearly showed the necessity for an integrated and automated process to support the decision-makers and security and emergency crews on site and also offer realistic counts of people within defined areas.

## 2 System Overview and Workflow

Based on the described requirements and the current situation, the aim within the research project was to develop airborne monitoring methods based on video data to enable computer-based analysis of potentially critical movement patterns in crowds as well as an accurate count estimation of people within defined areas. Behaviour monitoring is used to prevent crisis situations by helping security personnel to influence group dynamics early enough so it does not end in critical situation for event attendees. The components developed within the project make it possible to monitor extensive areas and trigger specific alarms in the control centre whenever potentially hazardous situations are detected. All relevant information required for decision-making and crowd control can thus be made available to the stakeholders involved. The following components were developed on the basis of a high-performance multi-sensor video system:

- Video analysis and near real-time geo-processing with the aim of deriving geo-referenced data on crowd density and pedestrian movement.
- Behaviour analysis and simulations for assessing the hazard potential of crowds.
- Intelligent operations control (location based information management).
- Support of field operations through information exchange with mobile units.

Figure 1 gives a rough overview of the included modules and data flows.

The image quality provided by modern airborne and terrestrial video systems enables precise monitoring of crowds and automated analysis of key parameters based on HDTV images taken with appropriate sensor configurations. Figure 2 shows the processing steps required for deriving security relevant parameters from airborne crowd surveillance videos to acquire the current situation providing a better assessment of a current situation for the support of time-critical decision-making processes.

---

<sup>1</sup><http://www.diamond-sensing.com/index.php?id=da42mppguardian>.

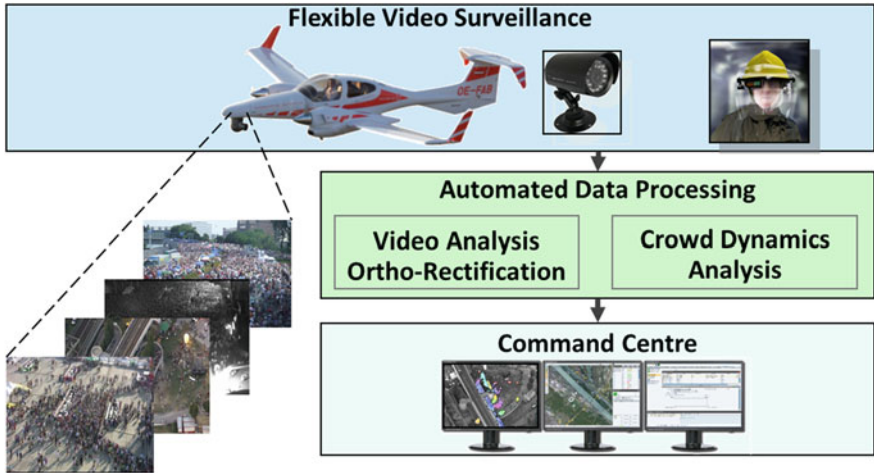


Fig. 1 System components and schematic workflow

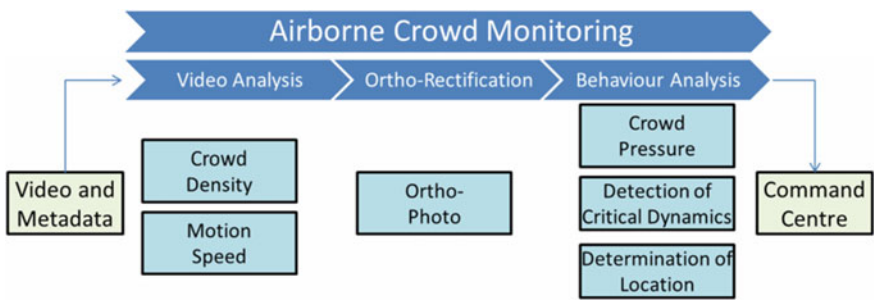


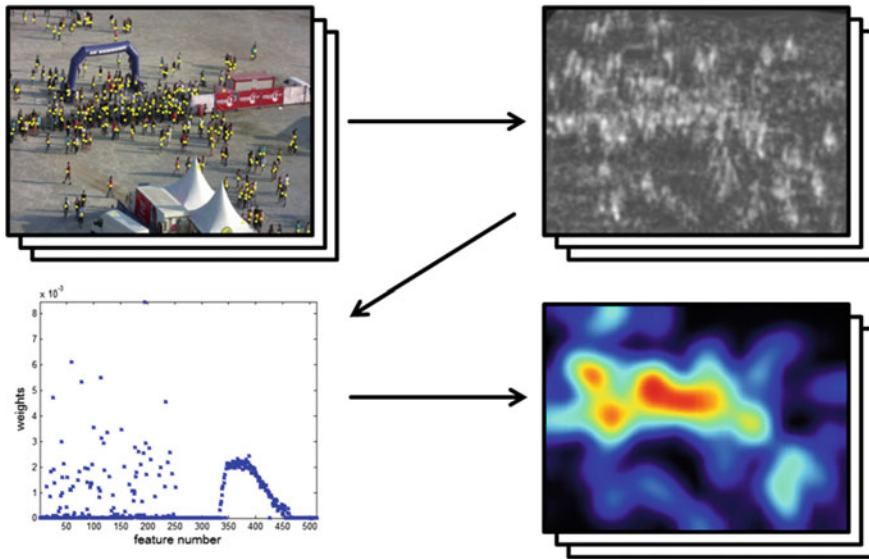
Fig. 2 Workflow for deriving safety and security relevant parameters

### 3 Airborne Crowd Monitoring

#### 3.1 Video Analysis

The main goal of the video analysis is to fully automatically extract the crowd density, the human count and the crowd motion from a given video stream for each frame. This information is later used to derive the human pressure, which is a crucial parameter to detect critical situations in human crowds. The proposed approach for density estimation and counting is sketched in Fig. 3 and discussed in the next section. The main idea is to extract image features which are then related to the human density by employing machine learning techniques.

The motion is estimated based on the variational description of optical flow in image geometry (Zach et al. 2007). To get a more robust estimate the flow is not



**Fig. 3** Proposed workflow for crowd density estimation: An image with annotated humans (yellow dots), discretized features (in this specific case the results of an object detector), the learned weights for each feature and the estimated human density function (estimated count equals 250) are shown (Perko et al. 2013)

gathered from two adjacent video frames but from frames with a temporal distance of 10 frames. In addition a given number of those flows are averaged to ensure smooth motion vectors (see Fig. 7). The resulting flow estimate is however in the same temporal sampling rate as the input video, i.e. 25 or 30 frames per second depending on the video camera specification.

### 3.2 Crowd Density Estimation

#### 3.2.1 Workflow

The presented methodology builds upon preceding work (Perko et al. 2013) which was limited to applications using one specific oblique viewing angle. In such cases pedestrians can be detected based on their silhouettes. However, when the view point changes to more nadir views this method will not produce useful results. Thus, an extension and two novel data sets for testing are presented.

The main idea is to calculate object detection scores from the given images and relate them to the human density by machine learning techniques. As object detector we propose a customized version of the histogram of oriented gradients (HoGs) detector (Dalal and Triggs 2005). The resulting scores are discretized such

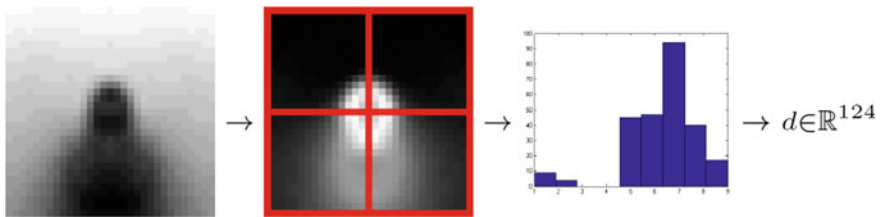


that the density estimation method is able to learn a weight for each of the scores. Thus, after learning the density function can be calculated by simple multiplications. In addition, the density estimate is a real density function, meaning that the integral over the density yields the object count (therefore, the integral over a subregion holds the number of objects in this particular region). Example images, the object detection scores and the density estimates are visualized in Fig. 5.

### 3.2.2 Object Detection

To enable a view invariant person detection we stick to detecting human heads in images, since those are visible in nadir views as well as in side views. Our proposed object detector is based on the construction of a useful descriptor for an image patch. Those descriptors are then used to train a support vector machine (SVM) that is later employed to calculate a confidence score for each location in the image. As basic descriptor we use the well-known HoG descriptor (Dalal and Triggs 2005) which describes an image patch by the occurrence of gradient orientations for a given number of local cells, thus encoding the silhouette of an object. We use the HoG variant reported in Felzenszwalb et al. (2010), since it yields slightly better object detection results while simultaneously having a lower dimensional descriptor compared to the original variant in Dalal and Triggs (2005). For each image patch this implementation results in a vector of dimension  $4 + 3*o$  with  $o$  being the number of orientations within the gradient histogram. After initial tests we use 9 orientations which results in a 31-dimensional vector for one HoG cell. The size of a HoG cell is set to  $15 \times 15$  pixels. As one cell would result in a weak descriptor we use  $2 \times 2$  HoG cells centered on our object and stack those 4 descriptors which finally yield a 124-dimensional feature vector. It can be considered as a rather low-dimensional description especially when compared to the original HoG-based pedestrian descriptor of Dalal and Triggs (2005) with 3780 dimensions. Figure 4 sketches the main concept of our descriptor. Shown is a patch holding a person (left), the gradient magnitude with the spatial arrangement of the 4 HoG cells (center) and one resulting gradient descriptor (right).

For learning we need positive examples extracted from manually labelled objects and negative examples (not holding a person). The positive descriptors are



**Fig. 4** Sketch of the proposed object descriptor. Four HoGs cells are stacked to gather a 124-dimensional feature vector

calculated for our manually labelled objects, where we are also incorporating a vertically flipped image to double the number of training data. For each image the same number of negative samples is gathered randomly from the image. To avoid that a negative sample also holds a person, a distance transform is calculated from the positive locations. Then a negative sample must have a distance larger than 1 % of the image diagonal (i.e. 18 pixel for an image of size  $1440 \times 1080$ ). The descriptors of positive and negative samples are employed to train an SVM, where the resulting model is later used for object detection.

### 3.2.3 Object Counting and Density Estimation

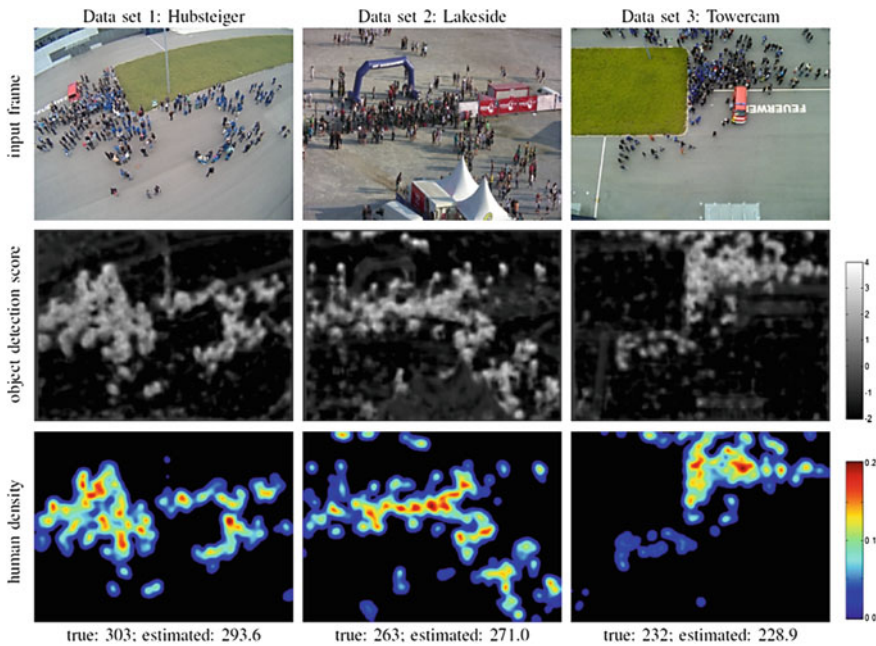
For counting objects and estimating their density we employ the method in Lempitsky and Zisserman (2010). This method takes densely extracted confidences from our detector and learns the density estimate via a regression to a ground truth density. Thus, each pixel has to be described by a feature vector of the following form  $f = (0, 0, \dots, 0, 1, 0, \dots, 0)$ , which is 1 at the dimension of the corresponding discretized feature and otherwise 0. For density learning our confidences have to be discretized, which is done by setting the minimal value to  $-2$  and the maximal value to  $+4$ . These bounds are used to scale the confidences to  $[0, 255] \in \mathbb{N}$ . Now, each of the possible 256 values defines a feature vector, as discussed above, which is 1 at the position of the confidence value. Therefore, it yields 256 individual features.

The training itself minimizes the regularized *Maximum Excess over SubArrays* (MESA) distance (cf. Lempitsky and Zisserman 2010) where we use two distinct approaches to solve the resulting linear or quadratic equation system, namely the  $L_1$  and the Tikhonov regularization (i.e.  $\min_x \|Ax - b\|$  or  $\min_x \|Ax - b\| + \|(x' \Gamma x) / 2\|$ ) with  $\|x\| \geq 0$  and Tikhonov matrix  $\Gamma$  being the identity matrix in our case). All details of this methodology are given in Lempitsky and Zisserman (2010). The result is a weight for each of the discretized features and the resulting human density is calculated by multiplying the according weight with the extracted feature value. Thus, for each pixel the density function is given and the sum over all pixels represents the number of objects in the image, i.e. our person count.

Therefore, in the testing phase the discretized features, i.e. our object detection scores, are extracted for each image and multiplied by the learned weight vector, directly resulting in the density estimation per pixel and corresponding person count. It should be noted that this approach introduces virtually no overhead over feature extraction (Lempitsky and Zisserman 2010).

### 3.2.4 Results of Object Counting and Density Estimation

**Test Data** Test Data for evaluation of the presented concept videos from three different scenarios were acquired in HD quality. Only individual video frames were used to simulate our envisioned airborne acquisition. Data from other tests showed



**Fig. 5** Exemplary results for the three different data sets. The *top row* shows a representative input video frame for each data set. *Middle row* gives the densely extracted object detection scores which are in the range from  $-2$  to  $+4$ . *Bottom row* visualizes the calculated human density functions scaled from  $0.0$  (blue) to  $0.2$  (red) in persons/pixel. The true number of persons in the images and the estimated count are stated below

that the images are analogous to images taken by an aerial platform. Exemplary images are shown in Fig. 5. This figure shows input images, the object detection score and the density estimate. The first scenario, referred as Hubsteiger, originates from a fire drill where we positioned an AXIS P3364 camera on a picker at approximate 25 m above ground. The images of this camera contain fish eye distortion and persons are observed under a slightly oblique look angle. The second one, referred as Lakeside, originates from a music festival in Styria, Austria. A Canon HV30 video camera was mounted on a tower (approximately 30 m above ground). Here the crowd is sensed under a flat look angle of about  $14^\circ$ , such that the whole silhouettes of persons are visible. The third one, referred as Towercam, originates from the same fire drill as Hubsteiger but here a NOKIA Lumia 710 mobile phone was mounted on the top of a building at about 40 m to capture the crowd in nadir direction. Finally, as we want to show the ability of generalization we constructed a combined data set that contains all images from data set 1–3. Even though the presented sequences are not taken from an airborne platform, the images have very similar properties as expected from UAVs or other sensing devices. Therefore, the presented workflow is supposed to yield appropriate results also on airborne imagery. We manually labelled 170 images to get the ground truth person

**Table 1** Manually labelled persons in the three data sets together with their statistics

ID	Number of images	Persons				
		Total	Min	Max	Mean	Std
DS1:Hubsteiger	45	11508	15	317	255.7	84.7
DS2:Lakeside	80	22300	249	319	278.8	13.4
DS3:Towercam	45	9468	144	263	210.4	33.4
ALL	170	43276	15	319	254.6	55.1

This information serves as ground truth for training and for testing

counts for training and later for the testing phase (overall more than 43000 persons were annotated, cf. Table 1). From the standard deviation of the people count in Table 1 it can be seen that DS1 Hubsteiger is the most difficult data set, as the number of people changes most dramatically.

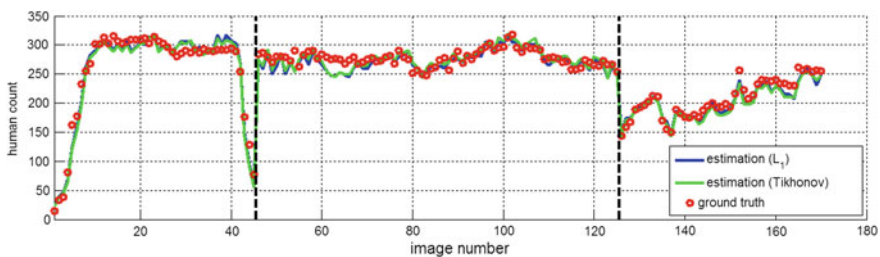
**Object Detection** To evaluate the object detection accuracy we extracted descriptors from positive and negative samples, for each data set and for the combined set (note, that the learning of the combined set involves huge amounts of data, i.e. more than 173000 124-dimensional vectors holding positive and negative samples). Then, we learned Support Vector Machine (SVM) models (Cortes and Vapnik 1995) and calculated the average accuracy by a 5-fold cross validation. For each run 4-folds were used to train the model and 1-fold served for testing. We also compared a linear SVM to a SVM with a radial basis function (RBF) kernel. For the RBF case we also varied two parameters  $\gamma \in [0.5, 1, 2]$  and  $c \in [2, 4, 8]$  (with  $\gamma$  being a parameter of the RBF kernel function for two samples  $x_i$  and  $x_j$  with  $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$  and  $c$  being a regularization parameter). While the linear SVM yields accuracies from 93 to 97 %, the RBF SVM performs better with 98.5–99.6 % (best results were achieved with  $\gamma = 2$  and  $c = 4$ ). Since the RBF SVM always achieves higher accuracies, this kernel was used for the density estimation later on. The detector for the combined data set gives nice results, with an accuracy close to 99 % using the RBF SVM. Therefore, one detector will be enough to process all given data sets. For training the final object detector we randomly selected 20 % of positive and negative samples from all data sets and trained a RBF SVM with the parameters stated above.

**Object Counting and Density Estimation** The accuracy for counting by density estimation of the training and testing process is listed in Table 2. We first used all available images to train the density estimation model. Then we took every second image, then every fourth etc., while the testing of the model was performed on the remaining images. It can be observed that the accuracy of training increases with a lower number of training samples. This makes sense, as the model adapts more and more to the specific samples but loses its ability for generalization (the well-known over-fitting problem). That is why the accuracy of testing is decreasing with a lower number of training samples. Thus, we can learn that about 20 images are sufficient for training the system. We can also observe that the regularization has a small effect on the testing results. Overall, an average error of human counts of about 12 can be

**Table 2** Accuracy of density learning and testing

Step	Training			Testing		
	#	$L_1$	Tikhonov	#	$L_1$	Tikhonov
1	170	10.3	10.7	0	–	–
2	85	10.0	10.9	85	11.3	11.0
4	43	8.7	9.6	127	11.2	11.3
8	22	7.1	7.4	148	12.2	12.0
16	11	6.2	5.3	159	13.1	13.2
32	6	4.7	3.8	164	13.2	12.9
64	3	2.0	2.4	167	14.9	13.2
128	2	1.2	0.5	168	37.2	23.1

Given are the average errors of the total human count over the training and test images, for two regularization options and different training and test set splits. A count error of 10 represents a relative error of 4 %

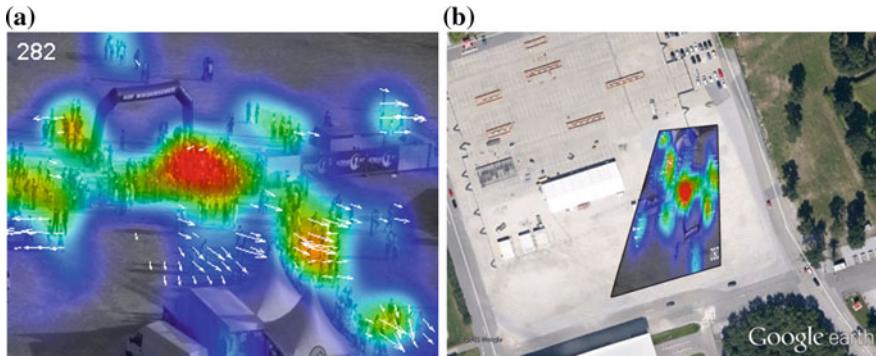


**Fig. 6** Person counting: Estimated person count using  $L_1$  regularization (blue) and Tikhonov regularization (green). The red dots indicate the manually measured ground truth. The vertical lines show the transition between the data sets 1, 2 and 3

reached, which correspond to a relative error below 5 %. Figure 5 visualizes some density estimates and Fig. 6 shows the results when using every 4th image for learning. Shown are the estimated human count for the two regularizations given in blue and green colour, together with the manually measured counts shown as red dots. The dashed black lines show the separation between the three data sets. Overall, it can be seen that the estimation is quite close to the ground truth data. Especially for data set 1 *Hubsteiger* our framework is also able to get good estimates when a lower number of people populates the scene (e.g. when people are entering the area in the first few images and when they leave from image number 40–45; cf. Fig. 6).

### 3.3 Ortho-Rectification

Geo-referencing, also called ortho-rectification, is a standard method in photogrammetry and in remote sensing (cf. e.g. Kraus and Harley 2007) which projects



**Fig. 7** Geo-referencing of a given image, the human density and motion estimate: **a** input image with superimposed color coded human density function, motion, and estimated number of individuals and **b** the geo-referenced version of **(a)** shown as Google Earth overlay. This conceptual figure is taken from our previous work (Perko et al. 2013)

the image onto the earth's surface with a given map projection. To be able to handle the distortions due to the topography a digital surface model (DSM) is used (global digital surface models like SRTM<sup>2</sup> or ASTER GDEM<sup>3</sup> are freely available). If the terrain is rather flat the DSM can be replaced by the knowledge of the mean terrain height.

All gathered information from the video analysis is geo-referenced and can therefore be visualized and processed in any geographic information system (GIS) system. Figure 7 shows a video frame superimposed with the estimated density and motion and the same information geo-referenced and overlaid in Google Earth. For the ortho-rectification of single images from a video stream, the runtime performance is an important factor. We used an indirect ortho-rectification approach where, based on the camera parameters and the position/orientation of the camera, the position on the ground is calculated. Thereby, we used the raw GPS and IMU (inertial measurement unit) data without post processing steps to enable a near real-time processing. Therefore, GPS and IMU need to deliver positioning estimates with an accuracy which should be sufficiently high for this kind of processing. There are many possibilities for getting the image data and corresponding meta-data from a video sensor, which mainly depends on the used system. For example high end video systems support KLV embedded meta-data in video transport streams like defined within NATO STANAG 4609.<sup>4</sup> For the extraction of single images as well as the synchronized meta-data, we used a commercial product of CarteNav

<sup>2</sup><http://srtm.csi.cgiar.org>.

<sup>3</sup><http://gdem.ersdac.jspacesystems.or.jp>.

<sup>4</sup>NATO Motion Imagery (MI) STANAG 4609 (Edition 3) [http://www.gwg.nga.mil/misb/docs/nato\\_docs/STANAG\\_4609\\_Ed3.pdf](http://www.gwg.nga.mil/misb/docs/nato_docs/STANAG_4609_Ed3.pdf)

Solutions,<sup>5</sup> which offers this functionality within the product. Depending on the gained format, the data is converted into the internal used meta-data format and used for further geo-data processing. Within this process, using a digital surface/elevation model allows the generation of an ortho-image with an acceptable accuracy for this purpose. The ortho-image itself is projected into a target coordinate system and projection which is appropriate for the area. Thereby we used a metric system as this is easier to use for the following calculation steps.

To keep it simple we define a common map frame for each of our test sites in WGS84 UTM 33 North projection (EPSG Code 32633) since our sites are located in Austria, Europe. Then for each image and for each column/line coordinate the according world coordinate is calculated which are used to rectify the density and motion information.

**Density** For projecting the density we use a forward transformation and project each density pixel into the common frame. If a pixel gets hit more than one time the values are summed up. This ensures that the sum of the density, i.e. the human count, stays the same in image and world coordinates. Since it happens that some pixels are hit more often than their neighbours due to rounding effects, the whole geo-referenced density is smoothed using a Gaussian kernel.

**Motion** Rectifying the motion is a bit tricky. In image geometry we cannot differentiate between object motion and camera motion. However, when transforming the reference image coordinate into the common frame using the reference transformation and the corresponding matched image coordinate with the search transformation, absolute world coordinates can be extracted. These two world coordinates define the real object motion independent of the camera movement.

### 3.4 *Crowd Behaviour Analysis*

The behaviour of crowds has been studied extensively for many years and researchers have conducted various experimental studies in order to understand human behaviour in different situations. Parameters such as crowd density, speed, flow and crowd pressure, see Helbing et al. (2007), and Steffen and Seyfried (2010) for definitions, are determined either manually (Seyfried et al. 2005) or by means of digital image processing (Johansson et al. 2008; Liu et al. 2009). Many models have been proposed and it was demonstrated that these models can describe the observed self-organizing behaviour like lane formation in crossing flows, intermittent flows at shared bottlenecks, arching at bottlenecks or the transition from laminar to stop-and-go flows (Cristiani et al. 2014). Due to the lack of viable data regarding critical crowd situations, insights of turbulent crowd flows has only been discovered in the last few years.

---

<sup>5</sup><http://www.cartonav.com>.



### 3.4.1 Turbulent Crowd Flows

Observations of dense crowds showed characteristic motion patterns of mass behaviour like stop-and-go waves or crowd turbulences (Helbing and Johansson 2009). Stop-and-go waves in a dense crowd are first indicators of dangerous overcrowding and can be used for the automatic detection of critical situations that may get out of control and entail disaster. Already Fruin (1993) reported high densities up to 7 persons per square meter, conditions where local crowd density is so high that the all the available space is filled full with human bodies. Under these conditions the individual control is lost and pedestrians become an involuntary part of the mass. In studies of pilgrim flows in Makkah from Helbing and Johansson (2009), stop-and-go waves have been observed even in areas without any obvious bottlenecks. They occur when the pedestrian density reaches a high level such that unobstructed pedestrian flow is inhibited. While the transition from laminar to stop-and-go flow was already well understood the insights of the subsequent transition into turbulent crowd dynamics were first revealed in Helbing et al. (2007). Therein variables were identified which are useful for an early warning of critical crowd conditions. Turbulent crowd flow occurs in situations of extremely high densities and is characterized by movements into all possible directions. It is caused by people who move involuntarily and induce sudden movements of other people nearby. As a consequence, people are pushed around and fall down. They are trampled down and, moreover, they turn into obstacles for others leading to more stumbling people.

### 3.4.2 “Pressure” in the Crowd

In contrast to purely density-based assessments, critical situations like shock waves and crowd turbulences are characterized by high variance of motion magnitudes under high densities. Helbing et al. (2007) introduced the crucial parameter to detect critical situations in human crowds as the so called local “pressure”  $P(\vec{r})$  defined by

$$P(\vec{r}) = \rho(\vec{r}) \text{Var}_{\vec{r}}(\vec{V}) \quad (1)$$

with the local pedestrian density  $\rho(\vec{r})$  times the local velocity variance  $\text{Var}_{\vec{r}}(\vec{V})$  of the velocities  $\vec{V}$  at the location  $\vec{r} = (x, y)$ . The local density is calculated by the average circular region of radius  $R$  at a given location in conjunction with a Gaussian distance-dependent weight function centred at the location, see Helbing et al. (2007) for details. For instance, the criticality in the crowd depends on the local conditions and therefore close measurements within a radius of  $R = 1$  m were used.



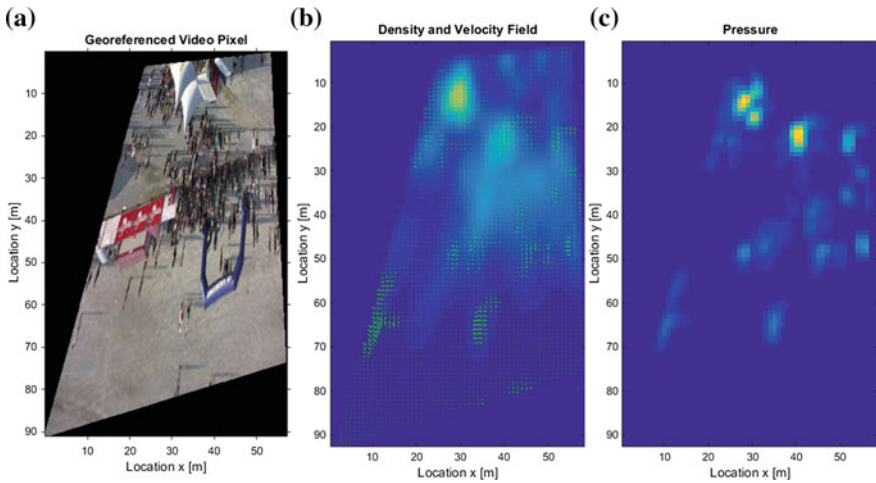
### 3.4.3 Method Employed

As described before, to evaluate critical conditions from a video stream, the crowd density and crowd motion has to be extracted for each frame through video analysis. The novelty of this approach was the use of airborne video instead of terrestrial video to provide an accurate crowd density and motion estimate for automated crowd behaviour analysis.

Accordingly, the results of the video analysis with pedestrian density and motion estimates for each video frame are transformed into a world-coordinate system using ortho-rectification and aggregated cell-by-cell in a two dimensional Cartesian grid with a cell size of  $1 \times 1$  m. For each cell  $c_{ij}$  the values of pedestrian density, velocity and pressure are stored in a two-dimensional array. In Fig. 8a, the video frame transformed into world-coordinates is shown. The contour plot in Fig. 8b illustrates the values for density and the green arrows represent the pedestrian velocity in the form of a vector field. In the shown example the grid has an overall dimension of  $58 \times 92$  m and the transformation in the world-coordinate system enables the use of physical units directly, e.g. number of persons per square meter and velocity in meters per second.

To calculate the local velocity variance, the velocities  $\vec{V}(c_{ij}, t)$  of the current and the eight directly adjacent cells (Moore neighbourhood) are used instead averaging over a circular region with a radius  $R$ , covering an area of  $9 \text{ m}^2$ . For each cell  $c_{ij}$  of the grid the velocity variance is

$$\text{Var}_{c_{ij}}(\vec{V}) = \langle [\vec{V}(c_{ij}, t) - \bar{U}(c_{ij})]^2 \rangle_t \quad (2)$$



**Fig. 8** Geo-referenced video frame (a), density with velocity field (b) and the computed “pressure” (c)

with  $\vec{U}(c_{ij})$  as mean velocity in the Moore neighbourhood. By adapting Eq. 1 to the grid structure, the pressure for each cell  $c_{ij}$  is calculated with

$$P(c_{ij}) = \rho(c_{ij}) \text{Var}_{c_{ij}}(\vec{V}) \quad (3)$$

whereas  $\rho(c_{ij})$  denotes the corresponding cell density. The contour plot in Fig. 8c displays the pressure in the cells (blue colour corresponds to low values, orange and yellow reflect higher values), where regions with higher values indicate higher crowd pressure.

Due to the lack of video data from critical situations, the hot spots shown in Fig. 8c reflect the highest pressure in the current video frame with an uncritical situation only. Here, the highest pressure values reaches  $0.002/s^2$  and were only a tenth of the threshold for dangerous situations as mentioned in Helbing et al. (2007) with  $0.02/s^2$ .

If a critical threshold for the crowd pressure will be exceeded an alert is sent to the command center, including the severity of the situation with the values normal, dense or critical depending on crowd density and pressure, the world-coordinates and a video frame of the video sequence that triggered the alarm.

## 4 Data Management

The results from the data processing and analysis modules are stored within a geo-oriented data archive. It allows geo-based as well as time based access to the available data and so provide all available data to command and control systems in the command room as well as to mobile units in the field. Thereby the command room provides access to all information required to support the safety and security management of large-scale events. The key aim is to generate a comprehensive current situation map giving a clear overview of the situation while also ensuring a high level of detail.

The main focus at the control room contains three key tasks: (1) Situation awareness, (2) decision support and (3) command & control. Being able to provide accurate information at the needed level and for the relevant target group is an important factor to support decision making processes. Therefore an application was developed to provide the information to the end user in an easy and intuitive way while displaying the position of sensors, infrastructure objects as well as mobile units combined with related sensor data. Also live video streams can be displayed by connecting the client to the streaming server on the sensor. The geo-referenced data is displayed as layers on a base map. These layers include current aerial images, the hazardous areas automatically calculated from the data as well as event-specific maps (e.g. stages, etc.). Access to the data archive allows the integration of historic data and so also allows the analysis how situations arose.

The creation of a common operational picture focused on providing fast and continuous situation awareness, role based data distribution (including commanders,

field commanders and field staff) and a concentration on a high usability for an effective support within crisis situation. Next to spreading information to mobile units, the data distribution also include interfaces to already existing command and control systems (e.g. see Ruatti Commander<sup>6</sup>) or local GIS systems.

## 5 System Demonstration

The number of high-performance video systems for civilian applications is growing rapidly in many European countries. The stakeholders involved and representatives of the air police (Federal Ministry of the Interior) have confirmed in personal communications that automated processing of video data and crowd behaviour analysis are essential in deploying these expensive multi-sensor video systems for purposes of safety and security management. The approach of multi-sensor data acquisition, geo-oriented data processing, automated behaviour analysis and target group specific representation in an integrated control room has clearly confirmed this potential. The system components were tested in two events. For organisational and legal reasons, only a limited amount of airborne video data was available for the development and evaluation of the system components. Additional video data were recorded from a tower during the Lakeside Festival in Austria in 2011 to simulate the airborne sensor configuration. In June 2012, project partners recorded extensive video data using a FLIR Star Safire 380-HD video system<sup>7</sup> during an overflight of the Donauinsel Festival.

## 6 Conclusions and Outlook

The presented system for decision support on large scale events yields promising results and helps to detect critical situations in near-real time. However, the limited availability of the video sensor system for legal and organisational reasons turned out to be a key problem in the project. This made it necessary to test different airborne video systems and to simulate airborne imaging situations using terrestrial video systems. Licence problems on the part of the supplier prevented access to the camera control software and the geo-sensor data (GPS, IMU). This reduced the amount of data available for development and testing in the fields of geo-processing, video and behaviour analysis and real-time data processing. The results therefore do not provide reliable information about the performance of the planned video-based crowd monitoring system but show results from the performed tests and different video systems. Even though the exemplary datasets hold only a

---

<sup>6</sup>Ruatti Systems GmbH, <http://www.ruatti-systems.de/en>.

<sup>7</sup>[www.flir.com/surveillance/display/?id=64505](http://www.flir.com/surveillance/display/?id=64505).

small number of people (around 300) and the crowd density is far from reaching panic densities, the presented approach is scalable provided that the image resolution in cm/pixel stays similar. In case of denser crowds the best image acquisition scenario will be a bird-eye view (close to DS3) as in this case human head are still visible and less occlusions occur in comparison to oblique look angles. In addition, a given larger area of interest could be observed by either using a higher-resolution camera with appropriate lens to assure the envisaged ground sampling distance or by employing the movement of the airborne vehicle to cover the area with time-delayed acquisitions. A challenge in the future is to get open interfaces to the sensory of such high end camera systems. These are not only of technical nature because, they principally existing but are not open accessible and not implemented on the different systems. The optimally suited imaging sensor obviously depends on the specific application. When the focus is on observing large scale events it would be beneficial to use a very high resolution camera that is still able to capture at least 3 frames per second, e.g. a Prosilica GT6600 with 29 MPixel and 4 fps. This frame rate is sufficient for motion estimation. When the area of interest to be observed is smaller any consumer grade full-HD video camera will fulfil the requirements. In a succeeding research project, this part as well as the correct synchronisation between individual video frames and the correct metadata from the GPS/IMU sensory will be focused on.

The results obtained from the video analysing component allow to conclude that the motion of crowds can in principle be well monitored and analysed at large scale events, and that the estimates of crowd densities which are crucial for an alerting system are sufficiently accurate to consider the application at future events. This first proof of concept requires future work to investigate the performance of the proposed methodology more deeply, i.e., using additional video sequences and to estimate the potential to adjust it to various illumination and imaging situations. Controlling the image gathering and tailoring it to the specific situation will additionally and substantially improve the results of the proposed image analysis. Oblique imaging also poses serious challenges, as persons may be partly hidden and geo-referencing in uneven terrain (bridges, installations etc.) requires high-precision surface models. These questions could not be dealt with in detail as the available resources were required for testing the different sensor systems. The final challenge will be to reduce analysis time by developing algorithms for near real-time analysis.

Based on the performed demonstrations, feedback from involved security staff as well as end users (police) have shown that the presented system will be a huge support for time critical decision making processes of the security staff in charge. Getting objective and reliable information about crowds and their dynamics allows better responses to critical situations and can so drastically enhance the safety of attending people.

**Acknowledgments** This work has been partially funded by the Ministry of Austria for Transport, Innovation and Technology within the Austrian Security Research Programme KIRAS: Project 845479: "MONITOR: Near real-time multisensor monitoring and short-term forecasts to support the safety management at mass events".

## References

- Cortes C, Vapnik V (1995) Support-vector network. *Mach Learn* 20:1–25
- Cristiani E, Piccoli B, Tosin A (2014) Multiscale modeling of pedestrian dynamics, vol 12. Springer
- Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Proceedings of the conference on computer vision pattern recognition, vol 2, pp 886–893
- Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D (2010) Object detection with discriminatively trained part based models. *IEEE PAMI* 32(9):1627–1645
- Fruin JJ (1993) The causes and prevention of crowd disasters. *Engineering for crowd safety*, pp 99–108
- Helbing D, Johansson A, Al-Abideen HZ (2007) The dynamics of crowd disasters: an empirical study. *Phys Rev E* 75:046109
- Helbing D, Johansson A (2009) Pedestrian, crowd and evacuation dynamics, encyclopedia of complexity and systems science, vol 16
- Johansson A, Helbing D, Al-Abideen HZ, Al-Bosta S (2008) From crowd dynamics to crowd safety: a video-based analysis, [arXiv:0810.4590](https://arxiv.org/abs/0810.4590)
- Kraus K, Harley IA (2007) Photogrammetry: geometry from images and laser scans, vol 1, 2nd edn.
- Lempitsky V, Zisserman A (2010) Learning to count objects in images. *Advances in neural information processing systems (NIPS)*, number 23, pp 1324–1332
- Liu X, Song W, Zhang J (2009) Extraction and quantitative analysis of microscopic evacuation characteristics based on digital image processing. *Physica A* 388(13):2717–2726
- Perko R, Schnabel T, Fritz G, Almer A, Paletta L (2013) Airborne based high performance crowd monitoring for security applications, scandinavian conference on image analysis (SCIA), vol 7944, pp 664–674. Springer LNCS
- Steffen B, Seyfried A (2010) Methods for measuring pedestrian density, flow, speed and direction with minimal scatter. *Physica A* 389(9):1902–1910
- Seyfried A, Steffen B, Klingsch W, Boltes M (2005) The fundamental diagram of pedestrian movement revisited. *J Stat Mech: Theory Exp*
- Zach C, Pock T, Bischof H (2007) A duality based approach for realtime TV-L1 optical flow. In: Symposium on pattern recognition (DAGM), pp 214–223

# Probabilistic Framework for Modelling the Evolution of Geomorphic Features in 10,000-Year Time Scale: The Eurajoki River Case

Jari Pohjola, Jari Turunen, Tarmo Lipping and Ari T.K. Ikonen

**Abstract** In this paper the long-term evolution of the catchment area of Eurajoki River, situated in Western Finland, is studied. The modelling area, nearly 1000 km<sup>2</sup> in size, is at present mostly covered by sea. Probabilistic digital elevation model and land uplift model form the basis for the future catchment area modelling. A land uplift model is required due to the ongoing post-glacial rebound especially in the western parts of Finland. The maximum rate of land uplift in Finland is 1 cm per year while in the modelling area the land uplift rate is about 6 mm per year. The digital elevation model and land uplift model have been calculated using Monte Carlo simulation where the uncertainties in the source data have been taken into account. The probabilistic nature of these models enables also the river catchment area and river network analyses probabilistically. The analyses are done for the next 10,000 years in 1000-year intervals and 100 realizations are estimated for each time point. The results show that the catchment area expands towards the west as the land rises. An alternative river branch flowing northwards from the main course will form with a significant probability. Also, a delta area with multiple river branches is expected to form at about 7000 years after present.

**Keywords** Catchment area · River network · Monte Carlo simulation · Land uplift · Probabilistic modelling

---

J. Pohjola (✉) · J. Turunen · T. Lipping  
Tampere University of Technology, Pori, Finland  
e-mail: jari.pohjola@tut.fi

J. Turunen  
e-mail: jari.j.turunen@tut.fi

T. Lipping  
e-mail: tarmo.lipping@tut.fi

A.T.K. Ikonen  
Environmental Research and Assessment EnviroCase, Ltd., Pori, Finland  
e-mail: ari.ikonen@envirocase.fi

# 1 Introduction

Modelling of the evolution of geomorphic features in an area where post-glacial land uplift is still in progress requires a land uplift model and a digital elevation model. Olkiluoto Island in Satakunta has been selected as a site of nuclear waste repositories, and therefore the development of the geomorphic landscape in the Olkiluoto modelling area (Fig. 1) is studied intensively. Presently, the post-glacial land uplift in Satakunta is approximately 6 mm/year (Poutanen et al. 2010).

Post-glacial land uplift refers to the rebound of the earth's surface after the melting of the ice formed during a glacial period. Holocene land uplift has been known to coastal residents of Northern Baltic Sea for centuries. Perhaps the earliest recorded history of land uplift in Fennoscandia can be dated back to 1491, when in the city of Östhammar in Sweden, the channel from the Baltic Sea became too shallow to maintain the traffic and the residents of Östhammar were resettled to the

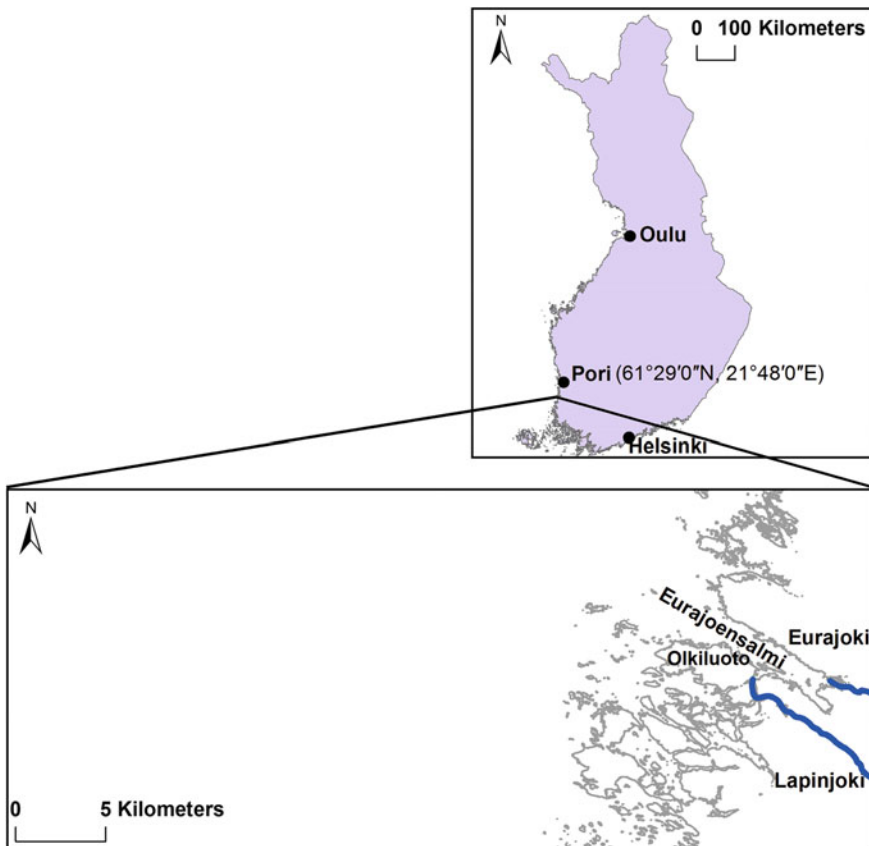


Fig. 1 The location and extent of the modelling area

nearby city of Öregrund which also got the city privileges of Östhammar (Ekman 1991; Harlén and Harlén 2003). Swedish professor of astronomy, Anders Celsius, measured in about 1743 the vertical sea water withdrawal to be 1.3 m in 100 years near the city of Gävle, but he had no information about the cause of the water withdrawal (Beckman 2001; Ekman 1991; Mörner 1979). Gerard de Geer showed, in late 1800s, the dependencies between ice age, ice recession, land uplift in Fennoscandia and land sinking in southern Sweden (Cato and Stevens 2011; Ekman 1991; Mörner 1979; Nordlund 2001). Nowadays land uplift can be monitored using precise GPS-station networks and gravimetric measurements from satellites (Lidberg et al. 2010; Müller et al. 2012; Poutanen et al. 2010; Timmen et al. 2004).

The causes of ice ages have been linked to solar radiation and its effects on the Earth's climate. Such climatic patterns were studied in the 1920s by a Serbian geophysicist Milutin Milanković. In his theory named as Milankovitch cycles, he concluded that there has been a sequence of ice ages caused by eccentricity, axial tilt and precession in the earth's orbit (Milanković 1998). The history of the Baltic Sea and its surroundings has been studied extensively. For example, reconstruction of the Baltic Sea basin from 130 ka BP to the present has been presented in Andrén et al. (2011) while the surrounding vegetational history and palaeoproductivity have been studied in Barnekow (2000), Kuhry and Turunen (2006) and Itkonen et al. (1999), respectively. In Berglund et al. (2009) the post-glacial land uplift effects near two nuclear power plant sites have been studied. The authors have assessed nuclear waste release scenarios up to 9000 AD including various statistical land uplift, hydrology and future inhabitant scenarios.

The land uplift rate in Finland has been studied using information on lake isolations from the sea. In Eronen et al. (2001) the isolations of several lakes were examined using sediment samples taken from the bottom of the lakes. The samples were dated with  $^{14}\text{C}$  radio-carbon dating and the age of the depth layer where saltwater algae changed into freshwater algae, was determined. The resulting shore-level displacement curves show that the land uplift rate has not been steady and there have also been local variations in the uplift. In this study the challenge is to extrapolate the uplift as truthfully as possible 10,000 years into the future. The empirical curve fitting method developed by Pässe (2001) is used.

The digital elevation model (DEM) underlying the study was created based on ground elevation and water depth measurement data from several sources. Several methods such as the inverse distance weighting method and various modifications of kriging are available for creating DEMs through interpolation. In a previous study (Pohjola et al. 2009, 2014) a set of interpolation methods were compared and it was found that the thin plate spline method (Donato and Belongie 2002) produced the narrowest confidence limits. Another advantage of the thin plate spline method is the preservation of natural shapes of the landscape while the methods relying on the weighted sum of data values tend to flatten hill tops and valley bottoms. Spatial correlation of the elevation values is taken into account by introducing histogram-indexing described in Sect. 2.4.



The objective of this work is to study the formation of future river network and catchment areas in the Olkiluoto modelling area (see Fig. 1) using a digital elevation model and a land uplift model taking into account the uncertainties in the estimation process using Monte Carlo simulation. The results will be used in future biosphere assessment and in modelling of the migration of radionuclides in the biosphere. The paper is outlined as follows: the digital elevation model and the land uplift simulation procedures are presented in Sect. 2, the river network modelling in Sect. 3, the simulation results in Sect. 4 and the conclusions in Sect. 5.

## 2 Inputs for River Network Modelling

In order to model the characteristics of the catchment areas and river network in the modelling area in the next 10,000 years, information about the current river system, the elevation model of the area, as well as the future land uplift process need to be considered. River network modelling, presented in Sect. 3, is then applied based on the models of future elevation. The whole modelling process is described in the flowchart presented in Fig. 2.

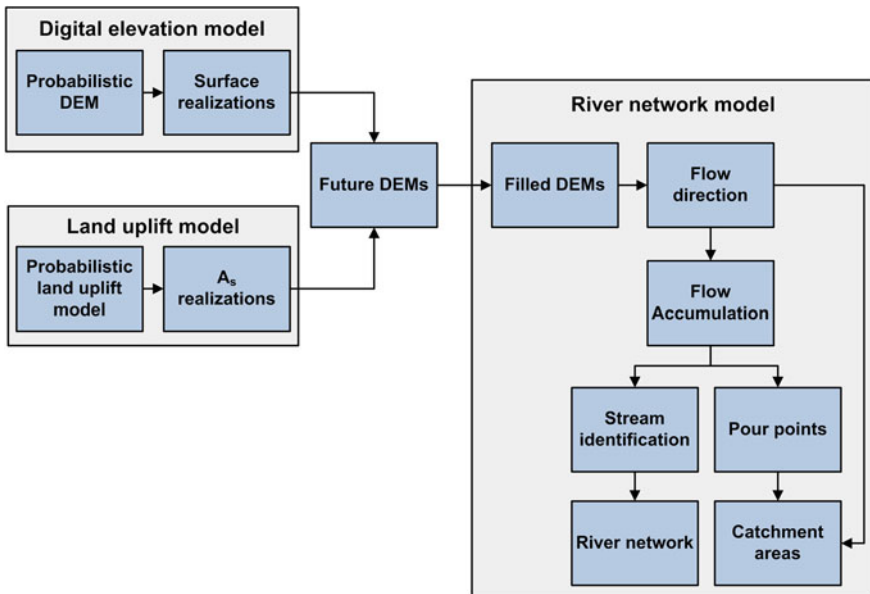


Fig. 2 Flowchart of the modelling process

## 2.1 *Eurajoki and Lapinjoki Rivers—Past and Present*

Eurajoki River is the main outflow channel from Lake Pyhäjärvi and Lake Köyliönjärvi. The measured annual mean discharge (1991–2000) of Eurajoki River is  $9.2 \text{ m}^3/\text{s}$ . Lapinjoki River is the outflow channel for nearly 40 small lakes, and its measured annual mean discharge (1991–2000) is  $3.6 \text{ m}^3/\text{s}$ . The total catchment area of Eurajoki River is  $1336 \text{ km}^2$  and that of Lapinjoki River  $462 \text{ km}^2$  (Ympäristö 2015).

Historically, it is known that both Eurajoki and Lapinjoki Rivers have discharged near to each other at the bottom of the present Eurajoensalmi Bay. There are remains of a small castle on an island of the time, founded to monitor the traffic of Eurajoki and Lapinjoki Rivers. The oldest remains of the castle are from late 1200s, but the  $^{14}\text{C}$  and dendrochronologically dated remains of the logs from newest layers are from 1348–1390 AD (Salminen 2009). Nowadays Lapinjoki River ends to a narrow strait called Karhukarinrauma that then connects to the Eurajoensalmi Bay. In Pohjola et al. (2014) it was found that at about 3000 AP the Lapinjoki River will flow again directly into the Eurajoensalmi Bay.

## 2.2 *Digital Elevation Model*

The digital elevation model (DEM) of the modelling area (Fig. 1) was created using measurement data from various sources. Measurements always involve error, which can be described by the error density function and  $p = 0.95$  confidence limits. Depending on the measurement method, confidence limits may be within a few centimetres (as in precise levelling, for example) or they can be as wide as 13 m (as in satellite based gravimetric data acquisition, for example). In many cases the error density function related to the particular measurement method was not known, so it was assumed that the error followed Gaussian distribution. For certain underwater sonar measurements the error density function followed exponential distribution. In the analysis of future landscape development, the DEM measurement errors were taken into account by using Monte Carlo simulation. Data values at the locations of the measurement points were drawn from the corresponding error distributions, and the thin plate spline interpolation method was used to obtain a probabilistic DEM at the regular grid of  $2.5 \times 2.5 \text{ m}$ . A detailed description of the methods used in the DEM creation process is given in Pohjola et al. (2014).

## 2.3 *Land Uplift Model*

When studying the effect of the post-glacial land uplift on a river catchment area in Western Finland, a land uplift model is required. In this study we used the empirical land uplift model presented by Pässe (2001) to extrapolate the past shore level displacement into future. This model was locally optimized specifically for the

study area using older and newly available lake isolation dating and archaeological data presented in Pohjola et al. (2014). This land uplift model is based mainly on curve fitting to the lake isolation datings. In the model, shore level ( $S$ ) can be expressed as the sum of two processes, the glacio-isostatic uplift ( $U$ ) and the eustatic sea-level rise ( $E$ ):

$$S = U + E. \quad (1)$$

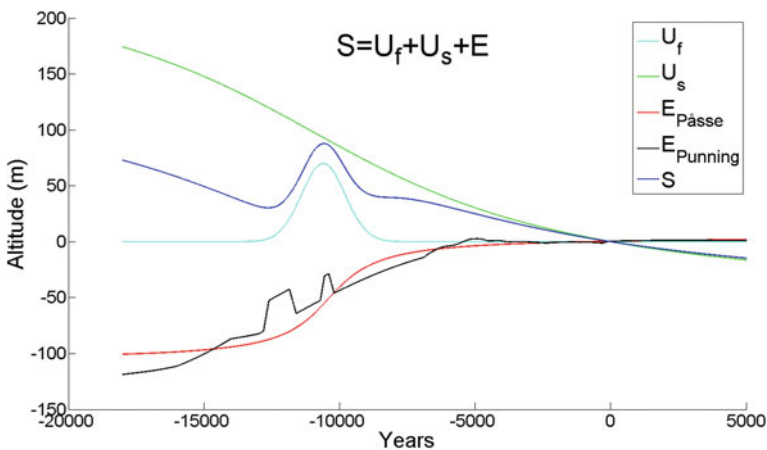
The components of the model are presented in Fig. 3. The glacio-isostatic uplift  $U$  has been divided into two components, the slow uplift  $U_s$  and the fast uplift  $U_f$ .

The eustatic sea-level rise ( $E$ ) is due to the melting of the ice; the water level has risen over 100 m from its lowest point after the last ice age. In Punning (1987) a more accurate sea-level curve than in Pässe (2001), taking into account the oscillations in the Baltic Sea area, was presented. This curve, modified to include the more accurate data on the Baltic Sea lake phases (Björck 2008) (see the eustatic curve of  $E_{Punning}$  in Fig. 3), was used as the eustatic component in the land uplift calculations underlying the river network and catchment area modelling presented in this paper, replacing the estimate used in Pässe (2001).

The slow uplift can be expressed by:

$$U_s = \frac{2}{\pi} A_s \left( \arctan\left(\frac{T_s}{B_s}\right) - \arctan\left(\frac{T_s - t}{B_s}\right) \right), \quad (2)$$

where  $A_s$  is the ‘download’ factor (in metres),  $B_s$  is the ‘inertia factor’ (in year<sup>-1</sup>),  $T_s$  is the time for the maximal uplift rate (in years) and  $t$  denotes the time variable (in years). The fast uplift can be expressed by:



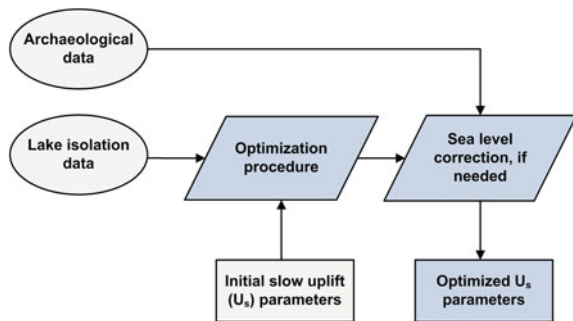
**Fig. 3** Shore level displacement, land uplift and eustatic sea level rise according to the land uplift model by Pässe (2001).  $E_{Punning}$  denotes the alternative eustatic model (Punning 1987) used in the catchment area and river network modelling in this study

$$U_f = A_f e^{-0.5 \left( \frac{t - T_f}{B_f} \right)^2}, \tag{3}$$

where  $A_f$  is the ‘total subsidence’ (in metres),  $B_f$  is the ‘inertia factor’ (in year<sup>-1</sup>),  $T_f$  is the time for the turning point from subsidence to uplift (in years) and  $t$  denotes the time variable (in years). A more detailed description of the model parameters is presented in Pässe (2001).

In this study the land uplift modelling part focuses on the optimization of the  $A_s$  and  $B_s$  parameters of the slow uplift  $U_s$ . The  $T_s$  parameter is taken from Geological Survey of Sweden (2016). The fast uplift  $U_f$  took place over 10,000 years ago, so its impact on the catchment area and river network modelling process is only marginal. Two kinds of input data were available for the land uplift model parameter optimization: one collected from lake basins, indicating the age of the sediment level where the environment changes from brackish water to fresh water, corresponding to lake isolation from the sea, and the other collected from archaeological sites of prehistoric human activity, indicating the time when the particular location definitely represented dry land. Both data sets involved the usage of the <sup>14</sup>C radiocarbon dating procedure, from which a probability density function can be derived for the calendar age of the sample (Bronk Ramsey 2009). The archaeological datings provide upper limit check for the water level in Monte Carlo simulations. In addition to the dating, another source of uncertainty is the elevation value of the data points. This uncertainty was taken into account by applying Gaussian distribution (standard deviation of 3 m was considered sufficient to count for the uncertainty due to, e.g., erosion related to lake tilting) to each elevation datum in the data sets. Monte Carlo simulation involving 1,000 realizations was then used to obtain the probabilistic estimates of the  $U_s$  parameter values of Pässe’s shoreline displacement model (Eq. 2). A flowchart of the land uplift model parameter estimation process is presented in Fig. 4. This Monte Carlo based parameter estimation of Pässe’s model can be thought as a curve fitting process where the minima of squared error of a set of random variables is sought. The random variables in this case are the lake isolation points and the archaeological data points with confidence limits. In Monte Carlo simulation the new values are taken randomly within their respective confidence limits and the most probable

**Fig. 4** Flowchart of the land uplift model parameter estimation process (redrawn and simplified from Pohjola et al. 2014)



value for the model parameters are the maxima of the histogram. More detailed analysis and discussion of the method are presented in Pohjola et al. (2014).

## ***2.4 Combining Probabilistic DEM and Land Uplift Models to Obtain Realizations of Future Land Surface***

Every point in the probabilistic DEM contained  $p = 0.95$  confidence limits from Monte Carlo simulation as described in Sect. 2.2. This information was utilized to create 100 realizations of the land surface that remained within the individual confidence limits at each point of the DEM grid. Especially, with each of the surface realizations it was important to maintain natural landscape forms such as hills and depressions, at their proper locations.

To ensure natural-looking surface realizations, the following histogram-indexing method was developed. The modelling area was divided into  $2 \text{ km} \times 2 \text{ km}$  blocks, each block containing  $800 \times 800$  grid points of the probabilistic DEM. A random location within each block was then drawn and random numbers between 1 ... 100 to represent the indexes into the histograms of the elevation values at these locations were generated. Thin plate spline interpolation was applied subsequently to the indexes to determine the index values of the elevation histograms at all the grid points of the probabilistic DEM. This ensures that the values of the surface realizations will change spatially smoothly within the individual DEM point error tolerance. As the interpolation process yields fractional numbers, the values had to be rounded to get valid indexes. The elevation of the particular surface realization was obtained by looking up the elevation values from the histograms of all the grid points according to the interpolated and rounded histogram indexes. To obtain 10 surface realizations, the procedure was repeated 100 times, that is, within each  $2 \text{ km} \times 2 \text{ km}$  block 100 different locations with corresponding random bin numbers were drawn as the starting points for surface generation. The same technique was used with the probabilistic land uplift model to obtain realizations for the  $A_s$  parameter and thus for the land uplift curve according to Pässe's model. The most probable value was chosen for the  $B_s$  parameter.

A GIS-based toolbox UNTAMO (Posiva 2013) was then used to combine the surface realizations and the probabilistic land uplift model including the realizations for the  $A_s$  parameter. The resulting future elevation models were calculated in 10,000 years' time span with a 1000 year interval.

## **3 River Network Modelling**

Modelling of the flow accumulation (catchment areas and the river network) presented in this paper is based on combination of probabilistic DEMs and land uplift models (Fig. 2). The resulting 100 realizations of the future land surfaces at each

time point in the time scale from 1000 AP to 10,000 AP with a 1000-year interval were used as an input to the modelling tool for river network and the catchment area modelling. In the first phase of the actual hydrological modelling, filling of the sinks in the future DEMs was performed. Filling of the sinks removes the areas with internal drainage which could cause problems later in the river network and catchment area modelling processes.

The flow direction model was calculated from the resulting filled DEM. The flow direction model implies the direction of outflow from each cell in the DEM to its steepest down slope neighbour. Flow direction is resolved using two parameters: the difference in the elevation values between the neighbouring cells and the distance between the centre points of the cells. There are eight possible directions for the outflow. This procedure follows the so-called eight-direction D8 algorithm presented in Jenson and Domingue (1988).

Flow accumulation model was derived from the flow direction model. The value of a cell in the flow accumulation model is obtained as the number of cells that flow into this particular cell in the DEM. The annual precipitation was set to 532 mm/year typical to the modelling area. From the flow accumulation model, a river network can be comprised by setting a threshold for throughput. In this case only the future Eurajoki-Lapijoki river channel was relevant so the threshold was set to correspond the mean annual discharge of 1 m<sup>3</sup>/s.

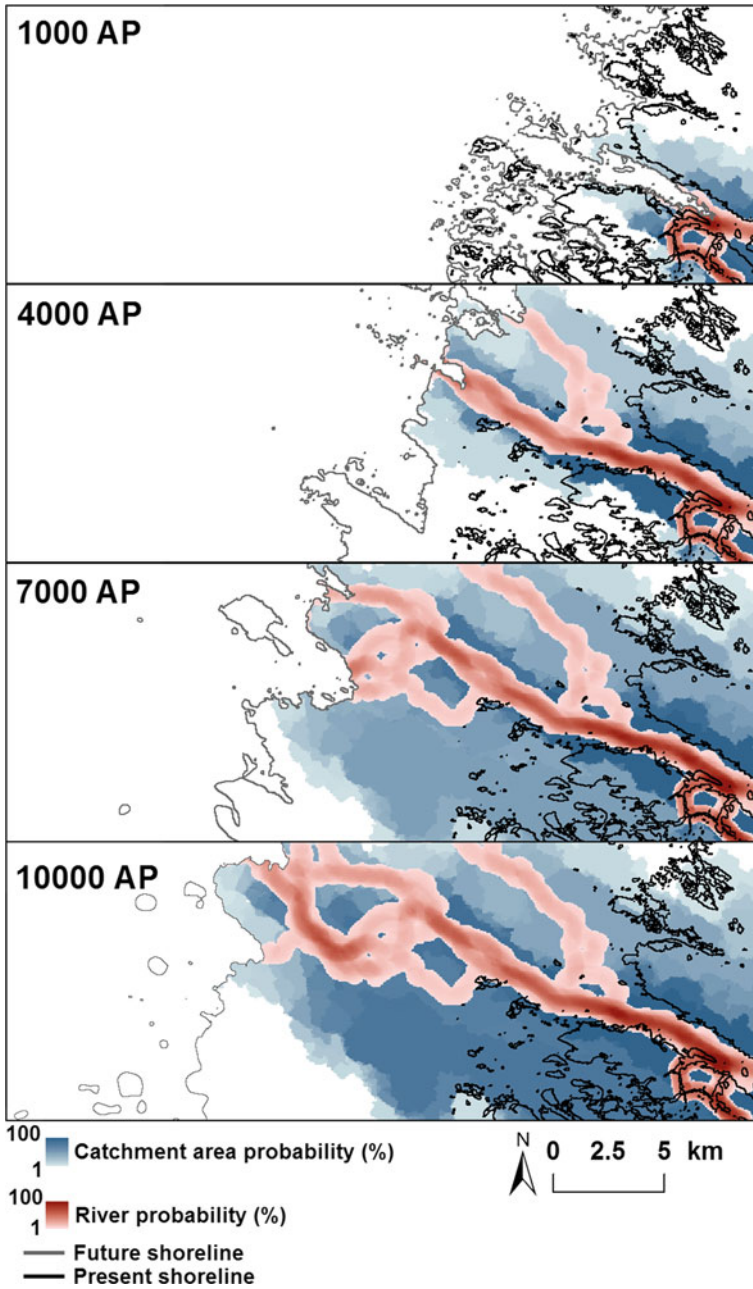
Pour point locations are determined based on the flow accumulation model. They were located at points of highest accumulated flow and therefore are the points that act as the outlets from the contributing areas. Delineation of the catchment area is determined based on the flow direction model and the pour point locations.

## 4 Results

The results include 100 realizations of catchment areas and river networks for the time scale from 1000 AP to 10,000 AP with an interval of 1000 years. The development of the size of the catchment area is presented in Table 1. Mean and standard deviation of the catchment area size have been derived from the results of the river network modelling for each time instant. The size of the catchment area

**Table 1** Catchment area size and standard deviation in the modelling area from 1000 AP to 10,000 AP

Years AP	1000	2000	3000	4000	5000
Area (mean, std) (m <sup>2</sup> )	3.8*10 <sup>7</sup> , 0.9*10 <sup>7</sup>	6.4*10 <sup>7</sup> , 0.7*10 <sup>7</sup>	8.2*10 <sup>7</sup> , 0.3*10 <sup>8</sup>	8.7*10 <sup>7</sup> , 0.2*10 <sup>8</sup>	1.1*10 <sup>8</sup> , 0.5*10 <sup>8</sup>
Years AP	6000	7000	8000	9000	10000
Area (mean, std) (m <sup>2</sup> )	1.4*10 <sup>8</sup> , 0.6*10 <sup>8</sup>	1.8*10 <sup>8</sup> , 0.8*10 <sup>8</sup>	2.1*10 <sup>8</sup> , 0.8*10 <sup>8</sup>	2.4*10 <sup>8</sup> , 0.7*10 <sup>8</sup>	2.4*10 <sup>8</sup> , 0.7*10 <sup>8</sup>



**Fig. 5** Catchment area delineation and river network probabilities in the modelling area at 1000 AP, 4000 AP, 7000 AP and 10,000 AP

increases as the land rises from the sea but the increase starts to slow down as the land uplift rate begins to approach zero between 9000 AP and 10,000 AP.

In Fig. 5 the evolution of the river network and the catchment area are presented as a map of the probability. The catchment area delineation and the river network location probability are shown for the time instances 1000 AP, 4000 AP, 7000 AP and 10,000 AP. Also, the present shoreline (black line) and the shoreline at the time point in question (grey line) are presented. The shoreline is estimated to shift westwards about 20 km by 10,000 AP. Probably the most interesting feature of the future development of the river network is the formation of an alternative branch at about 3000 AP flowing northwards from the main river channel. Also, the model implies that at around 7000...10,000 AP a delta area with multiple branches could form at the estuary of Eurajoki River similar to that seen at present about 40 km north of the study area at Kokemäenjoki River.

## 5 Discussion and Conclusions

Modelling the development of future geomorphic landscape is a challenging task especially in an area where land uplift is still an ongoing process. There are several unknown variables involved such as global eustasy, impact of human influence, and uncertainties in the underlying data.

Probably the most important source of uncertainty when considering the scenarios of future evolution of geomorphic features in the study area is related to the global sea level development. In this study, a mild increase in the sea level after the present day has been included in the eustatic model to counterbalance the land uplift phenomenon. In majority of the future sea level estimates (such as Meehl and Stocker 2007, for example) it is predicted that the sea level rises slightly in the future. The future sea level at the Olkiluoto site has been discussed in Posiva (2014). A less probable scenario, according to which the sea level rises nearly 4.5 m in the next 500 years and then starts gradually decreasing, has also been presented alongside with scenarios based on more conservative sea level rise. If this less probable scenario would be taken into notice, the geomorphic landscape evolution process would be slower as the sea level rise would counteract the effect of the land uplift.

Other sources of uncertainty are related to the various data used in the modelling process. The ground above the sea level is easy to model from lidar data; however, uncertainties and sparseness of the sonar data will have an effect on the accuracy of the sea bottom surface DEM. In order to obtain a regular DEM raster, interpolation techniques have to be applied. No single interpolation method can be considered superior as the goodness of the model depends on the form of the surface and there tends to be a compromise between interpolation accuracy and similarities of the resulting surface model to natural geomorphic shapes (Oksanen 2006). The latter is important from the point of view of hydrological modelling.



Another source of uncertainty is introduced by the land uplift model and its implementation. In principle, two approaches can be taken to model the land uplift process: (1) geodynamical modelling based on physical properties of the Earth's crust (e.g., Whitehouse 2009) or (2) empirical modelling based on fitting a mathematical function to existing data on the land uplift process in the past (e.g., Pässe 2001). The latter approach was considered more suitable for high resolution modelling of a relatively small study area. However, also the empirical approach relies on data from various sources such as archaeology or lake sediments. These data contain uncertainties related to measurement errors as well as dating uncertainties. For example, in some cases contradictions were observed in the data so that a past settlement remained well below the sea level as modelled according to the rest of the available data at the time point corresponding to the dating of its archaeological findings (see also Tiitinen 2011). It can be argued, however, that postglacial land uplift is a slow process and errors in the land uplift model of a relatively small area mainly affect the speed of the geomorphic evolution while the general pattern mainly remain unaffected.

The results from river network and catchment area analysis can be used in radionuclide transport modelling. Scenarios, where a possible radionuclide leak from a repository takes place, have to be examined in order to find out the risks for humans. In Avila et al. (2013) the long-term transport and accumulation of radionuclides was studied based on a landscape development model for the next 9000 years. These kinds of studies are essential in assessing the risk for people and biota living near water bodies.

Despite the uncertainties discussed, the model of future development of the landscape presented here, including river networks and catchment areas, can be considered as a valuable and reliable basis for safety studies of the nuclear waste disposal in the vicinity of the Olkiluoto Island. The model is currently used in the calculations of radionuclide dose conversion factors and in the creation of compartment models for radionuclide transport as a part of the safety analysis of the spent nuclear fuel repository.

## References

- Andrén T, Björck S, Andrén E, Conley D, Zillén L, Anjar J (2011) The development of the Baltic Sea basin during the Last 130 ka. In: Harff J, Björck S, Hoth P (ed) *The Baltic Sea basin*. Springer, Berlin, pp 75–97. doi:[10.1007/978-3-642-17220-5\\_4](https://doi.org/10.1007/978-3-642-17220-5_4)
- Avila R, Kautsky U, Ekström P-A, Åstrand P-G, Saetre P (2013) Model of the long-term transport and accumulation of radionuclides in future landscapes. *Ambio* 42:497–505. doi:[10.1007/s13280-013-0402-x](https://doi.org/10.1007/s13280-013-0402-x)
- Barnekow L (2000) Holocene regional and local vegetation history and lake-level changes in the Torneträsk area, Northern Sweden. *J Paleolimnol* 23:399–420. doi:[10.1023/A:1008171418429](https://doi.org/10.1023/A:1008171418429)
- Beckman O (2001) *Anders celsius*. Uppsala University. <http://www.astro.uu.se/history/celsius.pdf>. Accessed 17 Nov 2015

- Berglund S, Kautsky U, Lindborg T, Selroos J-O (2009) Integration of hydrological and ecological modelling for the assessment of a nuclear waste repository. *Hydrogeol J* 17:95–113. doi:[10.1007/s10040-008-0399-6](https://doi.org/10.1007/s10040-008-0399-6)
- Björck S (2008) The late quaternary development of the Baltic Sea basin. In: Bolle H-J, Menenti M, Rasool I (ed) *Assessment of climate change for the Baltic Sea basin*. Springer, Berlin, pp 398–407. doi:[10.1007/978-3-540-72786-6](https://doi.org/10.1007/978-3-540-72786-6)
- Bronk Ramsey C (2009) Bayesian analysis of radiocarbon dates. *Radiocarbon* 51(1):337–360
- Cato I, Stevens RL (2011) Gerard De Geer—a pioneer in quaternary geology in scandinavia. *Baltica* 5(1):1–22. doi:[10.5200/baltica.2012.25.01](https://doi.org/10.5200/baltica.2012.25.01)
- Donato G, Belongie S (2002) Approximate thin plate spline mappings. In: Heyden A, Sparr G, Nielsen M, Johansen P (eds) *Proceedings of the 7th European conference on computer vision-part III*. Springer, Berlin, pp 21–31
- Ekman N (1991) A concise history of postglacial land uplift research (from its beginning to 1950). *Terra Nova* 3(4):358–365. doi:[10.1111/j.1365-3121.1991.tb00163.x](https://doi.org/10.1111/j.1365-3121.1991.tb00163.x)
- Eronen M, Glückert G, Hatakka L, Van De Plassche O, Van Der Plicht J, Rantala P (2001) Rates of Holocene isostatic uplift and relative sea-level lowering of the Baltic Sea in SW Finland based on studies of isolation contacts. *Boreas* 30:17–30. doi:[10.1111/j.1502-3885.2001.tb00985.x](https://doi.org/10.1111/j.1502-3885.2001.tb00985.x)
- Geological Survey of Sweden (2016) Kartgenerator. <http://apps.sgu.se/kartgenerator>. Accessed 27 Jan 2016
- Harlén H, Harlén E (2003) *Sverige från A till Ö: geografisk-historisk upplagsbok*. Kommentus, Stockholm
- Itkonen A, Marttila V, Meriläinen JJ, Salonen V-P (1999) 8000-year history of palaeoproductivity in a large boreal lake. *J Paleolimnol* 21:271–294
- Jenson SK, Domingue JO (1988) Extracting topographic structure from digital elevation data for geographic system analysis. *Photogram Eng Remote Sens* 54(11):1593–1600
- Kuhry P, Turunen J (2006) The postglacial development of boreal and subarctic peatlands. In: Wieder RK, Vitt DH (eds) *Boreal peatland ecosystems*. Springer, Berlin, pp 25–46
- Lidberg M, Johansson JM, Scherneck H-G, Milne GA (2010) Recent results based on continuous GPS observations of the GIA process in Fennoscandia from BIFROST. *J Geodyn* 50(1):8–18. doi:[10.1016/j.jog.2009.11.010](https://doi.org/10.1016/j.jog.2009.11.010)
- Meehl G, Stocker T (2007) Global climate projections. In: Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt K, Tignor M, Miller H (eds) *Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change, 2007*. Cambridge University Press, Cambridge, pp 748–845
- Milanković M (1998) *Canon of insolation and the ice-age problem*. Textbook Publishing Company, Belgrade
- Müller J, Naeimi M, Gitlein O, Timmen L, Denker H (2012) A land uplift model in Fennoscandia combining GRACE and absolute gravimetry data. *Phys Chem Earth* 53:54–60. doi:[10.1016/j.pce.2010.12.006](https://doi.org/10.1016/j.pce.2010.12.006)
- Mörner N-A (1979) The Fennoscandian uplift and late cenozoic geodynamics: geological evidence. *GeoJournal* 3(3):287–318. doi:[10.1007/BF00177634](https://doi.org/10.1007/BF00177634)
- Nordlund C (2001) *Det upphöjda landet: vetenskapen, landhöjningsfrågan och kartläggningen av Sveriges förflutna, 1860–1930*. Dissertation, Umeå University
- Oksanen J (2006) *Digital elevation model error in terrain analysis*. Dissertation, University of Helsinki
- Pohjola J, Turunen J, Lipping T, Ikonen ATK (2009) Creation and error analysis of high resolution DEM based on source data sets of various accuracy. In: Lee J, Zlatanova S (eds) *3D geo-information sciences*. Springer, Berlin, pp 341–353
- Pohjola J, Turunen J, Lipping T, Ikonen ATK (2014) Landscape development modeling based on statistical framework. *Comput Geosci* 62:43–52. doi:[10.1016/j.cageo.2013.09.013](https://doi.org/10.1016/j.cageo.2013.09.013)
- Posiva (2013) *Safety case for the disposal of spent nuclear fuel at olkiluoto—terrain and ecosystems development modelling in the biosphere assessment BSA-2012*. Posiva Report 2012-29. Posiva Oy, Olkiluoto

- Posiva (2014) Safety case for the disposal of spent nuclear fuel at olkiluoto—data basis for the biosphere assessment BSA-2012. Posiva Report 2012-28. Posiva Oy, Olkiluoto
- Poutanen M, Nyberg S, Ahola J (2010) GPS measurements in Satakunta area. Posiva Working Report 2010-61. Posiva Oy, Olkiluoto
- Punning YM (1987) Holocene eustatic oscillations of the Baltic Sea level. *J Coastal Res* 3(4):505–513
- Påsse T (2001) An empirical model of glacio-isostatic movements and shore-level displacement in Fennoscandia. Report R-01-41. Swedish Nuclear Fuel and Waste Management Co., Stockholm
- Salminen T (2009) Kumo castle, Aborch and Vregdenborch—sources and past scholarship. In: Talvio T Suomen Museo 2008 (115. vuosikerta), Suomen Muinaismuistoyhdistys, Helsinki, pp 21–82
- Tiitinen T (2011) Liikettä ajassa ja paikassa—Lounais-Suomen muinaisrannat tarkastelussa. In: Uotila K (ed) *Avauksia Ala-Satakunnan esihistoriaan*. Eura Print Oy, Eura, pp 47–80
- Timmen L, Gitlein O, Müller J, Denker H, Mäkinen J, Bilker M, Wilmes H, Falk R, Reinhold A, Hoppe W, Pettersen BR, Omang OCD, Svendsen JGG, Øvstedal O, Scherneck H-G, Engen B, Engfeldt A, Strykowski G, Forsberg R (2004) Observing Fennoscandian geoid change for GRACE validation. In: *Proceedings of the joint CHAMP/GRACE science meeting*. GeoForschungsZentrum Potsdam, July 6–8, 2004
- Whitehouse P (2009) Glacial isostatic adjustment and sea-level change. Technical Report TR-09-11. Swedish Nuclear Fuel and Waste Management Co., Stockholm
- Ympäristö (2015) Eurajoki-Lapinjoki-ryhmä. <http://www.ymparisto.fi/fi-FI/Satavesi/Toiminta/Vesistoalueryhmat/EurajokiLapinjokiryhma>. Accessed 18 Nov 2015

# GeoPipes Using GeoMQTT

Stefan Herle and Jörg Blankenbach

**Abstract** The integration of common OpenGIS Web Services (OWS) into the Internet of Things and Service (IoTS) paradigm is a difficult task since they are based on HTTP with all its weak points. E.g. coupling small sensing devices or real-time processes with these services takes an enormous effort due to the different domain requirements. This paper focuses on extending existing geo web services with a push-based messaging mechanism to overcome their major drawbacks. We introduce the concept of GeoPipes and an exemplary implementation of them using the GeoMQTT protocol. The latter one is an extension of the MQTT protocol which is presented in this paper. Application examples show that with this concept a lot of technological issues can be solved easier.

**Keywords** Spatio-temporal publish/subscribe • IoT • Web services • Sensor web • Event processing

## 1 Introduction

In modern web applications location data are becoming more and more important. The Open Geospatial Consortium (OGC) provides specifications of different geo web services and encoding standards to include geospatial data into these web applications in a standardized way. However, these geo web services have some major weak-points when it comes to the integration of the new paradigm of the Internet of Things and Service (IoTS). In the IoTS various kinds of devices and applications, which have different requirements and constraints, have to be interconnected via the Internet. Furthermore, the predicted IoTS revolution in the years

---

S. Herle (✉) · J. Blankenbach  
Geodetic Institute and Chair for Computing in Civil Engineering and Geo Information Systems, RWTH Aachen University, Aachen, Germany  
e-mail: herle@gia.rwth-aachen.de

J. Blankenbach  
e-mail: blankenbach@gia.rwth-aachen.de

ahead will increase the amount of data and especially of spatial data which have to be integrated into modern web applications. Additionally, the velocity of newly generated data, by e.g. sensors, increases rapidly which means that storing and processing of this data must be adapted to the new requirements. For instance, in time-critical scenarios the processing of the data has to be accomplished in real-time to make decisions in time. The existing and implemented OGC services are not prepared for these new requirements yet.

Our approach, which we call *GeoPipes*, tries to couple the OGC services with a novel mechanism to add real-time and asynchronous functionalities. We implemented a push-based protocol for sharing geospatial data between different types of devices and applications based on the Message Queue and Telemetry Transport (MQTT) protocol. We call this extension *GeoMQTT*, a simple spatial notification and data streaming transfer mechanism between instances. This extension allows us to close some of the gaps and drawbacks in the standardized geo web services of the OGC but also to create some sophisticated real-time processes.

The paper is structured as follows. Section 2 gives an overview of related work in push-based geo services. Subsequently, the paper illustrates the concept of GeoPipes and specifies the requirements for the implementation. The MQTT protocol is introduced briefly before we describe the implementation details of GeoMQTT in the third section. We implemented some example applications which are shown in Sect. 4. In these applications, we enhance existing OGC services by the GeoPipes concept to overcome their drawbacks. Finally, a conclusion and outlook is given.

## 2 Related Work

Most OGC geo web service standards are based on Hyper Text Transport Protocol (HTTP) which uses the typical request-response pattern. Push-based protocols for sharing geospatial data between instances are, however, a much better approach in IoT architectures. Producers of data are able to push their data in real-time to consumers. Thereby, producers might be e.g. measurement units or processes which create new geospatial data.

There have been some efforts by the OGC to introduce a push-based approach for geospatial data. In Westerholt and Resch (2014) these approaches were compared for their general usability to enhance existing OGC services, especially the Web Processing Service (WPS). An early solution is the Sensor Alert Service (SAS) (Simonis 2007) and its successor the Sensor Event Service (SES) (Echterhoff and Everding 2008) created by the Sensor Web Enablement (SWE) initiative. In Westerholt and Resch (2014) it is concluded that especially the latter one is far beyond the basic purpose of a notification service and only suitable in the sensor domain. Secondly, the Web Notification Service (WNS) is a transport protocol-agnostic approach by the OGC to implement a more general approach (Simonis and Echterhoff 2007). However, according to Echterhoff and Everding

(2008) it is outdated and not under proactive development anymore. Last, the OGC Event Service is a generalization of the SAS and SES services beyond the sensor web domain (Echterhoff and Everding 2011). It is based on SOAP and WS-Notification with an arbitrary payload and the means of choice for Westerholt and Resch (2014). Recently, the OGC also focuses on pushed-based delivery of geospatial data and founded the PubSub Standards Working Group. The resulting Publish/Subscribe candidate is currently under review (OGC 2015).

However, in IoT environments with small devices, different services and various end-users these protocols do not match the requirements, mostly because they are still too high-level or heavyweight to run on any kind of hardware. For instance, processing of large messages such as XML payloads cannot be accomplished by devices with limited resources which reveal the previously mentioned services as impractical for a holistic solution.

### 3 GeoPipes with GeoMQTT

We introduce the concept of GeoPipes to facilitate the linkage of every type of device, application and software to each other. GeoPipes are channels between distributed instances and enable the sharing of geospatial data streams in a standardized manner. Producers can stream their geospatial data to consumers in a push-based way so that the latter one can process the data instantly. Producers and consumers can be of different types and hardware.

Initially, the main purpose of the concept of GeoPipes covers the sophisticated connection between data producers in the IoT such as sensor nodes in wireless geosensor networks (WGSN) and high level geo web services. The geospatial data is pushed through the pipe to a consumer which offers the received data as a web service to end users. But GeoPipes can also be used for messaging between different applications and visualization instances. That way, GeoPipes can also be utilized to implement a geospatial data enriched message-oriented middleware (MOM).

The requirements for a protocol, which implements the concepts of GeoPipes, are multifaceted since it is designated to connect different kinds of systems and has multiple objectives and issues to solve. The following chosen requirements are crucial for an implementation and are determined by different domains:

- **Messaging paradigm:** First, the messaging between instances should follow a push-based manner, meaning, if a message is published to the pipe it should be pushed to the consumer automatically. The publisher does not have to know the consumers of the pipe. Additionally, GeoPipes should be consumable by multiple consumers. All instances which are interested in the data published to a pipe receive the corresponding data. The publish/subscribe pattern is most suitable for these purposes (Eugster et al. 2003).

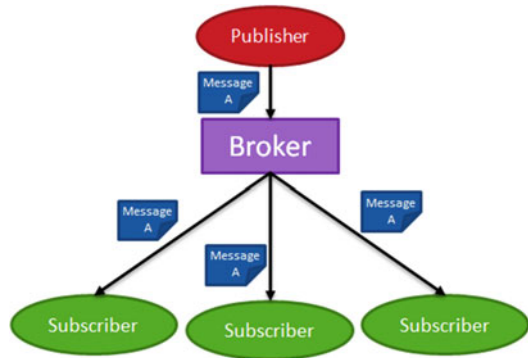
- **Open:** Since we do not want to create a totally new protocol, we need an open protocol to extend with our intended spatial functionalities.
- **Scalability:** GeoPipes should be consumed and produced by a large number of instances. This also means that the system should be able to deal with an increasing amount of instances.
- **Efficiency:** GeoPipes should facilitate the exchange of geospatial data in real-time between instances. Therefore, efficiency is a crucial requirement since we want to realize real-time applications with GeoPipes.
- **Interoperability:** Instances, which consume and produce data in GeoPipes, might be of any type. Especially resource-constrained devices should have the same rights as any other device. This directly leads to the next requirement.
- **Lightweight:** The protocol, which implements GeoPipes, should be lightweight, meaning, the overhead should be as small as possible. Devices with limited resources are not able to handle large protocol headers or data. So, a protocol with a minimal footprint is needed.
- **Security:** Authorization and authentication is also a big issue in IoTS environments. In the GeoPipes concept, this means in particular which instances are allowed to publish and which instances are allowed to consume from the pipe.
- **Reliable:** Message loss should be avoided if possible. Therefore, the protocol must support a reliability mechanism.

Our prototype implementation uses an extension of the lightweight IoTS protocol MQTT which basically meets all of the mentioned requirements with some trade-offs. We introduce new message types in the original protocol which are enriched with geospatial information. The new types of messages are described in Sect. 3.2. Thus, we call this extension GeoMQTT. Before describing the added features in detail, we illustrate the basic features of MQTT in the next section.

### 3.1 *Message Queue and Telemetry Transport (MQTT)*

The Message Queue and Telemetry Transport (MQTT) protocol is a very lightweight protocol and implements the so-called publish/subscribe interaction scheme. This scheme is an event-based communication model between *publishers* which produce certain information and *subscribers* which register to information. Subscribers are able to express their interest in a specific event or a pattern of events. If an incoming event, generated by a publisher, matches the registered subscribers' interests, they are notified subsequently. Producers publish information on a software bus (an event manager) and consumers subscribe to the desired information to receive from that bus (Eugster et al. 2003). The term *event* is typically used for publishing information on this bus whilst the term *notification* denotes the act of delivery to consumers. The central component in a publish/subscribe system is the *broker* which distributes the incoming data, issued by publishers, to all interested subscribers. The central broker in the architecture coordinates and manages all the

**Fig. 1** Publish/subscribe mechanism



subscriptions registered by the subscribers. This way, publisher and subscriber are connected by events and notifications but decoupled in time, space and synchronization. Figure 1 illustrates the general structure of pub/sub systems.

According to Eugster et al. (2003), three principal types of publish/subscribe systems exist: topic-based, type-based and content-based. MQTT uses a topic-based publish/subscribe scheme. Every message published by producers has a topic which is used in the broker to notify all interested subscribers of that topic. In MQTT the topic can be structured in different ways which is covered in Sect. 3.1.1.

Recently, MQTT is becoming more popular, since its extremely simple and lightweight nature designates MQTT to be one candidate for an IoT and machine-to-machine (M2M) standard. Its properties also support constrained devices as well as low-bandwidth, high-latency and unreliable networks (MQTT.org 2015). Thus, it fits our requirements for GeoPipes perfectly.

### 3.1.1 Addressing via Topics

An MQTT broker distributes the messages according to the interests of the subscribers and the topic names of the messages. A topic name is a string which can be almost arbitrarily chosen. Some restrictions can be found in the specification, e.g. it must be at least one character long etc. (OASIS 2014). Topic names in MQTT are structured by the so-called topic level separator, a forward slash, which divides the topic name into multiple “topic levels”. This way, a topic tree, which provides a hierarchical structure for the topic names, can be set up.

Furthermore, the specification distinguishes between *topic names* and *topic filters*. As mentioned before, topic names are used to assign every publish message to a specific topic whilst topic filters are used by subscribers to subscribe to one or a set of topics. Topic filters are similar in structure to topic names but can also contain wildcards between the topic separators to specify a bunch of possible topic names. Thereby, the number sign “#” is the so-called multi-level wildcard character that matches any number of levels within a topic. It represents the parent and any



number of child levels. For example, the topic filter “*sensor/dike/id5/node/#*” matches the following topic names:

- *sensor/dike/id\_5/node*
- *sensor/dike/id\_5/node/id\_3*
- *sensor/dike/id\_5/node/id\_3/temperature*

The single level wildcard “+” can be used multiple times in one topic filter and also in conjunction with the multi-level wildcard. It has to occupy one and only one entire level of the filter. For example, the topic filter “*sensor/dike/+/node/#*” matches the topic names:

- *sensor/dike/id\_1/node*
- *sensor/dike/id\_1235/node/id\_3/temperature*
- *sensor/dike/id\_9239/node/id\_2/humidity*

As mentioned, there are some restrictions for the topic name and the topic filter which are covered in the specification (OASIS 2014).

### 3.1.2 MQTT Features

The MQTT Version 3.1.1, which is an OASIS standard since 2014, offers some interesting core features supporting the topic-based pub/sub mechanism.

#### Quality of Service (QoS)

One of the most important features is the QoS level within MQTT which specifies the guarantees of delivering a message. Delivering a message in this sense means on the one hand publishing from a client to the broker and on the other hand forwarding from the broker to the subscribing clients. In the first case, every publish message is assigned a QoS level specified by the publisher. In the second case, the subscribed client has to set a QoS level during his subscription. Three levels are available: *at most one*, *at least one* and *exactly one*. According to the chosen level, different message types are sent between the involved components. The QoS mechanism is crucial especially in unreliable networks, since the delivery status of important messages can be monitored or in case messages are retransmitted.

#### Persistent or Clean Session Mechanism

In a persistent session the broker stores all information relevant for the client such as subscribed topics or undelivered messages, even if it is offline. As soon as the client reconnects, the undelivered messages are forwarded and the session is restored. The client is able to end the persistent session by setting a clean session flag.

#### Retained Messages

Retainable messages are marked with a flag in the publish message. The broker stores the last retained message and the corresponding QoS for the topic. Clients receive the retained message immediately after subscribing to this particular topic.

### Last Will and Testament (LWT)

The LWT feature notifies clients about an “ungracefully” disconnected client. The last will is basically a normal MQTT message and is specified by the client when it connects to the broker. Other clients can subscribe to this last will message by the topic name and the broker sends the last will, if it detects that the client is disconnected abruptly.

These features support the reliable exchange of messages. Thus, the reliability requirement is met. However, MQTT does not implement a security solution in its core specification but provides some guidance how to integrate authentication and authorization in the protocol. For scalability, some MQTT brokers implement a so-called *broker bridge* which can be used for scaling horizontally. For future developments we have to keep both features in mind.

### 3.1.3 MQTT for Sensor Networks

MQTT is based on TCP/IP but there is also an extension for connectionless communication protocols like UDP or ZigBee. This extension is especially useful in wireless sensor networks (WSN). Hence, it is called MQTT for Sensor Networks (MQTT-SN). MQTT-SN was designed in 2008 based on the following design principles (Hunkeler et al. 2008):

- As close as possible to MQTT
- Optimized for tiny battery-operated Sensor/Actuator devices
- Considerations of WSN constraints such as high link failure rate, low bandwidth and short message payload
- Network independent

The developed architecture for MQTT-SN introduces two more components in a MQTT system to bind MQTT-SN clients: the clients themselves and a gateway acting like a translator between the two protocols (see Fig. 2).

The gateway can be implemented using a *transparent* and *aggregating* approach. A transparent gateway establishes a new connection to the broker for every MQTT-SN client while an aggregating gateway has only one connection to the broker. Furthermore, MQTT-SN uses more message types than the basic protocol but the messages are similarly lightweight. More information about the supported features and implementation of MQTT-SN clients and broker can be found in the specification of the protocol (Stanford-Clark and Truong 2013).

## 3.2 GeoMQTT

Originally, the motivation for a geospatial extension for the MQTT protocol was to tag the publish messages with a timestamp and coordinates besides the topic name.

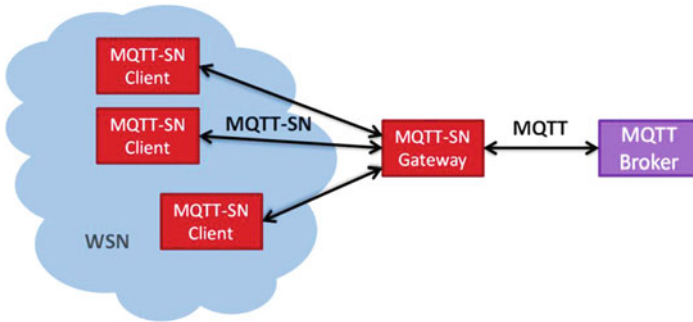


Fig. 2 MQTT-SN extension for sensor networks

Because the integration of this meta information into the payload of normal MQTT publish messages is quite an inadequate solution, we introduce a new publish message type in MQTT, the *GeoPublish* message. Furthermore, with this new meta information in a publish message—time and location, we can also introduce two new subscription filters. Using the geospatial extension, a client is able to use a *GeoSubscribe* message to narrow down interested events to temporal intervals and a specific spatial area in conjunction with the MQTT topic filter. The GeoMQTT broker forwards incoming publish messages to interested subscribers, if and only if all three filters are satisfied. This way, a lot of subscription options are introduced providing a wide range of expressing interests in events.

### 3.2.1 Temporal and Spatial Filters

We enhanced the subscription mechanism of MQTT with a temporal and a spatial filter. Spatial and temporal filters can be specified by the GeoMQTT clients in multiple ways. For instance, temporal filters can use the shape of a specific time span by setting start and end timestamps or repetitive periods by specifying a temporal pattern. The options for the filters are covered in the next sections in detail.

#### Temporal Filter

The newly introduced *GeoPublish* message contains a mandatory timestamp which can be set by the publishing client either in ISO8601 format or in Unix time with a 32 bit number. This new meta information allows us to implement a new subscribe mechanism based on a temporal filter. In the *GeoSubscribe* message, clients are able to specify this filter for a fixed period of time or for repetitive periods. The filter consists of two parameters—a start timestamp/expression and an end timestamp/interval length. Timestamps can be defined in ISO8601 format or Unix time. For repetitive periods, the start expression consists of a Cron expression and an interval length in seconds. Examples for the filter's parameter definition are given in Table 1.

**Table 1** Examples of the time filter used in GeoSubscribe messages

Example	Start timestamp/expression	End timestamp/interval length
Fixed (ISO)	2015-08-31T12:43:33	2015-09-15T10:00:00
Fixed (Unix)	1441017813	1442304000
Repetitive	0 0 8 ? * SAT	7200

The first two examples in the table define a fixed period between two timestamps in both cases. If the timestamp of the *GeoPublish* message lies within this period, the filter is evaluated to true. The start expression of the repetitive period is defined in Cron syntax according to Terracotta (2015). In the example, the start expression states “every Saturday morning at 8 am”. Given the interval length set to 7200 s, the temporal filter is satisfied if the timestamp of the *GeoPublish* message is between 8 am and 10 am on every Saturday morning. However, the client is also able to turn off the temporal filter completely.

### Spatial Filter

Besides the timestamp, the *GeoPublish* message also contains a mandatory coordinate as meta information. With this information, we can also integrate a spatial filter in our subscription mechanism. The spatial filter uses a simple “contains” relation. The client defines a geometry that it wants to subscribe to in the *GeoSubscribe* message. The broker checks for every incoming *GeoPublish* message which satisfies the topic and the temporal filter if the coordinates are located within the subscribed geometry. If it matches, the broker forwards the message to the subscriber.

Currently, the spatial filter is still in an early development phase enabling clients to subscribe to a 2D or 3D bounding box. This bounding box is specified by two coordinates in World Geodetic System 1984 (WGS84). If the bounding box contains the coordinate of the *GeoPublish* message, which is also defined in WGS84, the filter is satisfied. However, we are planning to replace this feature with a more comfortable solution using Extended Well-Known Text (EWKT) (The PostGIS Development Group 2015). This way, clients would be able to subscribe to more complex geometries specified in other coordinate systems as well.

### 3.2.2 Extending the MQTT Protocol

In detail, we introduced new message types in the MQTT protocol. The fixed header of every MQTT message consists of a 4-bit MQTT control packet type, so that MQTT allows up to 16 different message types. For the GeoMQTT extension we need three more message types: *GeoPublish*, *GeoSubscribe* and *GeoUnsubscribe*. Unfortunately, the core specification of MQTT already uses 14 of the 16 possible values with 0 and 15 reserved for future use. However, some of the core messages are only used in server to client flow direction (S2C), so that we can use the *SubAck* (9) and the *UnsubAck* (11) message types for our *GeoSubscribe* and

**Table 2** Message types in GeoMQTT

Name	Value	Direction of flow	Description
SUBACK_GEOSUB	9	C2S or S2C	Subscribe acknowledge and client geo subscribe request
UNSUBACK_GEOUNSUB	11	C2S or S2C	Unsubscribe acknowledge and client geo unsubscribe request
GEOPUBLISH	15	C2S or S2C	GeoPublish message

GeoUnsubscribe when sent from client to server (C2S). Since the direction of flow of the GeoPublish message type has to be bidirectional (C2S and S2C), we use one of the reserved values for this kind of message (value 15). The different message types introduced by GeoMQTT are illustrated in Table 2.

One advantage of this implementation is the backwards compatibility with existing MQTT brokers which do not support the GeoMQTT extension. The introduced message types are simply ignored since they are not known.

The *GeoSubscribe* and *GeoUnsubscribe* messages are of similar form like the ordinary Subscribe and Unsubscribe message types except for the two added filters. They force the same behavior in the broker, meaning, a geo subscribe or unsubscribe request by a client is acknowledged with a *SubAck* or *UnsubAck* message by the broker. Likewise, the *GeoPublish* message is similar to a simple *Publish* message except for the added timestamp and location. The broker verifies all three filters and forwards the messages respectively. However, all features of an ordinary *Publish* message are also supported for *GeoPublish* messages such as QoS or the retain feature.

### 3.2.3 Handling Conflicts Between MQTT and GeoMQTT

Conflicts may occur between the subscriptions of clients and MQTT *Publish* messages respectively GeoMQTT *GeoPublish* messages in the broker. For instance, if a client is subscribed to a topic filter with a MQTT subscription, and the broker receives a *GeoPublish* message with a topic name which matches the filter of the subscription, is it forwarded to the client or not?

In our prototype implementation of the GeoMQTT broker, we realized the following conflict handling strategies (see Table 3).

**Table 3** Conflict handling strategies between MQTT and GeoMQTT

	GeoMQTT message (GeoPublish)	MQTT message (Publish)
GeoMQTT subscription	Match by topic, spatial and temporal filter	Not forwarded
MQTT subscription	Temporal and spatial information ignored, match by topic filter and converted to base publish message	Match by topic

In the trivial cases the broker verifies the corresponding filters and forwards the messages respectively, meaning, if a client subscribed to a topic with a MQTT subscription and a MQTT publish message is being sent, the broker matches the topic name to the topic filter and forwards the message accordingly. Following this, if the client is subscribed by a topic, temporal and spatial filter (*GeoSubscribe*) and the broker receives a *GeoPublish* message, all three filters are matched to the meta information of the message and potentially forwarded. In case that a client is subscribed to a topic filter with a MQTT subscription and a *GeoPublish* message is published, the broker ignores the temporal and spatial information of the message and solely matches the topic filter. If it matches, the *GeoPublish* message is converted to an MQTT *Publish* messages and sent to the subscriber. Otherwise, if an ordinary publish message is received by the broker and a client is subscribed by a *GeoSubscribe* message, the message is not forwarded, even if the topic filter matches the topic name. This way, our conflict handling implementation is also compatible to MQTT clients that do not support the extension.

### 3.2.4 Extending the MQTT-SN Protocol

Currently, we are also working on a spatial extension for the MQTT-SN protocol which brings GeoMQTT to sensor nodes with limited resources. Like the MQTT-SN protocol, the GeoMQTT-SN protocol is based on the same message types but also adds the three new message types *GeoPublish*, *GeoSubscribe* and *GeoUnsubscribe*. The design guidelines of the MQTT-SN protocol are followed by this implementation (see Sect. 3.1.3).

### 3.2.5 GeoMQTT Implementation

A prototype broker is implemented by forking the “moquette” broker written in Java by Selva (2015). The entire filter matching logic and conflict handling strategy is realized by the broker, so our implementation follows the principle of thin clients and fat brokers.

We implemented several GeoMQTT clients in different programming languages such as Java and Python. Furthermore, we developed a QGIS plugin based on the Python client to enable GeoPipe streaming into a geographic information system (GIS). A JavaScript client is realized to use the GeoMQTT protocol via WebSockets and receive the data directly in the users’ web browsers of choice. The sensor nodes themselves use an avr-gcc compiler, thus, the GeoMQTT-SN client is implemented in C++ as well as the GeoMQTT-SN gateway.

## 4 Applications

As mentioned, the motivation behind the development of the GeoMQTT protocol extension is to integrate a timestamp and coordinates into the metadata of the MQTT publish message. This is achieved by introducing the *GeoPublish* message type. The original issue we want to solve is to bind high-level data storage services for measurements to sensor nodes in WGSN in a standardized way. In the course of development, we realized that this extension can be used for a wide range of real-time applications and enhance existing OWS services which are briefly discussed in the following sections.

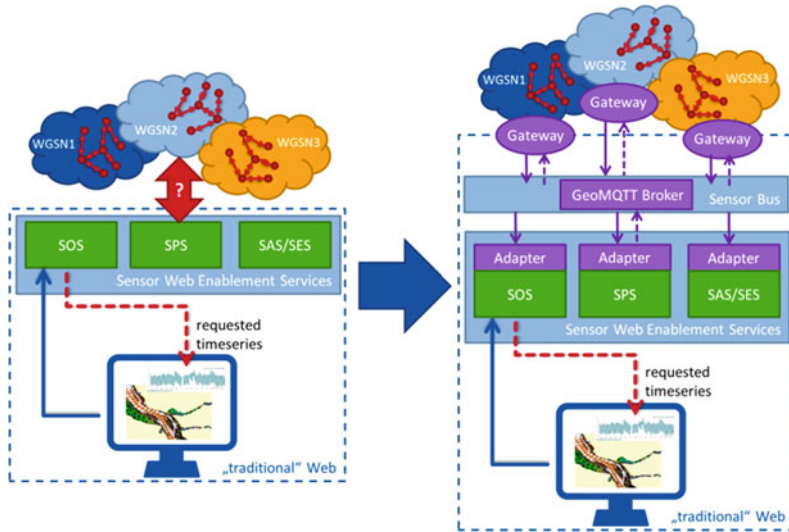
### 4.1 Closing the Interoperability Gap Between SWE Services

Sensor nodes in WGSNs are limited by nature in computing power, storage and energy consumption. For transmitting data among nodes or to a data sink, they are usually equipped with low-power digital radios like ZigBee which restricts the length of messages to a couple of bytes. On the other hand, sensor data should be stored in a way, so that they are connected to the World Wide Web and accessible in a standardized manner. This is sometimes termed as the “Sensor Web”. The SWE initiative of the OGC has already developed standards for accessing sensor data archives and metadata via web services like the Sensor Observation Service (SOS) (Grothe and Kooijman 2008). The objective of the Sensor Web is to hide “the underlying layers from the applications built on top of it” (Broering et al. 2011).

However, these standardized services mainly rely on the exchange of XML documents which is a huge disadvantage in environments with limited resources like sensor nodes in WGSNs. The installed hardware is not suitable for processing and transmitting large XML documents which lead to an interoperability gap (Walter and Nash 2009).

This interoperability gap between the low-level protocols used in WGSNs and the high-level protocols of e.g. a SOS for inserting or accessing the sensor data archives needs to be closed in a sophisticated way. In Broering et al. (2010) an intermediary layer is suggested to bypass the gap between the two systems. They introduce a so-called sensor bus which follows a message bus pattern. In the prototype implementation, they use different technologies like XMPP or IRC which are, however, still too high-level for a client implementation used by small sensor nodes.

With GeoPipes over GeoMQTT(-SN), we are able to close this gap between different devices and networks (see Fig. 3). The GeoMQTT broker serves as the sensor bus. The high-level services and the sensor nodes are connected with GeoMQTT clients and GeoMQTT-SN clients respectively. Since the payload of a GeoMQTT message is arbitrary, we can use different encoding standards to deliver the measurements adaptable to the capabilities of the sensor platform.



**Fig. 3** Sensor bus closing the gap between WGSNs and sensor web

One advantage of this architecture is that every client has the same features and rights, meaning even the sensor nodes can subscribe to messages and receive them. For example, we can alter the configuration on-the-fly by sending a message to the node. Since the GeoMQTT protocol allows clients to subscribe to a specific extent, sensor nodes are able to receive all data from the surrounding nodes even if they are not in the same physical network in near real-time. This enables the sensor nodes to self-validate their measured values or use other values to calibrate the connected sensors.

In addition, the raw sensor data can easily be postprocessed before stored persistently in the database. Ghobakhlou et al. (2014) suggest a so-called Sensor Data Processing Service (SDPS) which subscribes to a MQTT topic on which the raw sensor data is published by the sensor node. If the message matches a specific pattern, the raw data is postprocessed by the service and stored in a SOS database. With GeoPipes over GeoMQTT we can establish a spatial notification and data stream transfer mechanism. Post processors log into the sensor bus, subscribe to the (Geo)MQTT messages and receive the corresponding raw data published by a sensor node. After processing, the results are republished to the sensor bus and taken by the SOS adapter to store the filtered data persistently in the database.

### 4.2 Further Applications

By utilizing GeoPipes over GeoMQTT, we can also build more sophisticated real-time processing chains than simple postprocessors for sensor data. Process



chaining is not a newly introduced concept in geospatial workflows. With the Web Processing Service (WPS), the OGC already standardized an interface for coupling different spatial or non-spatial processes together (Stollberg and Zipf 2007). However, for real-time process chains it has some major drawbacks such as its deficiency of the asynchronous processing approach (Sagl et al. 2012). A WPS process can take another process's output as an input and compute a result but is not able to consider asynchronous incoming data. It requires modifications to run geo processes on streaming data since in the current version each processing task has a finite lifetime (McCullough et al. 2011).

GeoPipes using GeoMQTT might enhance the WPS standard with a notification mechanism as it is used in Echterhoff and Everding (2008) or Foerster et al. (2012) but with different technologies. Unlike the approach followed in Echterhoff and Everding (2008), we do not alter the WPS standard itself but use additional input parameters in the services to define the GeoPipes in an URL-like syntax. This can be used to send e.g. intermediate results of the process. The service can be invoked in the browser and with the GeoMQTT WebSocket implementation, intermediate results can be received. Additionally, GeoPipes can be defined as input pipes for the process, so that the process receives real-time geospatial data from other publishers like sensors.

Using the WPS interface usually means, that the service is invoked by a user. We can also think of process chains, similar to the sensor postprocessing mentioned earlier. The chain consists of different processes connected by GeoPipes waiting for the newest data to arrive. Data from other processes or sensor nodes etc. stream into these processes and the computed results are published back to an output pipe.

## 5 Conclusions and Future Work

GeoPipes using GeoMQTT offer a simple geospatial notification and data streaming transfer mechanism. With the spatial MQTT extension we introduce a publish/subscribe system for exchanging geospatial data in a pushed-based manner. We use this protocol for overcoming major disadvantages of standardized OGC geo web services which do not meet the requirements of IoT environments. Since GeoMQTT extends MQTT, all important and useful features are also available. Our requirements for GeoPipes in the IoT are met by this implementation, currently with some trade-offs. The main strength of the GeoPipes mechanism using GeoMQTT is that channels can be established to share geospatial data streams between distributed instances. These instances can be different kinds of devices, applications and software which can be linked easily in this way. The example applications already illustrate the capabilities of GeoPipes using GeoMQTT. As mentioned, it can especially be used to solve drawbacks in standardized geo web services. The interoperability gap between the services of the SWE and the sensor devices can be closed in an elegant way. Furthermore, asynchronous processing can

be added to the WPS interface. That way, we are also able to realize real-time geo processing chains which are invoked by events or the user.

The proposed MQTT extension is still under development. For example, we are planning to replace the spatial subscription mechanism with a more sophisticated one based on EWKT, also supporting other coordinate systems and more complex geometries. Furthermore, we need to implement security and scalability features which are not included in the MQTT core specification, but are already implemented in some brokers. So far, we have implemented prototypes of the GeoMQTT broker and some clients such as a Python and Java client as well as a JavaScript client using WebSockets. Currently, we are developing a GeoMQTT-SN extension to additionally be able to integrate sensor nodes in connectionless environments as clients. Further, our future work involves creating a REST interface to bridge GeoMQTT and HTTP as well as integrating geospatial data mining facilities using Apache Storm.

## References

- Broering A, Foerster T, Jirka S, Priess C (2010) An intermediary layer for linking geosensors and the sensor web. In: Proceedings of COM.Geo 2010, 1st international conference on computing for geospatial research and applications. ACM, Washington, DC
- Broering et al (2011) New generation sensor web enablement. *Sensors* 11(3):2652–2699
- Echterhoff J, Everding T (2008) OGC sensor event service interface specification. OpenGIS Discussion Paper. [http://portal.opengeospatial.org/files/?artifact\\_id=29576](http://portal.opengeospatial.org/files/?artifact_id=29576). Accessed 2 Dec 2015
- Echterhoff J, Everding T (2011) OGC event service: review and current state. [https://portal.opengeospatial.org/files/?artifact\\_id=45850](https://portal.opengeospatial.org/files/?artifact_id=45850). Accessed 2 Dec 2015
- Eugster P, Felber A, Kermarrec, Guerraoui A-M (2003) The many faces of publish/subscribe. *ACM Comput Surv* 35(2):114–131
- Foerster T, Baranski B, Borsutzky H (2012) Live geoinformation with standardized geoprocessing services. In: Gensel J, Josselin D, Vandenbroucke D (eds) Bridging the geographic information sciences, international AGILE'2012 conference, Avignon, 2012. Lecture notes in geoinformation and cartography. Springer, Heidelberg, pp 99–118
- Ghobakhlou A, Sallis P, Wang X (2014) A service oriented wireless sensor node management system. In: Proceedings of the instrumentation and measurement technology conference (I2MTC), Montevideo, 2014
- Grothe M, Kooijman J (2008) Sensor web enablement. Netherlands Geodetic Commission, Netherlands
- Hunkeler U, Truong H, Stanford-Clark, A (2008) MQTT-S: a publish/subscribe protocol for wireless sensor networks. In: Proceedings of the 3rd international conference on communication systems software and middleware and workshops (COMSWARE'08), pp 791–798
- McCullough A, Barr S, James P (2011) A typology for real-time parallel geoprocessing for the sensor web era. In: Foerster T et al (eds) Integrating sensor web and web-based geoprocessing. CEUR Workshop proceedings, vol 712, Utrecht 2011
- MQTT.org (2015) Frequently asked questions. <http://mqtt.org/faq>. Accessed 2 Dec 2015
- OASIS (2014) MQTT Version 3.1.1 OASIS Standard, 29 Oct 2014
- OGC (2015) OGC requests comment on publish/subscribe standards for geospatial data. <http://www.opengeospatial.org/standards/requests/138> Accessed 2 Dec 2015

- Sagl G et al (2012) Standardised geo-sensor webs and web-based geo-processing for near real-time situational awareness in emergency management. *Int J Bus Contin Risk Manag* 3 (4):339–358
- Selva A (2015) Moquette—Java MQTT lightweight broker. <https://github.com/andsel/moquette>. Accessed 2 Dec 2015
- Simonis I (2007) OGC Sensor alert service candidate implementation specification. [http://portal.opengeospatial.org/files/?artifact\\_id=15588](http://portal.opengeospatial.org/files/?artifact_id=15588). Accessed 2 Dec 2015
- Simonis I, Echterhoff J (2007) Draft OpenGIS web notification service implementation specification. [http://portal.opengeospatial.org/files/?artifact\\_id=18776](http://portal.opengeospatial.org/files/?artifact_id=18776). Accessed 2 Dec 2015
- Stanford-Clark A, Truong H (2013) MQTT for Sensor Networks (MQTT-SN), Protocol Specification, Version 1.2
- Stollberg B, Zipf A (2007) OGC web processing service interface for web service orchestration aggregating geo-processing services in a bomb threat scenario. In: Ware M, Taylor E (eds) *Web and wireless geographical information systems*, 7th international symposium, W2GIS 2007, Cardiff, 2007. Lecture notes in computer science, vol 4857. Springer, Heidelberg, pp 239–251
- Terracotta (2015) Quartz CronTrigger Tutorial, <http://www.quartz-scheduler.org/documentation/quartz-1.x/tutorials/crontrigger>. Accessed 2 Dec 2015
- The PostGIS Development Group (2015) PostGIS 2.2.1dev Manual, Chapter 4.1.2 PostGIS EWKB, EWKT and Canonical Forms. [http://postgis.net/docs/using\\_postgis\\_dbmanagement.html#EWKB\\_EWKT](http://postgis.net/docs/using_postgis_dbmanagement.html#EWKB_EWKT). Accessed 2 Dec 2015
- Walter K, Nash E (2009) Coupling wireless sensor networks and the sensor observation service—bridging the interoperability gap. In: Haunert J-H, Kieler B, Milde J (eds) *Proceedings of the 12th AGILE international conference on geographic information science*, Hannover, 2009, pp 99–118
- Westerholt R, Resch B (2014) Asynchronous geospatial processing: an event-driven push-based architecture for the OGC web processing service. *Trans GIS* 19(3):455–479

# Continuous Generalization of Administrative Boundaries Based on Compatible Triangulations

Dongliang Peng, Alexander Wolff and Jan-Henrik Haurert

**Abstract** Continuous generalization aims to produce maps at arbitrary scales without abrupt changes. This helps users keep their foci when working with digital maps interactively, e.g., zooming in and out. Topological consistency is a key issue in cartographic generalization. Our aim is to ensure topological consistency during continuous generalization. In this paper, we present a five-step method for continuously generalizing between two maps of administrative boundaries at different scales, where the larger-scale map has not only more details but also an additional level of administrative regions. Our main contribution is the proposal of a workflow for generalizing hierarchical administrative boundaries in a continuous and topologically consistent way. First, we identify corresponding boundaries between the two maps. We call the remaining boundary pieces (on the larger-scale map) *unmatched* boundaries. Second, we use a method based on so-called *compatible triangulations* to generate additional boundaries for the smaller-scale map that correspond to the unmatched boundaries. Third, we simplify the resulting additional boundaries. Fourth, we determine corresponding points for each pair of corresponding boundaries using a variant of an existing dynamic programming algorithm. Fifth, we interpolate between the corresponding points to generate the boundaries at intermediate scales. We do a thorough case study based on the provinces and counties of Mainland China. Although topologically consistent algorithms for the third step and the fifth step exist, we have implemented simpler algorithms for our case study.

**Keywords** Map · Scale · Topological consistency · Dynamic programming · Morphing · Rubber-sheeting · Mainland China

---

D. Peng (✉) · A. Wolff

Chair of Computer Science I, University of Würzburg, Würzburg, Germany  
e-mail: dongliang.peng@uni-wuerzburg.de

A. Wolff

URL: [http://www1.informatik.uni-wuerzburg.de/en/staff/wolff\\_alexander/](http://www1.informatik.uni-wuerzburg.de/en/staff/wolff_alexander/)

J.-H. Haurert

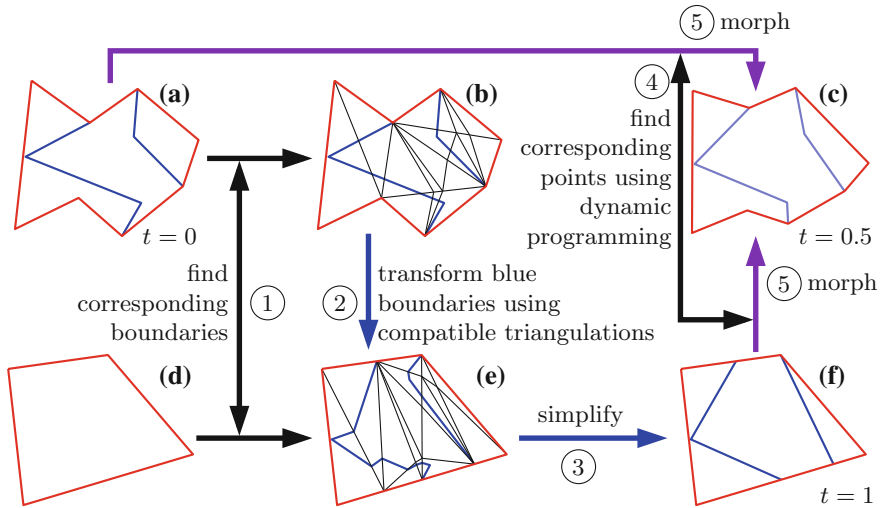
Institute for Geoinformatics and Remote Sensing (IGF),  
University of Osnabrück, Osnabrück, Germany  
e-mail: janhhaunert@uni-osnabrueck.de

# 1 Introduction

Nowadays people often browse through digital maps on computers or small displays to get geographic information. To understand maps better, users interactively zoom in and out to read maps from different levels. A typical strategy to support zooming is based on a multiple representation database (MRDB). Such a database stores a discrete set of levels of detail (LODs) from which a user can query the LOD for a particular scale (Hampe et al. 2004). A small set of LODs leads, however, to complex and sudden changes during zooming. Since this distracts users, hierarchical schemes have been proposed that generalize a more-detailed representation to obtain a less-detailed representation based on small incremental changes, e.g., the BLG-tree (van Oosterom 2005) for line simplification or the GAP-tree (van Oosterom 1995) for area aggregation. Such incremental generalization processes are represented in data structures that allow users to retrieve a map of any scale. Still, the generalization process consists of discrete steps and includes abrupt changes. Discrete steps can easily cause users to lose their “mental map” during interaction, which is annoying. To support continuous zooming, van Kreveld proposed five ways of gradual changes, which are *moving*, *rotating*, *morphing*, *fading*, and *appearing* (van Kreveld 2001). This is actually a step towards continuous generalization. Comparing to traditional generalization, continuous generalization has the advantage of generating maps without abrupt changes, which is ideal to improve zooming experience. To achieve continuous generalization, Sester and Brenner (2004) suggested simplifying building footprints based on small incremental steps and animating each step smoothly. With the same objective, Danciger et al. (2009) investigated the growing of regions, meanwhile preserving their topology, area ratios, and relative position. The strategy of using two maps at different scales to generate intermediate-scale maps has been studied in multiple representations, e.g., with respect to the selection of roads or rivers (Girres and Touya 2014). Actually, this strategy is the key idea of morphing-based methods for continuous generalization. For instances, several authors have developed methods for morphing between polylines (Cecconi 2003; Nöllenburg et al. 2008; Peng et al. 2013; Schneider and Hormann 2015) and methods for morphing between raster maps (Pantazis et al. 2009; Reilly and Inkpen 2004).

Topological consistency is a property that must be guaranteed in continuous generalization. In this paper, we investigate the problem of generalizing continuously a two-level hierarchical subdivision—from a larger-scale map of administrative boundaries to a smaller-scale map; see Fig. 1 for an example. Our aim is to generate maps at any intermediate scale without topological conflicts by generalizing continuously (going from Fig. 1a–d). Our method consists of the following five steps.

In Step ①, we find corresponding boundaries; see the red boundaries in Fig. 1a and d. We call the remaining boundaries on the larger-scale map *unmatched* boundaries; see the blue boundaries in Fig. 1a. In order to achieve continuous generalization, we *morph* (that is, deform continuously) between the corresponding boundaries (see, e.g., Nöllenburg et al. 2008). The unmatched boundaries must be morphed in



**Fig. 1** The framework of our method. **a** The larger-scale administrative boundaries of a region. **d** The smaller-scale administrative boundaries of the same region as in (a). **b, e** Compatible triangulations constructed for red polygons in (a) and (d), where the blue polylines in (e) are transformed from the blue boundaries in (b) based on the compatible triangulations. **f** The blue boundaries are simplified from the transformed ones in (e). **c** The result of continuous generalization when  $t = 0.5$ . Continuous generalization is achieved by morphing between (a) and (f), without the blue boundaries displayed in (f). The numbers in the circles indicate the step orders

a way that is consistent with what we do to the corresponding boundaries. As there is no correspondence for the unmatched boundaries, we generate the corresponding boundaries in Step ② and Step ③.

In Step ②, we transform the blue boundaries based on *compatible triangulations*; see Fig. 1b and e. Two triangulations are *compatible* if there is a correspondence between their vertex sets and the two triangulations are topologically equivalent (Surazhsky and Gotsman 2001). With compatible triangulations, we can transform a blue boundary in one triangulation (see Fig. 1b) to a boundary in the other triangulation by traversing the triangles correspondingly (see Fig. 1e). Therefore, if there is no conflict in one triangulation, then there is no conflict in the other triangulation.

In Step ③, we simplify the transformed (blue) boundaries, using the Douglas–Peucker algorithm (Douglas and Peucker 1973), so that the blue boundaries have the same complexities as the red ones in Fig. 1d; see the blue boundaries in Fig. 1f. We use the simplified boundaries as the correspondences for the blue boundaries in Fig. 1a. On this basis, we are able to generalize continuously by interpolating between each pair (both red pairs and blue pairs) of corresponding boundaries. This process of interpolation is also known as morphing. Since the blue boundaries do not exist on the smaller-scale map, we fade them out during morphing to make them disappear continuously.

In Step ④, we determine corresponding points for each pair of corresponding boundaries using a variant of an existing dynamic programming algorithm (Nöllenburg et al. 2008). Our variant minimizes the distance between corresponding points, which is favorable for Step ⑤.

In Step ⑤, we interpolate (or “morph”) between the corresponding points with straight-line trajectories to generate intermediate-scale administrative boundaries.

In order to achieve a topologically consistent workflow, we need to make sure that three of the above steps are topologically consistent, namely Step ②: transform (b–e), Step ③: simplify (e–f), and Step ⑤: morph (a–c–f); see Fig. 1. In this paper, we concentrate on the transformation step (Step ②); Fig. 1b–e) and show how to accomplish this step without introducing topological conflicts. We do not guarantee topological consistency of the other two steps *in our implementation*, but this can be done by integrating topologically consistent methods as proposed by Saalfeld (1999) for Step ③ and Gotsman and Surazhsky (2001) for Step ⑤.

Initially, we tested the rubber-sheeting method of Doytsher et al. (2001) (making all vertices “influential”). We soon noticed, however, that transformed boundaries often cross boundaries of the smaller-scale map. An example is shown in Fig. 2 (which corresponds to Fig. 1e), where the rubber-sheeting method leads to two crossings. Similar problems occurred when we applied other variants of rubber-sheeting (such as the one by Haunert 2005).

This is why we decided to search for a more robust method. It turned out that compatible triangulations (Aronov et al. 1993) can, by definition, realize the transformation without introducing topological conflicts. The (quite old) idea is as follows. Suppose that there is a point  $a_i$  inside a triangle  $\Delta p_1 p_2 p_3$ . Then, this point can be expressed as a *unique* convex combination of *simplicial coordinates*  $\lambda_{i,1}$ ,  $\lambda_{i,2}$ ,  $\lambda_{i,3}$  (Saalfeld 1985):

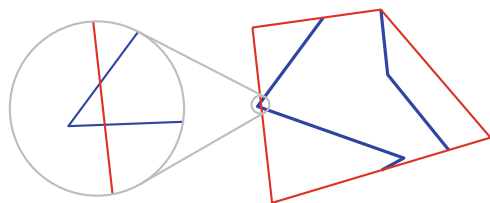
$$a_i = \lambda_{i,1} p_1 + \lambda_{i,2} p_2 + \lambda_{i,3} p_3,$$

where  $\lambda_{i,1}, \lambda_{i,2}, \lambda_{i,3} > 0$ , and  $\lambda_{i,1} + \lambda_{i,2} + \lambda_{i,3} = 1$ . We can *uniquely* locate the corresponding point  $b_i$  in a different triangle  $\Delta q_1 q_2 q_3$  by using  $a_i$ 's simplicial coordinates in  $\Delta p_1 p_2 p_3$ :

$$b_i = \lambda_{i,1} q_1 + \lambda_{i,2} q_2 + \lambda_{i,3} q_3.$$

This implies that if two distinct points  $a_i$  and  $a_j$  are in the same triangle, then we are able to locate their corresponding points  $b_i$  and  $b_j$  in another triangle such that  $b_i$  and  $b_j$  do not coincide. Moreover, if points  $a_i$  and  $a_j$  are in two different triangles

**Fig. 2** Crossings caused by the rubber-sheeting method of Doytsher et al. (2001)



of a triangulation, then  $b_i$  and  $b_j$  can be located in two different corresponding triangles of the compatible triangulation. As a result, once we manage to construct compatible triangulations of two maps, we can transform polylines consistently, which means the transformed polylines will not conflict with each other and will not conflict the boundaries already on the smaller-scale map. Indeed, compatible triangulations have been applied to compare maps from different time periods by hand (Fuse and Shimizu 2004). Instead, we construct compatible triangulations automatically, using an algorithm of Aronov et al. (1993).

Our contributions are as follows. We propose a workflow based on compatible triangulations for generalizing administrative boundaries in a continuous and topologically consistent way; see Sect. 2. We do a thorough case study with the boundaries of the counties and the provinces of Mainland China to test the effectiveness and the efficiency of our method; see Sect. 3. We conclude in Sect. 4.

## 2 Overall Algorithm

Suppose that we have two maps  $M_+$  and  $M_-$  of administrative boundaries of the same area, respectively at a larger scale and a smaller scale. We use a parameter  $t \in [0, 1]$  to define the continuous generalization process. We require that the generalization yields exactly  $M_+$  when  $t = 0$ ,  $M_-$  when  $t = 1$ , and that  $M_+$  is being generalized continuously into  $M_-$  when  $t$  increases from 0 to 1.

The map  $M_+$  is more detailed than  $M_-$ , and a region of  $M_-$  consists of several regions of  $M_+$ . Consequently, a boundary on  $M_-$  has a corresponding boundary on  $M_+$ , but a boundary on  $M_+$  may not have a smaller-scale correspondence. Thus we need to determine the corresponding boundaries between the two maps, and the leftovers on  $M_+$  are the unmatched boundaries. For a pair of corresponding boundaries, we use a dynamic programming algorithm similar to the algorithm OPTCOR (Nöllenburg et al. 2008) to determine corresponding points. Then, we morph between corresponding points on straight-line trajectories. For an unmatched boundary on  $M_+$ , we generate its corresponding boundary by first transforming the boundary using a compatible-triangulation-based method, and then simplifying the transformed boundary with the Douglas–Peucker algorithm (Douglas and Peucker 1973). As a result, we are able to morph between the unmatched boundaries and the generated boundaries (i.e., the simplified ones). We fade the morphing results of the unmatched boundaries out, such that they will disappear when  $t = 1$ .

The administrative regions are polygons. An administrative region usually shares different parts of its boundary with other administrative regions. These shared parts should be always shared during the continuous generalization process. When generalizing administrative boundaries, we should avoid processing a shared boundary twice. Therefore, we prefer to treat administrative boundaries as a set of consecutive polylines instead of polygons when it is possible. For the input of two maps, we first preprocess them to obtain the correspondences between the polylines.



## 2.1 Preprocessing

Given polygons on map  $M_+$  and map  $M_-$ , the preprocessing consists of three steps. Note that if the inputs are boundaries of the polygons, i.e., polylines, we can easily generate the polygons based on a technique that is known as doubly-connected edge list (Berg et al. 2008). First, we make a copy of the polygons on  $M_+$  and merge the copied polygons according to the polygons on  $M_-$ . For each copied polygon, we try overlapping it with every polygon on  $M_-$ , and record the one that has the largest overlap area. Then we merge all the copied polygons that record the same polygon on  $M_-$ . The merging step can be done more correctly with the help of semantic attributes if possible, but one should be careful about that there may be some separate parts (e.g., enclaves) with ambiguous semantic values.

Second, we obtain matched polylines respectively from the boundaries of the merged polygons and the polygons on  $M_-$ . We define that a vertex is an intersection if the vertex has degree at least 3. As the merged polygons and the polygons on  $M_-$  have corresponding intersections, we utilize these intersections to find matched polylines. We split the boundaries of the polygons at every intersection, respectively for the merged polygons and the polygons on  $M_-$ . Note that two polygons on the same map may share some parts of their boundaries, it is sufficient to take only one copy of the shared parts. Then we match the split boundaries of the two sources to get matched polylines. As we render these matched polylines using red color, we call them *the red polylines on  $M_+$*  and *the red polylines on  $M_-$* . Although there is a data-matching system (Mustière and Devogele 2008) available, we use a simple method to attain the matching. We match the split boundaries according to the overlap areas of their buffers. The buffer-based method works well in our case study as corresponding boundaries have relatively close positions.

Third, we obtain unmatched polylines on  $M_+$ . We split the boundaries of the polygons on  $M_+$  at every intersection, then we exclude all the split boundaries that overlay with the matched polylines on  $M_+$ . The remained polylines are *the blue polylines on  $M_+$*  as we render them using blue color.

After the preprocessing, we have three types of polylines. The first type consists of the red (matched) polylines on  $M_+$ , namely each of these polylines has a corresponding polyline on  $M_-$ ; see the red polylines in Fig. 1a. The second type consists of the blue (unmatched) polylines on  $M_+$ , each of which does not have a corresponding polyline on  $M_-$ ; see the blue polylines in Fig. 1b. The third type consists of the red (matched) polylines on  $M_-$ ; see the red polylines in Fig. 1d.

## 2.2 Morphing a Polyline to Its Corresponding Polyline

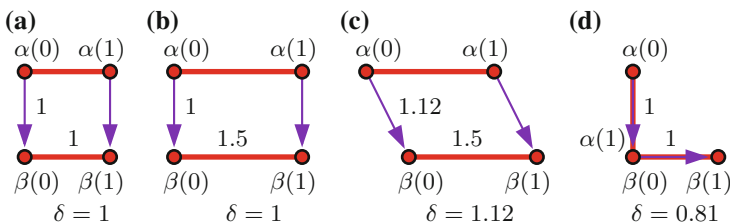
For a pair of corresponding polylines, one being a red polyline on  $M_+$  and the other being a red polyline on  $M_-$ , we use a variant of the dynamic programming algorithm

OPTCOR of Nöllenburg et al. (2008) to determine corresponding points (possibly injecting additional vertices).

The algorithm OPTCOR models the problem of finding corresponding points as an optimum correspondence problem, with respect to a cost function, between characteristic parts of each pair of corresponding polylines. The characteristic part is either a characteristic point or a subpolyline between a pair of neighbouring characteristic points. OPTCOR considers three cases of a correspondence for a subpolyline, namely, a subpolyline is mapped to a characteristic point, mapped to a subpolyline, or mapped to a merged sequence of subpolylines. We call all the three cases corresponding subpolylines as a point is a degenerate subpolyline and a merged sequence of subpolylines is still a subpolyline. Nöllenburg et al. (2008) describe a way to find characteristic points in order to speed up the dynamic program that computes the correspondences. However, they also mention that simply all vertices can be used as characteristic points, which is what we decided to do in order to obtain high-quality correspondences.

For a pair of corresponding subpolylines, Nöllenburg et al. (2008) define the cost as a combination of three values: (i) the distance between the corresponding points, (ii) the length difference of the pair of subpolylines, and (iii) the changes of the vectors of the corresponding points. Then OPTCOR determines the corresponding points by dynamically “looking back” to combine the at most  $k$  last subpolylines into a sequence of subpolylines, while minimizing the cost over all corresponding polylines. Here,  $k$  is the user specified look-back parameter, which gives a trade-off between quality and efficiency.

To make the problem simple, our variant considers only the first value as the cost, that is, the distance between corresponding points. Nöllenburg et al. denote this distance by  $\delta_i$ , but we will simply use  $\delta$ . In order to compute the cost function, we linearly interpolate between each pair of corresponding subpolylines so that each vertex on one subpolyline has a, possibly injected, corresponding vertex on the other one. The pairs of corresponding vertices subdivide the (sub)polylines into corresponding line segments. Note that a line segment is (part of) an edge of a polyline. The cost of a pair of (whole) polylines is the sum of the costs for each pair of corresponding line segments. The cost for a pair of corresponding line segments can be computed as follows (for a few examples, see Fig. 3).



**Fig. 3** Examples of computing  $\delta(f, g)$ . The values in the subfigures represent the lengths of the edges

Let polyline  $F$  on  $M_+$  and polyline  $G$  on  $M_-$  be a pair of corresponding polylines. Let  $f = \overline{\alpha(0)\alpha(1)}$  be a line segment on  $F$ , and let  $g = \overline{\beta(0)\beta(1)}$  be a line segment on  $G$  that corresponds to  $f$ . Let  $\alpha(0) = (x_1, y_1)$ ,  $\alpha(1) = (x_2, y_2)$ ,  $\beta(0) = (x_3, y_3)$ , and  $\beta(1) = (x_4, y_4)$ , which are already known. We define a pair of corresponding points  $\alpha(u) \in f$  and  $\beta(u) \in g$  as

$$\begin{aligned} \alpha(u) &= ((1 - u)x_1 + ux_2, (1 - u)y_1 + uy_2), \\ \beta(u) &= ((1 - u)x_3 + ux_4, (1 - u)y_3 + uy_4). \end{aligned}$$

We define the cost for the correspondence between  $f$  and  $g$  as the integral over the distances between the pairs of corresponding points (suppose that we move  $\alpha(u)$  to  $\beta(u)$ ), that is,

$$\delta(f, g) = \int_0^1 |\beta(u) - \alpha(u)| du,$$

where  $|\beta(u) - \alpha(u)|$  is the Euclidian distance between  $\alpha(u)$  and  $\beta(u)$ , and can be represented as  $\sqrt{au^2 + bu + c}$ . The coefficients  $a$ ,  $b$ , and  $c$  depend on the coordinates of  $\alpha(0)$ ,  $\alpha(1)$ ,  $\beta(0)$ , and  $\beta(1)$ , as follows.

$$\begin{aligned} a &= (x_1 - x_2 - x_3 + x_4)^2 + (y_1 - y_2 - y_3 + y_4)^2, \\ b &= -2((x_1 - x_3)(x_1 - x_2 - x_3 + x_4) \\ &\quad + (y_1 - y_3)(y_1 - y_2 - y_3 + y_4)), \\ c &= (x_1 - x_3)^2 + (y_1 - y_3)^2. \end{aligned}$$

Let  $X = au^2 + bu + c$ . We have  $a \geq 0$ , and since  $X \geq 0$  ( $X$  is the square of the Euclidian distance),  $\Delta = 4ac - b^2 \geq 0$ . Note that, if  $a = 0$ , then  $b = 0$ . Let

$$\delta(f, g) = \int_0^1 |\beta(u) - \alpha(u)| du = \int_0^1 \sqrt{X} du.$$

Then  $\delta(f, g)$  can be computed, according to Bronstein et al. (p. 1064, integrals 241 and 245) (2001), as follows:

$$\delta(f, g) = \begin{cases} \sqrt{cu}|_0^1 & \text{if } a = 0, \\ \frac{(2au+b)\sqrt{X}}{4a}|_0^1 & \text{if } a > 0, \Delta = 0, \\ \frac{(2au+b)\sqrt{X}}{4a}|_0^1 + \frac{\Delta}{8a\sqrt{a}} \ln(2\sqrt{aX} + 2au + b)|_0^1 & \text{if } a > 0, \Delta > 0. \end{cases}$$

Figure 3 shows a few examples where we compute  $\delta$ . We obtain the optimum correspondence by minimizing the sum of the distances between corresponding points weighted by the lengths of the line segments, that is,

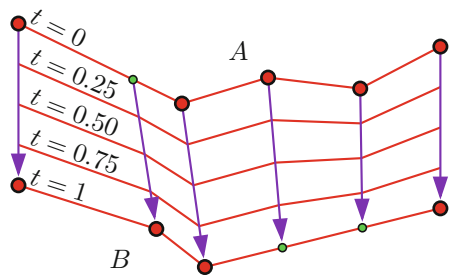
$$\delta(F, G) = \min_{\substack{\pi: \text{correspondence} \\ \text{between } F \text{ and } G}} \sum_{\substack{f \in F, g \in G, \\ \text{where } f \text{ corresponds to } g \text{ in } \pi}} \frac{|f| + |g|}{2} \delta(f, g).$$

Recall that OPTCOR considers three cases of a correspondence for a subpolyline. We find that the first case, i.e., a subpolyline is mapped to a characteristic point, may result in different numbers of vertices on the two polylines. Our major change to OPTCOR is that we remove this first case from the algorithm. This change ensures that the resulting pairs of corresponding polylines have the same numbers of vertices. This property is essential for the construction of compatible triangulations, which is a later step in our workflow. We name our modified version OPTCOR-S, where the letter *S* stands for *Simplified*.

Suppose that there are  $n_F$  vertices on  $F$  and  $n_G$  vertices on  $G$ , then OPTCOR-S requires that the look-back parameter  $k$  is bounded from below by  $n_F/n_G$  and  $n_G/n_F$ . Otherwise, there will be at least one segment that corresponds to a vertex. In our experiments, we always use a value of  $k$  that produces results with high quality (in the sense of the dynamic-programming algorithm). We morph by interpolating between corresponding points using straight-line trajectories. Figure 4 shows an example with 0.25, 0.5, and 0.75 for  $t$ .

Some other algorithms for determining corresponding points can be used (e.g., linear interpolation). We observed that an algorithm that determines corresponding points more carefully yields better results, meaning that the interpolated polylines are more similar to the two sources and it is less likely that crossings are introduced. This is why we decided to modify the algorithm OPTCOR. Some sophisticated algorithms can be considered to define the interpolation trajectories, such as geodesic shortest paths (Bereg 2005), b-morphs (Whited and Rossignac 2011), or a method based on least squares adjustment (Peng et al. 2013). Specifically, it is possible to use compatible triangulations not only for the transformation step (as in our method) but also to ensure the topological consistency in the morphing step (Surazhsky and Gotsman 2001).

**Fig. 4** Morphing a polyline  $A$  to its corresponding polyline  $B$ . The arrows show the moving trajectories of the vertices



### 2.3 Morphing a Polyline to Its Generated Corresponding Polyline During Fade-Out

For the blue polylines on  $M_+$ , morphing them must be consistent with what we do to the red corresponding polylines. To achieve this, we generate their corresponding polylines, that is, blue polylines on  $M_-$ . We first transform the blue polylines on  $M_+$  based on compatible triangulations, and then simplify the transformed polylines using the Douglas–Peucker algorithm.

For each pair of polygons correspondingly bounded by the red polylines on  $M_+$  and the red polylines on  $M_-$ , we construct a pair of compatible triangulations; see Fig. 1b and e. We call them the *triangulation on  $M_+$*  and the *triangulation on  $M_-$* . The construction of compatible triangulations requires that the two polygons have the same number of vertices. We have ensured this property using OPTCOR-S to determine corresponding points between the red polylines on  $M_+$  and  $M_-$ . We use the algorithm of Aronov et al. (1993) to construct compatible triangulations. For the two polygons both with  $m$  vertices, we first triangulate them independently. Then we create a regular  $m$ -gon, and map the chords of the two triangulations into the regular  $m$ -gon and overlay the mapped chords. The mapped chords may cross with each other, which results in some convex faces. We use the crossings as dummy vertices and split the mapped chords. We triangulate each convex face with more than three vertices. To triangulate, we select one vertex and add an edge between this vertex to each of the other vertices, except the two immediate-neighbouring ones. As a result, we have a *combined triangulation*. We map the combined triangulation (including dummy vertices and new edges) back to modify the two original triangulations. After the modification, we have a pair of compatible triangulations of the two original polygons.

With the compatible triangulations, we can transform a blue polyline in the triangulation on  $M_+$  to a polyline on  $M_-$ , using the same simplicial coordinates. To guarantee that the transformed polyline traverses exactly the “same” triangles as a blue polyline on  $M_+$ , we need to compute the crossings between the blue polyline and the triangulation edges, and also transform these crossings into the triangulation on  $M_-$ . Because of the additional crossings, the transformed polylines have higher complexity than the red polylines on  $M_+$ . While our aim is to generate polylines that have the same complexity as the red polylines on  $M_-$ . Therefore, we simplify the transformed polylines to the target complexity; see Fig. 1f. For a *blue hole* (polygon), we keep at least three vertices to avoid degenerating it into a straight line or a point. We call the simplified polylines the *blue polylines on  $M_-$* .

We use the algorithm OPTCOR-S again to determine corresponding points for each pair of corresponding blue polylines, which are respectively on  $M_+$  and  $M_-$ . We use straight-line trajectories to interpolate between corresponding points. As the blue polylines do not exist when  $t = 1$ , we fade them out during the morphing process. An example is shown in Fig. 1c, where  $t = 0.5$ .

## 2.4 Running Time Analysis

We analyze the running time for a pair of polygons correspondingly bounded by the red polylines on  $M_+$  and the red polylines on  $M_-$ . We use  $N$  to denote the vertex number of the polygon on  $M_+$ ,  $n$  the vertex number of the polygon on  $M_-$ , and  $N'$  the vertex number of all the blue polylines inside the polygon on  $M_+$ . For simplicity, we assume that  $O(N') \in O(N)$ .

Constructing the compatible triangulations takes  $O(N \log N + l)$  time, where  $O(l) \in O(N^2)$  is the number of dummy vertices inserted during the construction. Simplifying the transformed polylines, using the Douglas–Peucker algorithm, costs time  $O(N(N + l) \log N)$  (Hershberger and Snoeyink 1992). OPTCOR-S takes time  $O(k^2 N n)$  to determine corresponding points, where  $k$  is the look-back parameter. Note that, outputting a representation at a target zoom level only takes  $O(N)$  time (and, hence, is feasible in real time). In fact, we store, for each (possibly injected) vertex  $v$  on  $F$ , a representation such as  $v(t) = (1 - t)v + tw$ , where  $w$  is the vertex on  $G$  that corresponds to  $v$ . In our implementation, determining the corresponding points is the by far most time-consuming step.

## 3 Case Study

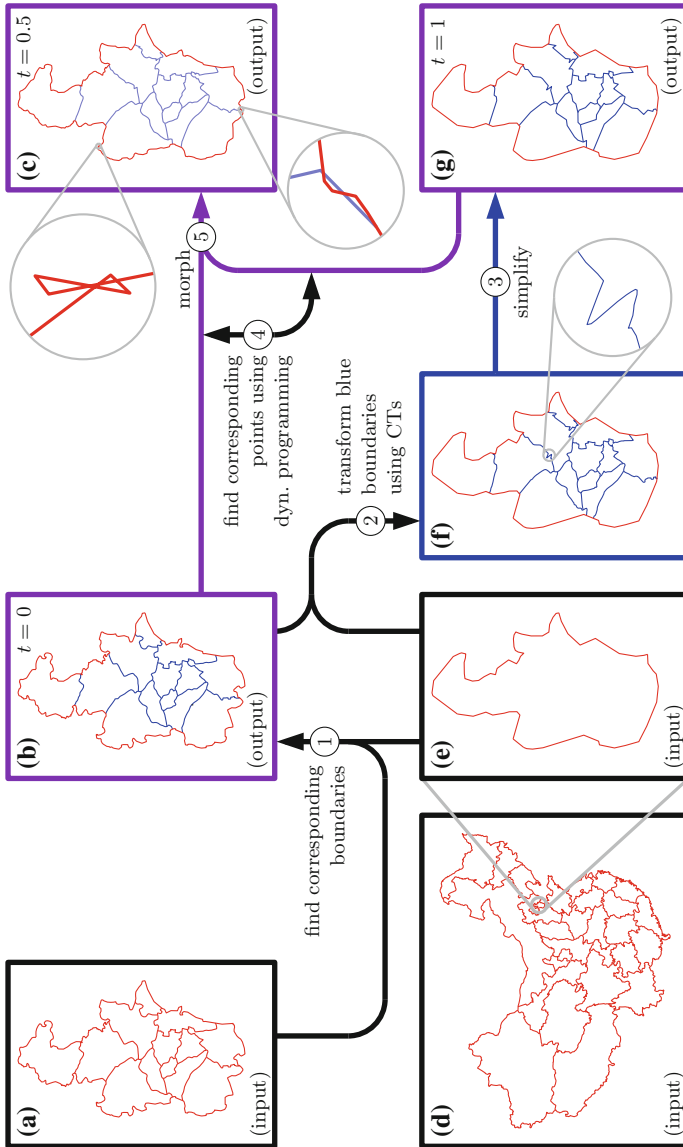
We implemented our method based on C# (Microsoft Visual Studio 2010) and ArcGIS Engine 10.1. We ran our case study under Windows 7 on a 3.3 GHz quad core CPU with 8 GB RAM. We measured time consumption by the built-in C# method `System.Environment.TickCount`.

We tested our method on the administrative boundaries of Mainland China, which are from the National Fundamental Geographic Information System, and based on the projected coordinate system *Krasovsky 1940 Lambert Conformal Conic*; see the provincial map in Fig. 5b. The tested data sets are obtained from the complete data sets of China by removing the only enclave in Gansu province and all the islands. We use a data set of county boundaries at scale 1 : 5,000 k with 493,625 vertices (5,909 polylines after preprocessing), and a data set of provincial boundaries at scale 1 : 30,000 k with 7,527 vertices (90 polylines after preprocessing). Since we can hardly see the details if we present the whole data set, we only show the results of the Tianjin province in Fig. 5.<sup>1</sup>

Our aim is to generalize continuously from map of counties  $M_+$  to map of provinces  $M_-$ . Using the provincial boundaries in Fig. 5e, we are able to distinguish the hierarchies of county boundaries in Fig. 5a, and then find the matched polylines (the red polylines in Fig. 5b and e) and the unmatched polylines (the blue polylines in Fig. 5b); see Step ①.

---

<sup>1</sup>Interactive animations of some provinces are available at <http://www1.pub.informatik.uni-wuerzburg.de/pub/data/agile2016/>. (We recommend opening the link with Google Chrome).



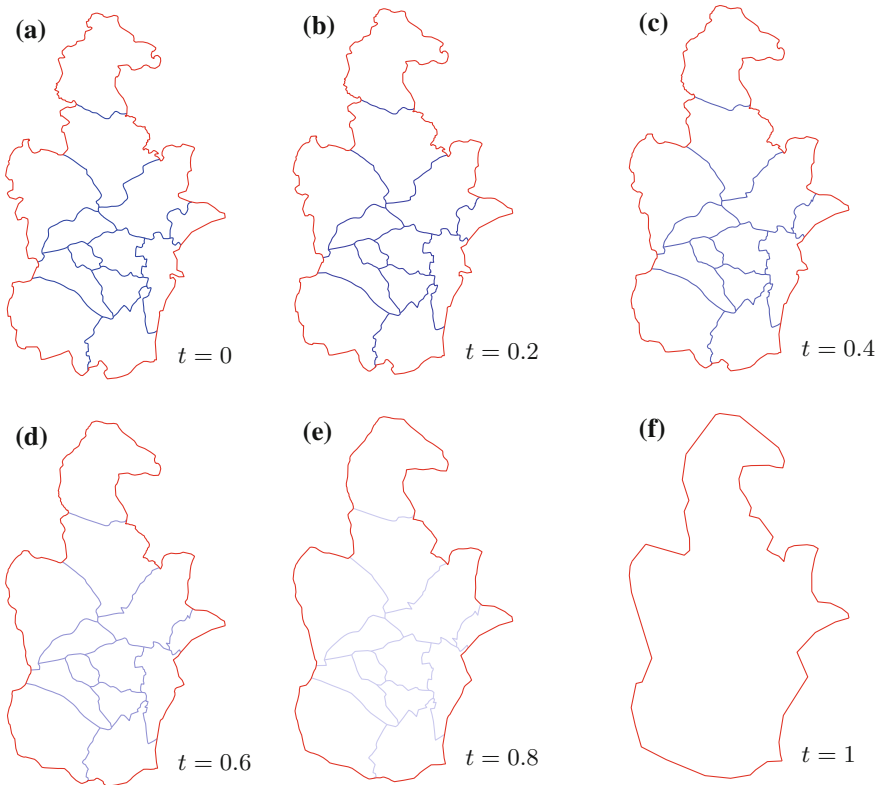
**Fig. 5** Case study on administrative boundaries of Mainland China. The numbers in the circles indicate the step orders, analogously to Fig. 1. For the sake of legibility, however, we do not display the compatible triangulations. Continuous generalization is achieved by morphing between (b) and (g). The blue boundaries in (g) are faded out in the transition from  $t = 0$  (in b) and  $t = 1$  (in g)

In Step ② we import the red polylines on  $M_-$  into Fig. 5f. Then we use the algorithm OPTCOR-S to determine corresponding points between the red polylines on  $M_+$  and  $M_-$ . We construct compatible triangulations for the polygons bounded by the red polylines of Fig. 5b and f so that we can transform the blue polylines on  $M_+$ ; the transformed polylines are the blue ones in Fig. 5f. During the determination of corresponding points, we used a fact that the 90 red polylines (with 55,533 vertices) on  $M_+$  and 90 red polylines (with 7,527 vertices) on  $M_-$  share many vertices ( $M_-$  may be generalized from  $M_+$ ). We split the red polylines of  $M_+$  and  $M_-$  into many subpolylines at the shared vertices so that we can compute corresponding points for each pair of subpolylines, which makes the computation much faster. Using a look-back parameter 145, the determination takes 266 s with cost  $\sum \delta(F, G) = 125,050 \text{ km}^2$ , where the look-back parameter 145 is the smallest value that yields the optimum result in the sense of the dynamic programming algorithm. The construction of the compatible triangulations takes 168 s. There is no conflict for the transformed polylines in the whole data set. However, a flaw is that there are zigzags caused by the compatible-triangulation-based transformation, e.g., the one shown in the enlarged figure next to Fig. 5f. We also tested the rubber-sheeting method of Doytsher et al. (2001), which, unfortunately, introduced 39 crossings.

In Step ③, we simplify the polylines, transformed by the compatible triangulations, to the blue polylines on  $M_-$  using the Douglas–Peucker algorithm, which takes 29 s. This simplification caused 8 crossings and 2 overlaps. We corrected the 10 conflicts manually. Note that we can avoid these conflicts by implementing a topologically consistent line simplification method, for example, the algorithm of Saalfeld (1999). In Step ④, we use OPTCOR-S again to determine corresponding points between the 5,819 blue polylines (with 438,092 vertices) on  $M_+$  and 5,819 blue polylines (with 58,105 vertices) on  $M_-$ . This time there are no shared vertices, and the computation takes about 16.5 h with look-back parameter 203, where this parameter is required by a certain pair of corresponding polylines to guarantee  $k \geq n_F/n_G$  (see Sect. 2.2). The cost of the resulting correspondence is  $\sum \delta(F, G) = 477,185 \text{ km}^2$ .

We morph from counties to provinces using straight line trajectories; see Step ⑤. We show the results of our continuous generalization of Tianjin Province in Fig. 6. Exporting the data set of Mainland China with 5,909 polylines (496,106 vertices) at any intermediate scale takes 45 s, mainly due to the slow creation of features in ArcGIS Engine. Unfortunately, this morphing causes conflicts in the intermediate-scale maps; two examples are shown in the enlarged figures next to Fig. 5c. For an instance, there are in total 41 crossings on the intermediate-scale map of Mainland China when  $t = 0.5$ . To avoid these crossings, we can use an algorithm which guarantees topological consistency during morphing, e.g., the algorithm of Surazhsky and Gotsman (2001).

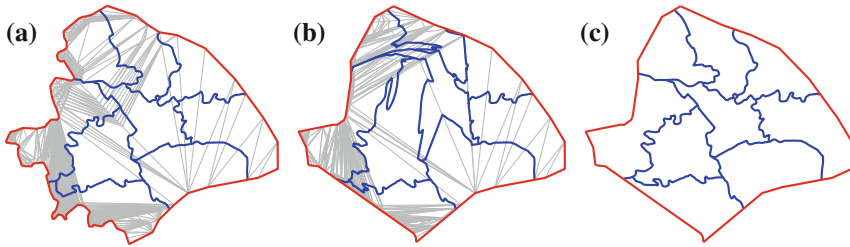




**Fig. 6** The continuous generalization of Tianjin province

## 4 Conclusions

In this paper, we have shown that rubber-sheeting, a popular method for data matching, can yield topologically inconsistent results. Therefore, we proposed a method to generalize continuously administrative boundaries based on compatible triangulations, which apparently have not been used in GIScience before (except by hand Fuse and Shimizu 2004). We have used compatible triangulations to transform unmatched polylines and managed to achieve topologically consistent. Although computing correspondences is slow, the computed result supports real-time interactions (e.g. zooming). By comparing our compatible-triangulation-based method to the rubber-sheeting method, we found that our method tends to yield results with larger distortions. An extreme instance is shown in Fig. 7. To decrease the amount of distortion, one could try constructing compatible triangulations that use the maximum number



**Fig. 7** A comparison of the compatible-triangulation-based method and the rubber-sheeting method for transforming the *blue* polylines on  $M_+$ , using the data sets of Shanghai as an instance. **a**  $M_+$  and the compatible triangulations of *red* polylines on  $M_+$ . **b**  $M_+$ , the compatible triangulations of *red* polylines on  $M_+$ , and the transformed polylines (*blue*) by the compatible-triangulation-based method. **c**  $M_+$  and the transformed polylines (*blue*) by the rubber-sheeting method of Doytsher et al. (2001)

of chords common to two triangulations. To that end, we could extend the dynamic programming algorithm mentioned by Aronov et al. (1993), [p. 34, bottom]. Whether this actually yields transformation results of higher quality is a question that requires further research.

Moreover, our current implementation of constructing compatible triangulations cannot deal with holes on a smaller-scale map  $M_-$ . Babikov et al. (1997) suggested a solution for this.

We used the Douglas–Peucker algorithm to simplify the transformed polylines. As expected, this led to some topological conflicts. To solve this problem, we need a topologically consistent algorithm, e.g., Saalfeld’s variant of the Douglas–Peucker algorithm (Saalfeld 1999). In the morphing process, we have used straight-line trajectories to interpolate between corresponding points. Again, these have introduced crossings. In order to guarantee topological consistency in the morphing process, we can use a compatible-triangulation-based algorithm to define the interpolation trajectories, e.g., the algorithm of Surazhsky and Gotsman (2001). With these two replacements, our workflow can generalize two-level hierarchical subdivisions (such as administrative boundaries) in a continuous and topologically consistent way.

**Acknowledgments** The authors thank Thomas C. van Dijk for proofreading an earlier version of this paper and the China Scholarship Council (CSC) for (partly) supporting this work.

## References

- Aronov B, Seidel R, Souvaine DL (1993) On compatible triangulations of simple polygons. *Comput Geom* 3:27–35
- Babikov M, Souvaine DL, Wenger R (1997) Constructing piecewise linear homeomorphisms of polygons with holes. In: *Proceedings of 9th Canadian Conference Computing Geometry (CCCG’97)*, pp 6–10

- Bereg S (2005) An approximate morphing between polylines. *Int J Comput Geom Appl* 15(2):193–208
- Bronstein IN, Semendjajew KA, Musiol G, Mühlig H (2001) *Taschenbuch der Mathematik*, 5th edn. Wissenschaftlicher Verlag Harri Deutsch
- Cecconi A (2003) Integration of cartographic generalization and multi-scale databases for enhanced web mapping. Ph.D. thesis, Universität Zürich
- Danciger J, Devadoss SL, Mugno J, Sheehy D, Ward R (2009) Shape deformation in continuous map generalization. *Geoinformatica* 13:203–221
- de Berg M, Cheong O, Kreveld MV, Overmars M (2008) *Computational geometry: algorithms and applications*, 3rd edn. Springer-Verlag TELOS
- Douglas DH, Peucker TK (1973) Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica* 10(2):112–122
- Doytsher Y, Filin S, Ezra E (2001) Transformation of datasets in a linear-based map conflation framework. *Surv Land Inf Syst* 61(3):159–169
- Fuse T, Shimizu E (2004) Visualizing the landscape of old-time Tokyo (Edo city). In: Gruen A, Murai S, Fuse T, Remondino F (eds) *The international archives of the photogrammetry, remote sensing and spatial information sciences*
- Girres JF, Touya G (2014) Cartographic generalisation aware of multiple representations. In: Duckham M, Stewart K, Pebesma E (eds) *Proceedings of 8th international Conference on Geographic Information Science—poster session*
- Gotsman C, Surazhsky V (2001) Guaranteed intersection-free polygon morphing. *Comput Graph* 25(1):67–75
- Hampe M, Sester M, Harrie L (2004) Multiple representation databases to support visualization on mobile devices. In: *Proceedings of ISPRS congress, volume XXXV–B4: IV of international archives of photogrammetry, remote sensing and spatial information sciences*, pp 135–140
- Hauert JH (2005) Link based conflation of geographic datasets. In: *Proceedings of 8th ICA workshop generalisation and multiple representation*
- Hershberger J, Snoeyink J (1992) Speeding up the Douglas–Peucker line-simplification algorithm. In: *Proceedings of 5th international symposium on spatial data handling*, pp 134–143
- Mustière S, Devogele T (2008) Matching networks with different levels of detail. *GeoInformatica* 12(4):435–453
- Nöllenburg M, Merrick D, Wolff A, Benkert M (2008) Morphing polylines: a step towards continuous generalization. *Comput Environ Urban Syst* 32(4):248–260
- Pantazis D, Karathanasis B, Kassoli M, Koukofikis A, Stratakis P (2009) Morphing techniques: towards new methods for raster based cartographic generalization. In: *Proceedings of 24th International Cartographic Conference (ICC'09)*
- Peng D, Hauert JH, Wolff A (2013) Morphing polylines based on least squares adjustment. In: *Proceedings of 16th ICA generalisation workshop (ICAGW'13)*, 10 pages
- Reilly DF, Inkpen K (2004) Map morphing: making sense of incongruent maps. In: Heidrich W, Balakrishnan R (eds) *Proceedings of graphics interface*, volume 62 of ACM international conference proceedings series. Canadian Human-Computer Communications Society, pp 231–238
- Saalfeld A (1985) A fast rubber-sheeting transformation using simplicial coordinates. *Am Cartogr* 12(2):169–173
- Saalfeld A (1999) Topologically consistent line simplification with the Douglas–Peucker algorithm. *Cartogr Geogr Inf Sci* 26(1):7–18
- Schneider T, Hormann K (2015) Smooth bijective maps between arbitrary planar polygons. *Comput Aided Geom Des* 35:243–354
- Sester M, Brenner C (2004) Continuous generalization for fast and smooth visualization on small displays. In: Altan O (ed) *Proceedings of 20th ISPRS congress, volume XXXV (Part B4) of international archives of the photogrammetry, remote sensing and spatial information sciences*, pp 1293–1298
- Surazhsky V, Gotsman C (2001) Controllable morphing of compatible planar triangulations. *ACM Trans Graph* 20(4):203–231

- van Kreveld M (2001) Smooth generalization for continuous zooming. In: Proceedings of 20th International Cartographic Conference (ICC'01)
- van Oosterom P (1995) The GAP-tree, an approach to on-the-Fly map generalization of an area partitioning. In: Mueller JC, Lagrange JP, Weibel R (eds) GIS and generalization, methodology and practice. Taylor & Francis, pp 120–132
- van Oosterom P (2005) Variable-scale topological datastructures suitable for progressive data transfer: the GAP-face tree and GAP-edge forest. *Cartogr Geogr Inf Sci* 32(4):331–346
- Whited B, Rossignac J (2011) Ball-Morph: definition, implementation, and comparative evaluation. *IEEE Trans Vis Comput Graph* 17(6):757–769