Lecture Notes
in Geoinformation and Cartography

LNG&C

Danny Vandenbroucke
Bénédicte Bucher
Joep Crompvoets *Editors*

Geographic Information Science at the Heart of Europe

Springer

Lecture Notes in Geoinformation and Cartography

For further volumes:
http://www.springer.com/series/7418

Danny Vandenbroucke
Bénédicte Bucher · Joep Crompvoets
Editors

Geographic Information Science at the Heart of Europe

Springer

Editors
Danny Vandenbroucke
Spatial Applications Division
Katholieke Universiteit Leuven
Heverlee
Vlaams-Brabant
Belgium

Joep Crompvoets
Public Management Institute
Katholieke Universiteit Leuven
Leuven
Vlaams-Brabant
Belgium

Bénédicte Bucher
COGIT
Institut Geographique National France
Saint-Mandé Cedex
Paris
France

ISSN 1863-2246 ISSN 1863-2351 (electronic)
ISBN 978-3-319-00614-7 ISBN 978-3-319-00615-4 (eBook)
DOI 10.1007/978-3-319-00615-4
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013938269

Printed on acid-free paper

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

Preface

Since 1998, the Association of Geographic Information Laboratories for Europe (AGILE) promotes academic teaching and research on GIS at the European level. Its annual conference reflects the variety of topics, disciplines and actors that make up the research scene on geographic information science in Europe and beyond. It has gradually become the leading GIScience conference in Europe.

For the seventh consecutive year, the AGILE conference full papers are edited in a book by Springer-Verlag. This year, 57 papers were submitted as full papers. The commitment of reviewers was quite impressive with sometimes up to 5 reviewers for one paper. In general 89.55 % of the assignments were performed in due time. We send our warm acknowledgements to the numerous reviewers who did a thorough review. All the reviews have been taken into account for the final selection of the very best papers. Reading and using the reviews made the work of the chairs easy and truly satisfactory. After the blind review process and a tough selection, 23 papers were selected that compose this book.

As the agenda for Europe 2020 is currently being set, this book demonstrates how geographic information science is at the heart of Europe. The paper contributions open perspectives for innovative services that will strengthen our European economy, that will inform citizens about their environment while preserving their privacy. Latest challenges of spatial data infrastructures are addressed such as the connection with the Web vocabularies or the representation of genealogy. User generated data (through social networks or through innovative camera and software) is also an important breakthrough in our domain. A trend to deal more and more with time, events, ancient data, and activities is noticeable this year as well.

As much as we congratulate authors for the quality of their work, we thank them for their contribution to the success of the AGILE conference and book series. In addition, we would like to thank the AGILE Council and members for their active support in making this volume a valuable publication in the field of Geoinformation Sciences and this conference a successful event. Special thanks go to Maribel Yasmina Santos, who made it possible that this book and previous AGILE Conference books are now indexed by the ISI Web of knowledge, which is necessary to make the contributions of our community more visible.

We would also like to thank our sponsors ESRI, Intergraph, Geosparc, the City of Leuven, the Province of Vlaams-Brabant, the National Geographical

Institute of Belgium, the Arenberg Doctoral School and Google for their kind contribution to this conference and Springer Publishers for their willingness to publish these contributions in their academic series Springer Lecture Notes in Geoinformation and Cartography.

March 2013

Danny Vandenbroucke
Bénédicte Bucher
Joep Crompvoets

Committees

Programme Committee

Programme Chair Danny Vandenbroucke
KU Leuven, SADL (Belgium)

Programme Co-Chair Bénédicte Bucher
IGN France, COGIT (France)

Programme Co-Chair Joep Crompvoets
KU Leuven, PMI (Belgium)

Local Organising Committee

Catharina Bamps, KU Leuven, SADL (Belgium)
Sofie Bruneel, KU Leuven, KU Leuven, E&ES (Belgium)
Philippe De Maeyer, Ghent University, CartoGIS (Belgium)
Jean-Paul Donnay, University of Liège, Geomatics Unit (Belgium)
Ludo Engelen, KU Leuven, SADL (Belgium)
Kristof Nevelsteen, KU Leuven, SADL (Belgium)
Thérèse Steenberghen, KU Leuven, SADL (Belgium) (Chair)
Jos Van Orshoven, KU Leuven, SADL (Belgium)

Scientific Committee

Trias Aditya, Gadjah Mada University (Indonesia)
Pragya Agarwal, Lancaster University (UK)
Rein Ahas, University of Tartu (Estonia)
Jagannath Aryal, University of Tasmania (Australia)

Yasushi Asami, University of Tokyo (Japan)
Peter Atkinson, University of Southampton (UK)
Fernando Bação, New University of Lisbon (Portugal)
Itzhak Benenson, Tel Aviv University (Israel)
Rohan Bennett, University of Twente (The Netherlands)
Lars Bernard, TU Dresden (Germany)
Michela Bertolotto, University College Dublin (Ireland)
Ralf Bill, Rostock University (Germany)
Roland Billen, University of Liège (Belgium)
Thomas Blaschke, University of Salzburg (Austria)
Lars Bodum, Aalborg University (Denmark)
Arnold Bregt, Wageningen University (The Netherlands)
Thomas Brinkhoff, Jade University Oldenburg (Germany)
Gilberto Camara, National Institute for Space Research (Brazil)
Tien-Yin Chou, Feng Chia University (Chinese Taipei)
Nicholas Chrisman, University of Laval (Canada)
Christophe Claramunt, Naval Academy Research Institute (France)
Arzu Coeltekin, University of Zurich (Switzerland)
Serena Coetzee, University of Pretoria (South Africa)
David Coleman, University of New Brunswick (Canada)
Lex Coomber, University of Leicester (UK)
Oscar Corcho, Universidad Politécnica de Madrid (Spain)
Helen Couclelis, University of California (USA)
Max Craglia, European Commission-Joint Research Centre (Italy)
Arie Croitoru, University of Alberta (Canada)
Rolf de By, University of Twente (The Netherlands)
Philippe De Maeyer, Ghent University (Belgium)
Michel Deshayes, IRSTEA (France)
Laura Diaz, Universitat Jaume I of Castellón (Spain)
Jürgen Döllner, University of Potsdam (Germany)
Jean-Paul Donnay, University of Liège (Belgium)
Matt Duckham, The University of Melbourne (Australia)
Sara Fabrikant, University of Zurich (Switzerland)
Peter Fisher, University of Leicester (UK)
Anders Friis-Christensen, National Survey and Cadastre (Denmark)
Stan Geertman, Utrecht University (The Netherlands)
Jérôme Gensel, University of Grenoble (France)
Yola Georgiadou, University of Twente (The Netherlands)
Michael Gould, ESRI Inc. (USA)
Carlos Granell, European Commission-Joint Research Centre (Italy)
Henning Sten Hansen, Aalborg University (Denmark)
Lars Harrie, Lund University (Sweden)
Francis Harvey, University of Minnesota (USA)
Gerard Heuvelink, Wageningen University (The Netherlands)
Stephen Hirtle, University of Pittsburgh (USA)

Hartwig Hochmair, University of Florida (USA)
Joaquín Huerta, Universitat 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)
Derek Karssenberg, Utrecht University (The Netherlands)
Tomi Kauppinen, Aalto University (Finland)
Marinos Kavouras, National Technical University of Athens (Greece)
Karen Kemp, University of Southern California (USA)
Thomas Kolbe, TU Munich (Germany)
Menno-Jan Kraak, University of Twente (The Netherlands)
Antonio Krüger, Saarland University (Germany)
Werner Kuhn, University of Münster (Germany)
Lars Kulik, The University of Melbourne (Australia)
Barend Köbben, University of Twente (The Netherlands)
Patrick Laube, University of Zurich (Switzerland)
Robert Laurini, INSA-Lyon (France)
Steve Liang, University of Calgary (Canada)
Miguel Luaces, University of A Coruña (Spain)
Sandra Luque, IRSTEA (France)
Michael Lutz, European Commission-Joint Research Centre (Italy)
Hans-Gerd Maas, TU Dresden (Germany)
Stephan Mäs, TU Dresden (Germany)
Miguel A. Manso Callejo, Universidad Politécnica de Madrid (Spain)
Bela Markus, University of West Hungary (Hungary)
Ian Masser, University of Sheffield (UK)
Kevin McDougall, University of Southern Queensland (Australia)
Filipe Meneses, University of Minho (Portugal)
Peter Mooney, National University of Ireland Maynooth (Ireland)
Adriano Moreira, University of Minho (Portugal)
Beniamino Murgante, University of Basilicata (Italy)
Pedro Muro Medrano, University of Zaragoza (Spain)
Zorica Nedovic-Budic, University College Dublin (Ireland)
Javier Nogueras Iso, University of Zaragoza (Spain)
Toshihiro Osaragi, Tokyo Institute of Technology (Japan)
Volker Paelke, Institute of Geomatics–Castelldefels (Spain)
Marco Painho, New University of Lisbon (Portugal)
Victor Pascual, Instituto Cartgráfico Cataluña (Spain)
Dieter Pfoser, Institute for the Management of Information Systems, RC Athena (Greece)
Alenka Poplin, HafenCity University Hamburg (Germany)
Poulicos Prastacos, Institute of Applied and Computational Mathematics FORTH (Greece)
Florian Probst, SAP Research CEC Darmstadt (Germany)

Hardy Pundt, University of Applied Sciences Harz (Germany)
Ross Purves, University of Zurich (Switzerland)
Martin Raubal, ETH Zurich (Switzerland)
Tumasch Reichenbacher, University of Zurich (Switzerland)
Wolfgang Reinhardt, Universität der Bunderwehr Munich (Germany)
Femke Reitsma, University of Canterbury (New Zealand)
Carmen Reyes Guerrero, CentroGeo (Mexico)
Claus Rinner, Ryerson University (Canada)
Stéphane Roche, University of Laval (Canada)
Julio Rojas-Mora, University of Avignon (France)
Maribel Yasmina Santos, University of Minho (Portugal)
Tapani Sarjakoski, Finnish Geodetic Institute (Finland)
Sven Schade, European Environment Agency (Denmark)
Christoph Schlieder, University of Bamberg (Germany)
Monika Sester, Leibniz University Hannover (Germany)
Takeshi Shirabe, Royal Institute of Technology (Sweden)
Spiros Skiadopoulos, University of Peloponnesse (Greece)
Bettina Speckmann, TU Eindhoven (The Netherlands)
Thérèse Steenberghen, KU Leuven (Belgium)
Emmanuel Stefanakis, University of New Brunswick (Canada)
Jantien Stoter, Delft University of Technology (The Netherlands)
Josef Strobl, University of Salzburg (Austria)
Juan Suárez, Centre for Forest Resources and Management (UK)
Maguelonne Teisseire, IRSTEA (France)
Marius Thériault, University of Laval (Canada)
Fred Toppen, Utrecht University (The Netherlands)
David Tulloch, Rutgers University (USA)
Nico Van de Weghe, Ghent University (Belgium)
Marc van Kreveld, Utrecht University (The Netherlands)
Bastiaan van Loenen, Delft University of Technology (The Netherlands)
Peter van Oosterom, Delft University of Technology (The Netherlands)
Jos Van Orshoven, KU Leuven (Belgium)
Lluis Vicens, Universitat de Girona (Spain)
Luis M. Vilches Blazquez, Universidad Politécnica de Madrid (Spain)
Marlène Villanova-Oliver, Grenoble University (France)
Monica Wachowicz, University of New Brunswick (Canada)
Robert Weibel, University of Zurich (Switzerland)
Stephan Winter, The University of Melbourne (Australia)
Mike Worboys, University of Maine (USA)
Bisheng Yang, Wuhan University (China)
Javier Zarazaga Soria, University of Zaragoza (Spain)

Contents

Contributors

Amin Abdalla Vienna University of Technology, Vienna, Austria

Benjamin Adams University of California, Santa Barbara, USA

Natasha Alechina The University of Nottingham, Nottingham, UK

Paulo Anciaes London School of Economics, London, UK

Michael Andorfer Research Studio iSPACE, Salzburg, Austria

Gennady Andrienko Fraunhofer IAIS, Sankt Augustin, Germany

Natalia Andrienko Fraunhofer IAIS, Sankt Augustin, Germany

Alkyoni Baglatzi School of Rural and Surveying Engineering, NTUA, Athens, Greece

Peter Bak IBM Research, Haifa, Israel

Mohamed Bakillah Institute for GI-Science, Ruprecht-Karls-Universität, Heidelberg, Germany

Lars Bernard TU Dresden, Dresden, Germany

Cai Cai University of Bremen, Bremen, Germany

Eliseo Clementini University of L'Aquila, L'Aquila, Italy

Thomas Couronne Orange Labs, Paris, France

René de Groot Institute for Geoinformatics, University of Münster, Münster, Germany

Heshan Du The University of Nottingham, Nottingham, UK

Mathieu Fauvel University of Toulouse INP-ENSAT, Toulouse, France

Andrew U. Frank Vienna University of Technology, Vienna, Austria

Christian Freksa University of Bremen, Bremen, Germany

Lutz Frommberger University of Bremen, Bremen, Germany

Lydia Gerharz University of Münster, Münster, Germany

Benedikt Graeler University of Münster, Münster, Germany

Glen Hart Ordnance Survey of Great Britain, Southampton, UK

Christin Henzen TU Dresden, Dresden, Germany

Pierre-Alexis Herrault University of Toulouse INP-ENSAT, Toulouse, France

Mohammed Imaduddin Humayun Institute for Geoinformatics, University of Münster, Münster, Germany

Corina Iovan Orange Labs, Paris, France

Enrico Ippoliti University of L'Aquila, L'Aquila, Italy

Mike Jackson The University of Nottingham, Nottingham, UK

Krzysztof Janowicz University of California, Santa Barbara, USA

David Jonietz University of Augsburg, Augsburg, Germany

Carsten Keßler Institute for Geoinformatics, University of Münster, Münster, Germany

Werner Kuhn Institute for Geoinformatics, University of Münster, Münster, Germany

Yoav Lerman Tel Aviv University, Tel Aviv, Israel

Steve H. L. Liang Department of Geomatics Engineering, University of Calgary, Calgary, Canada

Thomas Liebig Fraunhofer IAIS, Sankt Augustin, Germany

Paegelow Martin University of Toulouse INP-ENSAT, Toulouse, France

Stephan Mäs TU Dresden, Dresden, Germany

Grant McKenzie University of California, Santa Barbara, USA

Manfred Mittlboeck Research Studio iSPACE, Salzburg, Canada

Anna Monreale University of Pisa, Pisa, Italy

Ana-Maria Olteanu Raimond Orange Labs, Saint-Mandé, France

Itzhak Omer Tel Aviv University, Tel Aviv, Israel

Rebecca Ong University of Pisa, Pisa, Italy

Toshihiro Osaragi Tokyo Institute of Technology, Tokyo, Japan

Edzer Pebesma University of Münster, Münster, Germany

Dino Pedreschi University of Pisa, Pisa, Italy

Francesca Pratesi University of Pisa, Pisa, Italy

Chiara Renso CNR, KDDLab, Pisa, Italy

Daniela Richter Institute of Photogrammetry and Remote Sensing, Karlsruhe Institute of Technology, Karlsruhe, Germany

Kai-Florian Richter Department of Infrastructure Engineering, The University of Melbourne, Melbourne, Australia

Salvatore Rinzivillo ISTI-CNR, Pisa, Italy

Roberto Rösler Fraunhofer IAIS, Sankt Augustin, Germany

Falko Schmid University of Bremen, Bremen, Germany

Johannes Scholz Research Studios iSPACE, Salzburg, Austria

Tobias Schreck University of Konstanz, Konstanz, Germany

Wolfgang Schuster University of Augsburg, Augsburg, Germany

Angela Schwering Institute for Geoinformatics, University of Münster, Münster, Germany

David Sheeren University of Toulouse INP-ENSAT, Toulouse, France

Zbigniew Smoreda Orange Labs, Paris, France

Christoph Stasch University of Münster, Münster, Germany

Sabine Timpf University of Augsburg, Augsburg, Germany

Sayaka Tsuda Tokyo Institute of Technology, Tokyo, Japan

Monica Wachowicz UNB Fredericton, Fredericton, Canada

Wendy Hui Wang Stevens Institute of Technology, Hoboken, Italy

Paul Weiser Vienna University of Technology, Vienna, Austria

Stephan Winter Department of Infrastructure Engineering, The University of Melbourne, Melbourne, Australia

Alexander Zipf Institute for GI-Science, Ruprecht-Karls-Universität, Heidelberg, Germany

Part I
User Generated Data, Social Network Data

What You See is What You Map: Geometry-Preserving Micro-Mapping for Smaller Geographic Objects with MAPIT

Falko Schmid, Lutz Frommberger, Cai Cai and Christian Freksa

Abstract Geographic information is increasingly contributed by volunteers via crowdsourcing platforms. However, most tools and methods require a high technical affinity of its users and a good understanding of geographic classification systems. These technological and educational barriers prevent casual users to contribute spatial data. In this chapter we present MAPIT, a method to acquire and contribute complex geographic data. We further introduce the concept of *micro-mapping*, the acquisition of geometrically correct geometric data of small geographic entities. MAPIT is a method for micro-mapping with smartphones with high geometric precision. We show that MAPIT is highly accurate and able to reconstruct the geometry of mapped entities correctly.Please check and confirm the author names and initials are correct.

1 Introduction

Geographic data is the backbone of all geo-spatial applications. However, the collection of geo-spatial information is a resource intense task, traditionally performed by educated specialists employed in companies or national mapping agencies. This practice has changed fundamentally during the last decade: spatial

F. Schmid (✉) · L. Frommberger · C. Cai · C. Freksa
International Lab for Local Capacity Building (Capacity Lab), University of Bremen, Enrique-Schmidt-Str. 5, Bremen 28359, Germany
e-mail: schmid@informatik.uni-bremen.de

L. Frommberger
e-mail: lutz@informatik.uni-bremen.de

C. Cai
e-mail: cai@informatik.uni-bremen.de

C. Freksa
e-mail: freksa@informatik.uni-bremen.de

D. Vandenbroucke et al. (eds.), *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography,
DOI: 10.1007/978-3-319-00615-4_1,

information is increasingly collected and provided volunteers, a phenomenon also known as Volunteered Geographic Information (VGI) (Goodchild 2007). As a result geographic data became openly available for people and services. With the contribution of geospatial data by volunteers, also the nature of data drastically changed: the contributors decided what relevant information is and how to describe it (Haklay 2010). This requires not only the development of flexible services and classification systems, but also tools to collect the data volunteers intend to provide.

There are many different sources and types of VGI data available on different platforms spanning from hiking trails, photos, place information, online sensor data, to rather classical map-data. Volunteerly generated map data typically is collected by means of recording GPS trajectories which then are algorithmically transformed into street information (Ramm et al. 2010; Biagioni and Eriksson 2012). Depending on the accuracy of the received signal, the information has to be manually verified, corrected, and attributed with semantic information like street type, name, etc. to be transformed into useful geo-data.

One prerequisite for VGI being successful is that gathering of geo-data requires little effort. If the workflow is overly complicated, people will not spend their spare time to contribute. Intuitive user interfaces become especially important if the tasks go beyond mapping point- or trajectory-based data, e.g., when mapping extended objects. This usually requires to physically traverse the geographic object to be mapped. Some approaches combine satellite images as baseline data; in this case only immediately visible and conveniently reachable and traversable objects above a certain size are considered. The effort to map small objects is usually too high, as each object to be mapped requires thorough inspection, revision, and attribution.

In this chapter, we introduce the concept of *micro-mapping*, that is, mapping small geometric features in the plane. Examples of such features are vegetable fields in the garden, graves on a graveyard, areas on archaeological sites, fish ponds, flower beds, urban furnitures, etc. Retrieving the geometry of objects of this size can be a cumbersome procedure, as the precision of GPS sensors is not sufficient to provide accurate position information. As a result, such objects are usually mapped as points, or provided with a standard geometric shape.

These workaround solutions are not satisfactory, as in many cases the exact geometry of the objects provides important information. Later generations might want to precisely identify archaeological digging sites, GIS applications require the computation of crop of small agricultural parcels, urban planners benefit from detailed information of urban entities, even on smaller scale like benches or flower beds.

When mapping small features, an easy workflow of the mapping procedure is of even greater importance, because we must expect to have to map many of those small objects. Under this condition, it is infeasible to spend a lot of work on every single object. MAPIT, the method we present in this chapter, reduces the mapping procedure to the simple steps of taking a photo, drawing the outline of the object to be mapped, and label the object by means of a convenient user interface. This allows for mapping many objects in a short time and with high geometric precision.

2 Related Work

During the last decade, the generation, distribution, and usage of geographic information has dramatically changed. With the availability of affordable geodetic equipment, such as GPS devices and smartphones, and the availability of web-based data sharing platforms, the collection of geographic data became a phenomenon known as Volunteered Geographic Information (VGI) (Goodchild 2007; Sui 2008). VGI is a crowdsourcing movement, thus the collection of data by volunteers all over the world cooperating via web based platforms. One of the largest and most prominent VGI projects is OpenStreetMap[1] (OSM). OSM allows everybody to contribute geographic data of any kind. What started as a project collecting street map data quickly developed into a complex topographic mapping project with a huge amount of all kinds of mapped spatial entities. So far there are basically three different ways to provide geographic data to VGI platforms:

- *Geotagging*: Geotagging denotes the annotation of any kind of media or information (pictures, facts, etc.) with geographic coordinates to express its place of creation or relevance (e.g. Elwood 2008; Luo et al. 2011). Projects like EpiCollect (Aanensen et al. 2009) use geotagged form data to collect information about animal disease distribution, street art locations, archaeological digging sites, etc. Geo-tagging is an appropriate way to involve amateurs to provide non-spatial data, or whenever exact geo-spatial classification of the recorded entities is not important. However, this approach is limited in accuracy and expressiveness because no geometric information is provided. It is not possible to describe the geometry or orientation of the respective entities. We need additional information if the entity is supposed to be rendered on a map, or if properties need to be analyzed.
- *GPS-Trajectory*: A common practice to record geo-spatial data is to physically walk around the entity to be mapped and to record the complete GPS trajectory or only fixes that are required to describe the geometry of the entity, e.g., (Ramm et al. 2010; Turner 2006). The data of the tracks can then be analyzed and fused to richer data sets describing street networks or any other spatial entity to be mapped (Biagioni and Eriksson 2012).
- *Satellite Imagery Annotation*: An alternative method to create geo-spatial data is to analyze satellite imagery by means of crowdsourcing. With this approach, instead of physically traversing entities, contributers manually extract entities and their geometries from satellite images, e.g., (Maisonneuve and Chopard 2012).

With close range photogrammetry, a technique so far not explored for VGI, it is possible to obtain geographic data from camera systems (Luhmann 2010). Methods of this field are applied in traffic accident reconstruction (Du et al. 2009;

[1] http://www.openstreetmap.org

Fraser et al. 2008), and architectural engineering, e.g., bridge measurement (Jiang et al. 2008), 3D building reconstruction (Asyraf et al. 2011).

Once the data is collected it needs to be classified according to the addressed geographic specification system, such as OSM, CityGML,[2] ATKIS,[3] or the OS MasterMap.[4] However, those specifications are complex systems to formally describe possible spatial entities. Due to the complexity, it is hard for non-experts to contribute data with correct annotation. Once data is classified incorrectly, it will not be detected by algorithms for analysis or rendering. As studies on quality of OSM data show, the collection of complex geo-data by amateurs requires appropriate mechanisms to ensure quality (Goodchild 2009; Haklay 2010). However, human computer interaction aspects or human spatial conceptualizations of space are not very well studied and addressed in VGI literature and practice so far (Jones and Weber 2012). One approach to support this process is to incorporate ontological reasoning in the classification process, e.g., (Brando et al. 2011; Schmid et al. 2012). In addition to the classification process, geo-spatial editors and workflows are typically complex and hard to use. Without training and experience it is hard to collect, classify, and contribute spatial data to a platform like OSM. These educative and usability barriers prevent potential casual contributers to provide even only small bits of information to VGI platforms. However, in many cases, the inclusion of people at a grassroots level is the only possibility to gather and map expert data from agriculture, seasonal phenomena, land use, soil quality, disaster impacts, etc. (see Frommberger et al. 2012, for example).

3 mapIT: A Micro-Mapping Approach

In this chapter we present mapIT, a new approach for capturing, classifying, and contributing geographic data for Volunteered Geographic Information (VGI) initiatives. The purpose of mapIT is what we call *micro-mapping*: recording geometrically correct data of small geographic entities. *Small* entities in the context of micro-mapping are objects which size is too meaningful such that they could be represented as a point, but small enough to easily fit to one camera image (Schmid et al. 2012).

The design of mapIT follows the idea of WYSIWYG[5] editors: collecting, editing and contributing geo-spatial data are no separate steps in mapIT, but integrated in one seamless workflow in which users can directly contribute geometric data from camera images of any entity in the surrounding environment.

[2] http://www.citygml.org/

[3] http://www.adv-online.de

[4] http://www.ordnancesurvey.co.uk/oswebsite/products/os-mastermap/index.html

[5] WYSIWYG: "What-You-See-Is-What-You-Get"

MAPIT is designed to be barrier-free, i.e., it only requires little general knowledge to be used and no education in geographical classification systems. With MAPIT contributors can collect and contribute geo-data in situ, thus while being present in the environment. However, it is also possible to classify and contribute the data at any point in time.

3.1 What You See is What You Map

MAPIT is developed to integrate visual data capture, intuitive classification, and contribution in a single process. The idea of MAPIT is to enable the mapping of entities within the current vista space of users: that is, users can map spatial objects and phenomena when they see them. This method has several advantages compared to the alternative methods of GPS-trajectory based annotation, satellite image annotation, and geo-tagging.

- *Advantages Compared to Geo-Tagging* Geotagged information can be images, tweets, lexical entries, etc. Points of interest (POIs) are a particular form of geotagged entities, as they are primarily created to be used in geographic information systems. With geotagged information it is not possible to describe complex geometry. With MAPIT it is possible to contribute complex geometry also for entities until now considered to be to small to map geographically and geometrically correct.
- *Advantages compared to GPS-Trajectory recording* In areas with outdated, no, or not sufficient coverage by satellite imagery, entities have to be captured by means of recording the GPS trajectories of users surrounding the entity. With MAPIT users do not have to traverse the outline of the object and record GPS trajectories, but only have to take a photo of them. This is especially beneficial in cases when entities are hard to reach, e.g., when they are located in only hardly accessible marsh land. Additionally, due to the relatively small size of entities we are aiming at, the GPS data is usually highly scattered and require manual reconstruction of the geometry.
- *Advantages Compared to Satellite Imagery Annotation* In the last years, satellite images became the one of the most important sources for geo-spatial data. Data captured by GPS is verified with satellite images, but satellite images are also used to extract spatial features directly from the photo. These methods are very powerful to align scattered data correctly to the shapes of entities, and to create data along their visual outlines. However, this method is sensible to coverage, quality, and the frequency of updates of the underlying image material. In some cases the entities to be mapped cannot be recognized on the images due to resolution problems, in some cases the objects are occluded by entities located above (e.g., a pond located in a forest), in some cases the entities are not yet covered due to outdated images, and in some cases the entities are only seasonal phenomena (like flooding areas in rainy seasons), or even invisible (such as contaminated soil).

In general, mapIT is a suitable method whenever smaller entities have to be mapped that have to be geometrically accurate, up-to-date, are visually occluded for satellites, are changing or seasonal features, or are even invisible such as contaminated sites or archaeological digging sites.

4 mapIT: Workflow and Technical Details

mapIT is a VGI method to enable also casual contributors to provide correct geospatial data. We set three requirements for our method: (a) the mapping process should not require any geographical expert knowledge (for example, about geographic classification), (b) it should not be required to type in names or numbers, and (c) it has to run on low-cost smartphones with usual sensory capabilities (that is, GPS, compass, camera, tilt sensors). Thus, we develop a camera and speech-recognition based mapping application for Android phones (see Fig. 1): The user only has to take a picture of the entity to map (Fig. 1a), trace its outline on the touchscreen (Fig. 1b), and finally speak the type of the object into the phone (Fig. 1c). After these intuitive and barrier-free steps, the entity can immediately be uploaded to a server and is ready for further processing and inspection (Fig. 1d). Geometry and location of the entity are derived from GPS signal, geometric projection of the finger-trace on the touchscreen, camera lens properties, and information of the tilt sensors.

4.1 In-Situ, Ex-Situ, Online, and Offline Functionality of mapIT

Some of the components of mapIT obviously require internet connectivity, e.g. uploading data to a server. Mapping with mapIT is designed to work in-situ and ex-situ, as well as online and offline. Pictures of spatial entities taken with mapIT can be used at any point later to extract spatial information from it. This also holds for the speech recognition component based on the Android speech-to-text component[6]; in-situ speech labeling is only possible when a connection to the internet is available. If there is no connection available, the user can either label manually, by entering the label with a keyboard or use the speech functionality later when being connected to the internet again. Every picture taken with mapIT can be used to extract as many entities as wanted.

[6] http://developer.android.com/reference/android/speech/tts/TextToSpeech.html

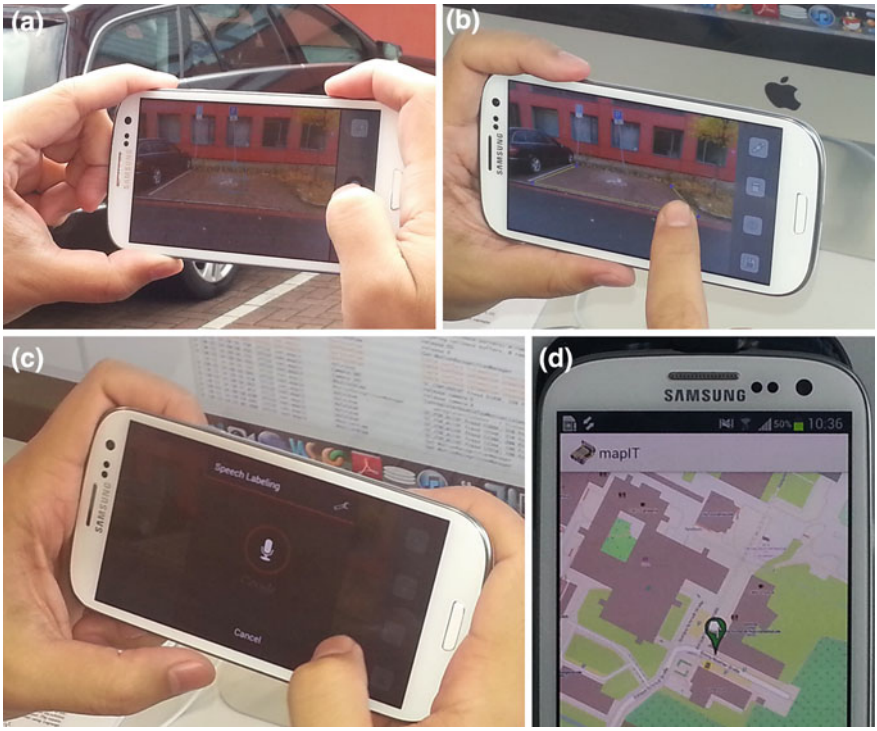


Fig. 1 Mapping requires little effort: The user just has to take a photo (**a**), outline the entity (**b**), annotate it via speech (**c**), upload it to a geo-server and check the entity on map (**d**)

4.2 Sensor Data Filtering

For geographic location calculation, a variety of sensor data is required, such as GPS and orientation sensor data. These sensor results are inevitably affected by hardware accuracy, environmental facts etc. Instead of simply accepting the raw sensor data, we adopt different methods of sensor fusion against different sensor types, in order to filter noise.

In particular, GPS signals inevitably suffer from noise and are known to be less reliable. To stabilize out readings, we record a series of GPS location estimates from the time the image capture starts until the shutter is pressed. From this series, we eliminate obvious outliers and smooth the remaining estimates by a weighted sum in favor of the latest estimates, following the assumption that later sensor readings provide a more reliable result. This considerably improved location estimates.

4.3 Projection to World Coordinates

In order to integrate the marked object into a geo-data set, it needs to be converted into a geographic location L. We now describe how we can retrieve the world coordinates of the marked object from the outline information of the camera image. The outline is a closed polygon, a set P^* of (x, y) coordinates in the image plane. In the first step, we reduce the number of points of P^* by applying the Douglas-Peucker algorithm (Douglas and Peucker 1973) for shape simplification and obtain a new, smaller set $P = \{P_0, P_1, \ldots, P_n\}$, $P_i = (x_i, y_i)$. This implements shape simplification directly on the smartphone.

The task to calculate the real world coordinates of the object outlined by P is an inverse perspective transformation (Foley et al. 1990, e.g.). For calculating the geographic location, distance and bearing angle from observer's location P_{camera} to every target point P_i is required. In the local coordinate system, $P_{\text{camera}} = (0, 0, h)$, with h being the height of the camera above the ground.[7]

Figure 2 depicts the projection from the image coordinate system to a local coordinate system, that is, from a point P_i in the image coordinate system to a point P_i' in the local coordinate system. The fact that the object is known to be in the xy-plane considerably reduces the complexity of the calculation.

The first intermediate step is to project P from the image coordinate system to a point $P_{3d} = (x_{3d}, y_{3d}, z_{3d})$ in a 3D coordinate system (defined by the blue axes in Fig. 2). For this we need the height *height* and width *width* of the camera image in pixels, the angle α from the phone's orientation sensor, and the camera lens parameter angles γ and δ that define the device's camera frustum. Then we get:

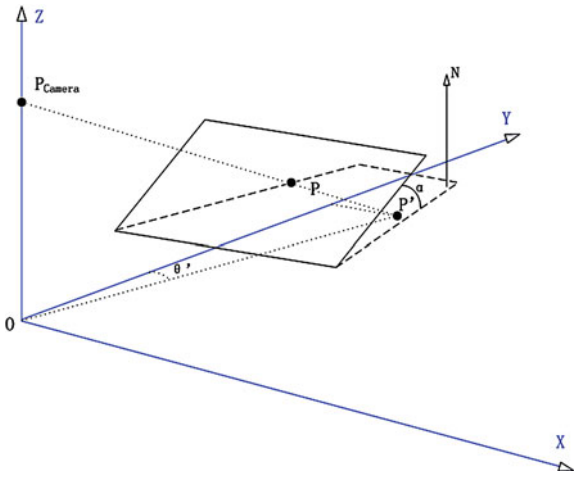


Fig. 2 Projection from image coordinate system to local coordinate system in the plane

[7] At the current state, the parameter h has to be set manually.

$$x_{3d} = x - \frac{\text{width}}{2} \tag{1}$$

$$y_{3d} = \frac{\text{width}}{2} \tan(\frac{\gamma}{2} + (\text{height} - y) \sin \alpha) \tag{2}$$

$$z_{3d} = (\text{height} - y) \cos \alpha \tag{3}$$

P'_i is the intersection point of the straight through P_{camera} and P_{3d} with the xy-plane. N is the normal vector of the xy-plane.

$$P' = P_{\text{camera}} + \frac{-P_{\text{camera}} \cdot N}{(P_{3d} - P_{\text{camera}}) \cdot N} (P_{3d} - P_{\text{camera}}) \tag{4}$$

With $P' = (x', y', z')$, we can finally retrieve θ':

$$\theta' = \arctan(\frac{x'}{y'}) \tag{5}$$

The distance d_i between P_{camera} and P'_i is

$$d_i = \frac{h \cdot \arctan \alpha'}{\cos \theta'} \tag{6}$$

with

$$\alpha' = \alpha - \frac{\delta}{2} + \delta | \frac{y' - y_{max}}{y_{max}} | \tag{7}$$

and y_{max} being the maximal y coordinate in the projected polygon.

With lat_o and lon_o being the latitude and longitude taken from the GPS estimate of the observer's position and R=6371004 being the equatorial radius of the earth in meter, we retrieve $L = (lat_i, lon_i)$ for every P'_i:

$$lat_i = \arcsin(\sin lat_o \cos \frac{d_i}{R} + \cos lat_o \sin \frac{d_i}{R} \cos \theta') \tag{8}$$

$$lon_t = lon_o + \text{ atan2 } (\sin \theta' \sin \frac{d_i}{R} \cos lat_o, \cos \frac{d_i}{R} - \sin lat_o \sin lat_t) \tag{9}$$

We repeat this procedure for every point in the object's outline and connect the points in the projection following the input sequence.

5 Evaluation

To show the feasibility of MAPIT we evaluated the accuracy of the collected geodata under everyday conditions. We collected defined real-world geometric data under controlled and varying conditions and compared the real world entity

Fig. 3 The satellite image used to verify the mapped parking lots. Each of them is 5×2.35 m

and the reconstructed entity with respect to accuracy of *angles*, *area*, and *perimeter*. We did not take the positional offset introduced by the GPS sensor into account, since potential errors are not introduced by the MAPIT approach but by the sensor unit itself. Any GPS-based approach depends on the accuracy of the sensor and cannot improve the physically limited result. However, we observed the usual offsets for the non-survey grade GPS units built in smartphones varying between near 0 m up to 10 m of positional displacement. With our filtering as outlined in Sect. 4.2, we obtained offsets of 1–4 m.

The aim of MAPIT is to enable micro-mapping, thus geometrically correct mapping of small spatial entities. For testing the precision we chose parking lots as reference objects, as they have a defined rectangular shape, are of a well-matching reference size of 5×2.35 m, and are visible on satellite imagery (see Fig. 3). All entities have been recorded using MAPIT running on a Samsung Galaxy Nexus 2 with Android 4.0. Our sample size was $p = 50$ measurements.

In order to evaluate MAPIT under realistic conditions we applied three different mapping variations:

- *Multiple perspectives*: we recorded 8 parking lots from 4 different perspectives in varying distances between 3 and 8 m in order to rule out influences on perspective adaptation of the method, see Fig. 4. This resulted in 32 individual measurements.
- *Multiple distances*: we recorded 4 parking lots from 3 different distances (5, 10, 15 m), however each from the same perspective (see Fig. 5). This resulted in 12 individual measurements.
- *Multiple entities from one photo*: we recorded 3 photos and mapped 2 distinct parking lots within each of them (see Fig. 6). This resulted in 6 individual measurements.

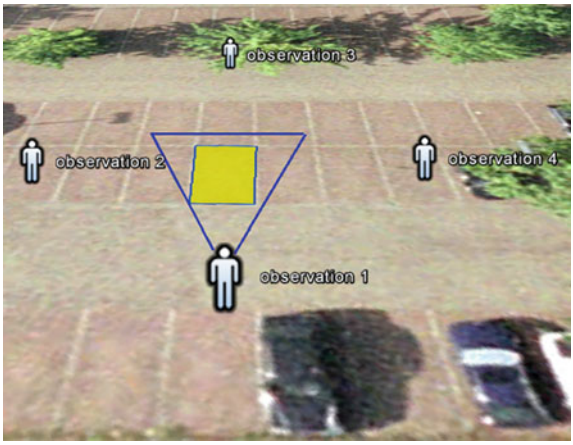


Fig. 4 Mapping an entity from 4 different perspectives

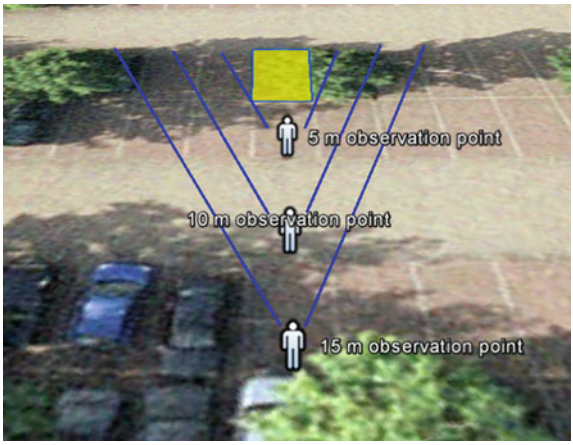


Fig. 5 Mapping an entity from three different distances (5, 10, 15 m)

After the recording of the photos with MAPIT (which includes the sensory information for geometric reconstruction) we manually set the four corner points of the parking lot on a large 20" screen with a computer mouse. This deviation from the original workflow (segmenting the entity directly on the touchscreen) was necessary to rule out errors introduced by inaccuracies of the touchscreen or during finger-based pointing. Defining the edges with greatest possible precision allows to analyze the results of the method without the influence of technical limitations of the interface. The four corner points where used as the input for the projection introduced in Sect. 4.3.

We then analyzed the resulting 50 reconstructed rectangles (32 multiple perspectives, 12 multiple distances, 4 multiple entities) with respect to their accuracy (each original parking lot is 5×2.35 m). We measured:

Fig. 6 Mapping of multiple objects from one photo

- the absolute angular deviation of each of the four inner angles of each reconstructed rectangular parking lot. Ideally each internal angle is exactly 90°. We compared 200 individual angles.
- the absolute areal deviation of each reconstructed rectangular parking lot. Ideally each entity has an area of 11.75 m^2. We compared 50 areas.
- the absolute side-length deviation of all sides of the reconstructed rectangular parking lot. Ideally each entity has two sides with 5, and two with 2.35 m length. We compared 200 individual side-lengths.

5.1 Results & Discussion

Table 1 shows the result of the evaluation for each mapping variation. It shows the mean deviation across all measurements of one mapping condition, and in the bottom row the overall mean across all mapping conditions. Although showing slight differences in the conditions, MAPIT performs uniformly well. The areal deviation across all three mapping variations is clearly below 4 %, in the multiple perspective condition even just 3.44 %. The same picture can be found in the side-length (perimeter) evaluation. In all three conditions the perimeter of the mapped entity is preserved to a very high degree: the summary deviation across all conditions is 4.33 %, in the multiple entities condition just 3.84 %. The angular accuracy of in all conditions is only slightly worse, however can still reconstruct

Table 1 Evaluation result

Variation	Area (%)	Perimeter (%)	Angle (%)
Multiple perspectives	3.44	4.46	5.99
Multiple distances	4.30	4.25	5.88
Multiple entities	4.90	3.84	6.26
Overall Deviation	3.82	4.33	5.99

angles to a very high precision; the deviations are only around 5.99 % across all conditions.

When we have a closer look at the distribution of the deviations across the conditions, we can get a better understanding of the composition of the results.

The chart for the angular error in Fig. 7a, shows that 50 % of the measurements only have a deviation between 0 to 2 %. The second peak is around a deviation of 10 %. After reviewing the data set, it turns out that the angles far from the observing points tend to have a greater deviation. In contrast, the angle of vertex turns out the better outcomes with 2 % in error. The chart for the side-length error in Fig. 7b shows a monotonly declining distribution of the deviation with about 70% of the measurements deviating below 5 %. A slightly different picture shows the chart for areal deviation in Fig. 7c.

Figure 8 shows visual results of the numbers presented above. All originally mapped parking lots are outlined by yellow line and filled with dots. The green polygons are the results of the mapping and geometric reconstruction with MAPIT. The geometry of the entities is very well reconstructed. We can observe a slight positional offset due to the GPS accuracy. However, even with the entities shifted from the original locations, our approach is capable of reconstructing complex topologies, such as neighboring parking lots on a parking ground.

6 Application Example

In this section we want to point to an application example where the approach shown in this chapter becomes a useful tool. We refer to a project we are working on in rural Laos where we are working with an educational program by the Lao government to enable poverty reduction work in the villages. Within the scope of this project, the problem of having an insufficient amount of protein in the daily food supply was prototypically tackled by installing small fishponds in the backyards of villager's houses to ensure an additional protein source (see Fig. 9). This idea turned out to be a successful and was quickly adopted by other villagers and across villages.

A major task in such kind of development work is monitoring the success and the impact of actions taken. In this case, this would mean to monitor the how the ponds spread over the area over time, their number, and the total amount of protein they can supply. To determine the latter number, it is essential to know the size of

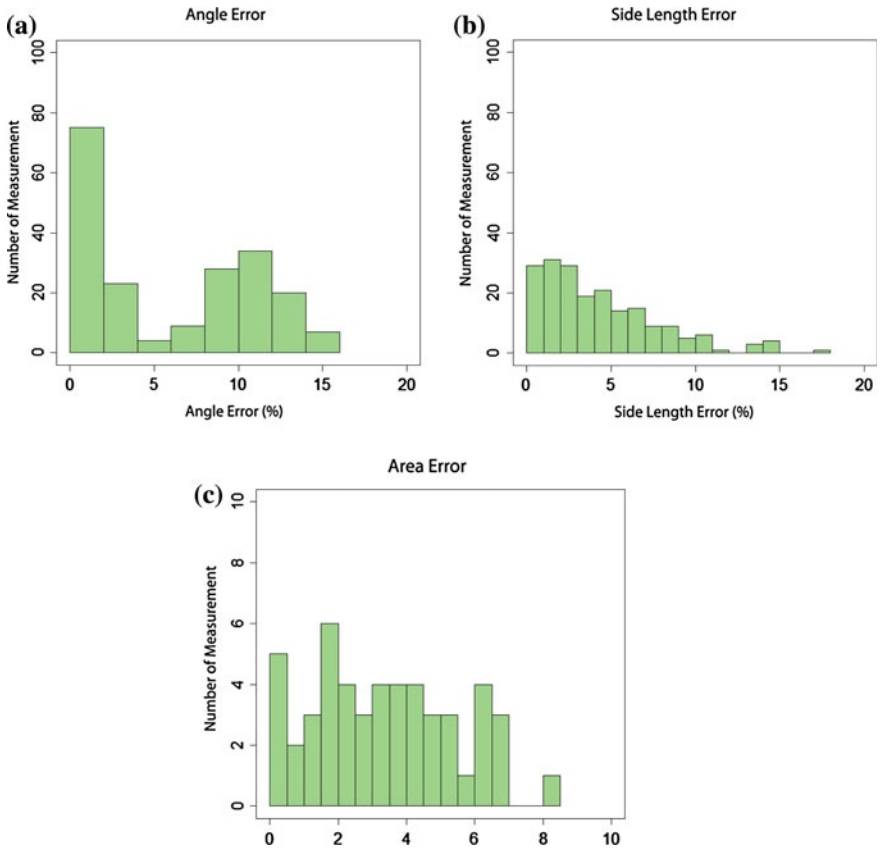


Fig. 7 The evaluation results: absolute internal angle deviation (**a**), absolute side-length (perimeter) deviation (**b**), absolute areal deviation (**c**)

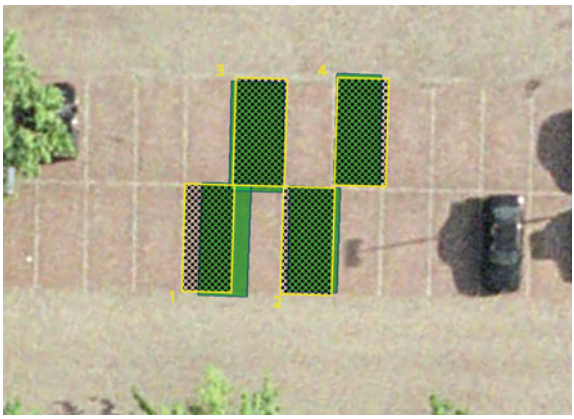


Fig. 8 This figure illustrates the reconstruction of four parking lots: the areas with *yellow outlines* and *filled with yellow dots* are the original parking lots to be mapped, the *green areas* are the same lots recorded and reconstructed with MAPIT. Note the accuracy of the geometric reconstruction. Only slight positional offsets are recognizable due to GPS filtering

Fig. 9 Backyard ponds in a village in rural Laos

the ponds in order to estimate how many fish can breed there. Thus, the application presented in this chapter can be a great help to perform this kind of monitoring. It runs on any low-cost Android smartphone, such that it is not expensive to equip local stakeholders with the needed technology. The mapping procedure is simple and intuitive and can be performed by laymen. People working there can assess the whole development of the pond project by simply going around, taking pictures, drawing the outlines, and label them as ponds. This data then can be aggregated on remote servers, visualized on dynamic maps, and protein supply can be estimated by determining the overall size of ponds in an area.

7 Conclusions & Future Work

Geo-spatial information is the basis for manifold applications in industry, development, and research. In contrast to past decades, the acquisition, availability, and usage of geographic data is anchored in a broad global movement of volunteers. Everybody can contribute the kind of data required for particular usage. However, until now there exists no easy-to-use method to record and contribute small geographic features with full geometric information. Due to the high effort required to map small entities, they have largely been ignored in the data collection process although there exist a large number of use cases for it.

In this chapter we introduced the concept of *micro-mapping*, the geometric correct acquisition of small spatial entities. We developed MAPIT, a method for rapid, barrier-free acquisition and contribution of small spatial entities with full geometric information. MAPIT is based on everyday smartphone technology and applies inverse perspective transformation to project coordinates of a photo to geographic space. We showed that the results of MAPIT are highly accurate, and we could reconstruct the original geometries of sample entities with high precision. The angular deviation between original and reconstructed entity is only 5.99 %, the side-length error 4.33 %, and the areal deviation is only 3.82 % between original and reconstructed entity.

MAPIT is designed to facilitate barrier-free contribution of geo-spatial data. It does not require more hardware than an ordinary smartphone and provides an easy workflow that allows acquisition of geo-data by non-experts. This can make mapIT a valuable tool in low-resource settings and facilitates the exploitation of geographical information in larger contexts that actually are unable to benefit from it—even in fields that usually do not much rely on technology, such as in agricultural development.

In order to make data processable it is necessary to classify it correctly. However geographic classification systems are complex and still hard to use for uneducated users. We plan to develop an ontological reasoning component to automatically translate natural language into a properly classified object.

Acknowledgments We gratefully acknowledge support by the University of Bremen, the German Research Foundation (DFG) within the SFB/TR8 Spatial Cognition and the International Research Training Group on Semantic Integration of Geospatial Information (GRK 1498), and support by the European Union Seventh Framework Programme—Marie Curie Actions, Initial Training Network GEOCROWD under grant agreement No. FP7-PEOPLE-2010-ITN-264994.

References

Aanensen D, Huntley D, Feil E, Spratt B et al (2009) Epicollect: linking smartphones to web applications for epidemiology, ecology and community data collection. PLoS One 4(9):e6968

Asyraf Hamdani M, Hashim K, Adnan R, Manan Samad A (2011) 3D images processing of structural building using digital close-range photogrammetric approach. In: IEEE 7th international colloquium on signal processing and its applications (CSPA), pp 318–321

Biagioni J, Eriksson J (2012) Map inference in the face of noise and disparity. In: Proceedings of the 20th international conference on advances in geographic information systems (ACM SIGSPATIAL GIS 2012), Redondo Beach, California

Brando C, Bucher B, Abadie N (2011) Specifications for user generated spatial content. Adv Geoinf Sci Chang World, pp 479–495

Douglas D, Peucker T (1973) Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. The Canadian Cartographer, 10(2):112–172

Du X, Jin X, Zhang X, Shen J, Hou X (2009) Geometry features measurement of traffic accident for reconstruction based on close-range photogrammetry. Adv Eng Soft 40(7):497–505

Elwood S (2008) Volunteered geographic information: future research directions motivated by critical, participatory, and feminist GIS. GeoJournal 72:173–183

Foley J, van Dam A, Feiner S, Hughes J, Phillips R (1993) Introduction to computer graphics. Addison-Wesley Professional

Fraser C, Cronk S, Hanley H (2008) Close-range photogrammetry in traffic incident management. In: Proceedings of XXI ISPRS congress commission V, WG V, Citeseer, vol 1, pp 125–128

Frommberger L, Schmid F, Cai C, Freksa C, Haddawy P (2012) Barrier-free micro-mapping for development and poverty reduction. Role of volunteered geographic information in advancing science: quality and credibility

Goodchild M (2007) Citizens as sensors: the world of volunteered geography. GeoJournal 69(4):211–221

Goodchild M (2009) Neogeography and the nature of geographic expertise. J Location Based Serv 3(2):82–96

Haklay M (2010) How good is volunteered geographical information? A comparative study of openstreetmap and ordnance survey datasets. Environ Plan B Plan Des 37(4):682–703
Jiang R, Jáuregui D, White K (2008) Close-range photogrammetry applications in bridge measurement: literature review. Measurement 41(8):823–834
Jones CE, Weber P (2012) Towards usability engineering for online editors of volunteered geographic information: a perspective on learnability. Trans GIS 16(4):523–544
Luhmann T (2010) Close range photogrammetry for industrial applications. ISPRS J Photogram Rem Sens 65(6):558–569
Luo J, Joshi D, Yu J, Gallagher A (2011) Geotagging in multimedia and computer vision—a survey. Multimedia Tools Appl 51:187–211
Maisonneuve N, Chopard B (2012) Crowdsourcing satellite imagery analysis: study of parallel and iterative models. Geograph Inf Sci, pp 116–131
Ramm F, Topf J, Chilton S (2010) OpenStreetMap—using and enhancing the free map of the world, UIT Cambridge
Schmid F, Cai C, Frommberger L (2012) A new micro-mapping method for rapid vgi-ing of small geographic features. In: Geographic information science: 7th international conference (GIScience, (2012) Columbus, Ohio, USA
Schmid F, Kutz O, Frommberger L, Kauppinen T, Cai C (2012) Intuitive and natural interfaces for geospatial data classification. In: Workshop on place-related knowledge acquisition research (P-KAR), Kloster Seeon, Germany
Sui D (2008) The wikification of GIS and its consequences: or Angelina Jolie's new tattoo and the future of GIS. Comput Environ Urban Sys 32(1):1–5
Turner AJ (2006) Introduction to neogeography. O'Reilly, Media

Trust as a Proxy Measure for the Quality of Volunteered Geographic Information in the Case of OpenStreetMap

Carsten Keßler and René Theodore Anton de Groot

Abstract High availability and diversity make Volunteered Geographic Information (VGI) an interesting source of information for an increasing number of use cases. Varying quality, however, is a concern often raised when it comes to using VGI in professional applications. Recent research directs towards the estimation of VGI quality through the notion of trust as a proxy measure. In this chapter, we investigate which indicators influence trust, focusing on inherent properties that do not require any comparison with a ground truth dataset. The indicators are tested on a sample dataset extracted from OpenStreetMap. High numbers of contributors, versions and confirmations are considered as positive indicators, while corrections and revisions are treated as indicators that have a negative influence on the development of feature trustworthiness. In order to evaluate the trust measure, its results have been compared to the results of a quality measure obtained from a field survey. The quality measure is based on thematic accuracy, topological consistency, and information completeness. To address information completeness as a criterion of data quality, the importance of individual tags for a given feature type was determined based on a method adopted from information retrieval. The results of the comparison between trust assessments and quality measure show significant support for the hypothesis that feature-level VGI data quality can be assessed using a trust model based on data provenance.

C. Keßler (✉) · R. T. A. de Groot
Institute for Geoinformatics, University of Münster, Münster, Germany
e-mail: carsten.kessler@uni-muenster.de

R. T. A. de Groot
e-mail: rta.de.groot@gmail.com

D. Vandenbroucke et al. (eds.), *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography,
DOI: 10.1007/978-3-319-00615-4_2,

1 Introduction

Professional geographic information is collected according to established standards, which allows the provider to guarantee levels of data quality, yet also results in high costs and sparse updates. Volunteered Geographic Information (VGI) has become an attractive alternative for use cases where professional geographic information is too expensive, not available for a theme,[1] or when timely updates are required (Goodchild 2007). In order to enable data consumers to get the best from both of these two data sources, a method is required that reliably filters VGI based on its quality. This approach would enable "cherry picking" on the data consumer's side, enabling them to identify high-quality features and leave features aside that seem problematic. At the same time, such a method would support the mapping community in spotting features that might need improvement.

Previous research has mostly focused on the *overall* quality of VGI datasets, e.g., by analyzing the average positional accuracy based on a dataset with guaranteed data quality. Such an overall analysis of a dataset, however, does not allow for quality assessments of single features in the dataset. Moreover, it requires access to the ground truth data that the data consumer is trying to do without. To overcome these problems, a model to assess the quality of VGI based on its provenance has been proposed (Keßler et al. 2011a, b; Mooney and Corcoran 2012b), where each feature's editing history is analyzed in order to assess its quality. These quality assessments are also referred to as *informational trust* (Bishr and Janowicz 2010) to denote the degree to which a data consumer can trust the information about the feature. Trust is a central principle in (real-world) social networks and has been shown to play a central role in online communities (Golbeck 2005). The trust analysis is based on intuitive notions such as the *many eyes principle* (Haklay et al. 2010), where the quality is more likely to be higher if more people have worked on a feature, as well as patterns of revisions such as vandalism or *edit wars* that indicate quality problems. In previous work, we have proposed a provenance vocabulary to annotate edits and enable the computation of a trust measure (Keßler et al. 2011a).

This chapter presents a practical application of this trust-based approach, along with an evaluation on a set of features selected from the OpenStreetMap (OSM) dataset for the city of Münster, Germany. The hypothesis for our work is that parameters derived from OSM provenance data determine a feature's trustworthiness, which is an indicator for data quality and therefore, trustworthiness and data quality are correlated. We test this hypothesis by carrying out two parallel assessments: one based on trust indicators derived from the OpenStreetMap history, and the other one based on high-quality reference data collected in a field survey. The outcome of both datasets is then tested for statistical correlation.

The remainder of the chapter is organized as follows: The Sect. 2 provides an overview of related work. Section 3 describes the study area. Section 4 analyzes

[1] See http://wheelmap.org, for example.

the trustworthiness of the features in the test dataset, followed by an analysis of their data quality based on a field survey in Sect. 5. Section 6 provides a comparison and statistical analysis of the correlation between the results of the two methods, followed by concluding remarks in Sect. 7.

2 Related Work

At the time of writing of this chapter, the OpenStreetMap community consisted of more than a million mappers who had generated more than 1.7 billion nodes, more than 160 million ways, and more than 1.7 million relations between them.[2] With the growing number of applications building on OSM, data quality has become an issue that has already been addressed in a number of articles. Haklay (2010) provides one of the first comprehensive studies, comparing OpenStreetMap data in the UK with government data provided by Ordnance Survey. Focusing on motorway objects, he found that about 80 % of the features from both classes overlap. Neis et al. (2012) present a similar study for the German street network by comparison with commercial data provided by TomTom and find that the overlap is at about 91 % (as of June 2011). Zielstra and Zipf (2010) show that the OpenStreetMap dataset converges towards commercial datasets in terms of completeness. A statistical comparison of junction points in OSM and commercial data from TeleAtlas by Helbich et al. (2010) also supports the statement that OSM presents a suitable alternative for commercial data. Koukoletsos et al. (2012) introduce an automated feature matching method for comparative studies focusing on positional accuracy such as the ones mentioned above. While the results for positional accuracy in OpenStreetMap are largely positive, a comprehensive study by Mooney and Corcoran (2012c) shows that most of the quality problems seem to lie in the thematic data, i.e., in the tags assigned to the features.

The International Organization for Standardization (ISO) defines data quality of geographic information as *the difference between the dataset and a universe of discourse* (International Organization for Standardization 2002; Jakobsson and Giversen 2009), where the universe of discourse is the real world view defined by a product specification or user requirements. The ISO 19113 standard lists five data quality elements: completeness and consistency, as well as positional, temporal, and thematic accuracy. While these elements are generally provided as metadata for professional geographic information to allow an informed judgment of the dataset's quality, they are missing both for OpenStreetMap and VGI in general.

In the absence of quality metadata for VGI, trust has been proposed as a proxy measure for data quality (Bishr and Kuhn 2007). Trust can be defined as "a bet about the future contingent actions of others" (Sztompka 1999, p. 25) and is closely related to reputation, defined as the subjective perception of trustworthiness

[2] See http://www.openstreetmap.org/stats/data_stats.html

inferred from information about the historical behavior of somebody/something (Mezzetti 2004). The concept of trust commonly refers to interpersonal trust, but can be extended to informational trust through people-object transitivity (Bishr and Janowicz 2010). Bishr and Mantelas (2008) have successfully applied this principle in a scenario dealing with urban growth data. Closely related to trust is the concept of *credibility*, which is slightly broader, as it comprises someone's expertise (covering aspects of accuracy, authority and competence) in addition to trustworthiness (with aspects of reputation, reliability and trust; Flanagin and Metzger 2008). Artz and Gil (2007) point out that provenance is a key factor for trust on the web. The approach on provenance we take in this chapter is *data-oriented*, as the focus is on the origins of specific data items, instead of the processes that generate the data (Simmhan et al. 2005).

3 Test Dataset

The OpenStreetMap history dump[3] contains all nodes, ways, and relations that were ever added to OSM, along with each of its revisions. From this file, we extracted the features within the area of interest for our study, namely Münster's old town (*Altstadt*), shown in Fig. 1. The map is based on a shape file extracted from OSM for easier processing in GIS environments.[4] The boundaries for the area of interest are based on the city of Münster's district boundaries, which contains the Altstadt area as an official administrative district.

This area of the city was selected because of its high density of points of interest for sightseeing, shopping, dining, et cetera. It is thus made sure that the input dataset for our analysis contains enough features with a long enough editing history to allow for a meaningful analysis. At the same time, the area is small enough to allow for a field survey that collects data of the features on-site to create a ground truth dataset. However, re-mapping all features in the area of interest to create a ground-truth dataset is clearly not feasible. We therefore selected a subset of the features in the area of interest based on the number of revisions that the respective feature has undergone. The selection was made based on the number of revisions, as our trust assessment will be based on the provenance of the feature. This approach thus ensures that the test dataset has a rich enough input in terms of feature history. We decided that including up to 100 features in the field survey is feasible. This criterion is met if we include features that have at least 6 versions, which applies to the 74 features highlighted in Fig. 2.

[3] See http://planet.openstreetmap.org/planet/full-history/

[4] See http://download.geofabrik.de/osm/europe/germany/nordrhein-westfalen/muenster.shp.zip

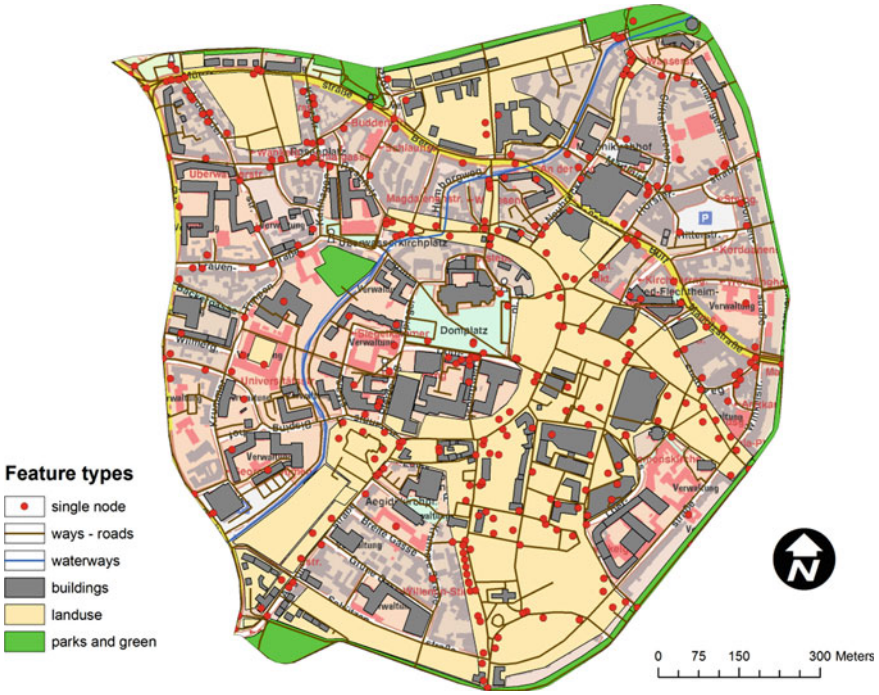


Fig. 1 Overview of Münster's old town based on the OSM data as of October 2011

4 Trust Assessment

This section introduces the trust assessment of the selected features. The parameters taken into account are introduced and the results of the assessment are discussed.

4.1 Parameters for Trustworthiness

First ideas on the parameters that influence trust assessments of features in OpenStreetMap have been discussed in Keßler et al. (2011a). However, not all proposed parameters can be applied to our test dataset, most notably user reputation. A measure for user reputation would require global knowledge of the OSM dataset to assess the user's experience (by counting the respective number of contributions) or even assessments of the quality of all features the user has been involved in. This is clearly not feasible for the small-scale test dataset used in this chapter, and is a complex research problem of its own. We are therefore taking the following parameters into account for our evaluation:

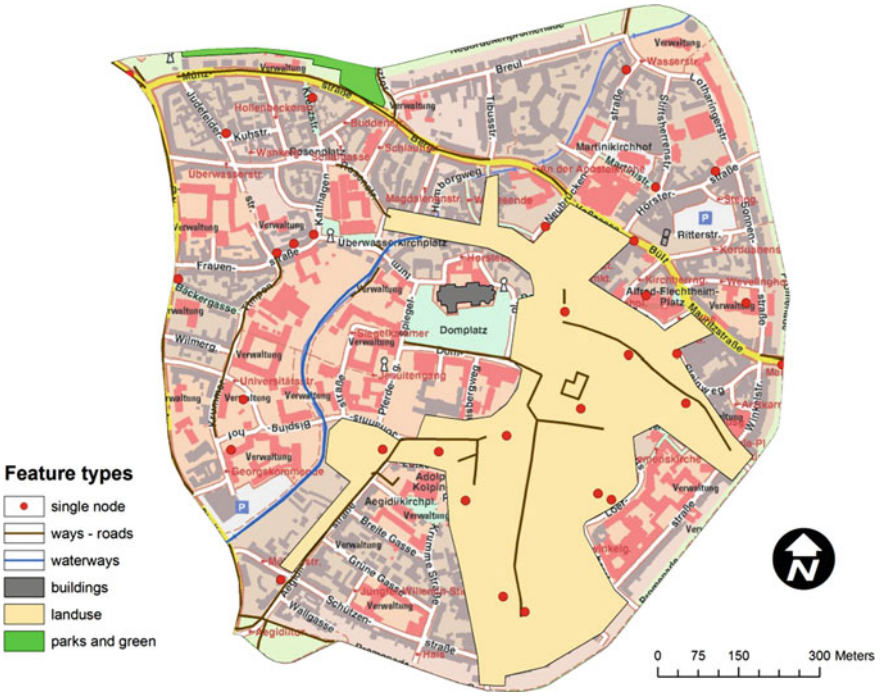


Fig. 2 Features with at least 6 versions highlighted

- **Versions**. As pointed out above, more versions essentially mean more provenance information to analyze. At the same time, they indicate that the feature has undergone a certain number of iterations with the goal to improve the feature's quality. We therefore assume that a higher version number adds to the feature's trustworthiness.
- **Users**. A higher number of users involved in creating a feature increases the trust measure, following the many eyes principle. Even though previous research found that the relation between number of users and data quality is not linear, we follow a linear approach here as the corresponding study by Haklay et al. (2010) has only taken positional accuracy into account and did not look at the quality of the thematic attributes.
- **Confirmations**. On top of the number of users that have been working directly on a feature, we also take *indirect* confirmations into account by looking at all revisions that have been made in the immediate vicinity of a feature *after* the last revision of a feature. The rationale is that it is very likely that someone who is editing features in a certain area also looks at features in the vicinity; therefore, we count all edits made within a 50 m buffer of a feature after its last revision as confirmations that increase trustworthiness.

- **Tag corrections**. Tags in OSM consist of a key and corresponding value.[5] Corrections occur when the value for a certain tag is changed, e.g., when the tag amenity = restaurant is changed to amenity = pub. We assume that this points to ambiguities in the feature classification and thus decreases trustworthiness.
- **Rollbacks**. A tag correction reverting a feature to its previous state is considered as a rollback and also decreases feature trustworthiness.

It is important to note that the parameters' influence on the trust measure as outlined above only covers the general case—there will always be exceptions to the rule. These exceptions, however, cannot be covered by generic statements, and can only be discovered by comparison with ground truth, which is not feasible in practice. We discuss in the following how well such a generic approach works.

4.2 Calculating Trust Assessments

For each of the parameters, we have created a classification into five equal intervals [1…5] that indicate a ranking from low to high trustworthiness. For those parameters with a positive influence on trustworthiness (numbers of versions, users, and confirmations, respectively), higher counts lead to a classification into a trustworthier interval. For the two parameters with a negative influence (corrections and rollbacks), *lower* counts lead to a classification into a trustworthier interval. As the combination of these parameters into a single trust assessment is still an open research question, we take a naïve approach here by assigning equal weights to all parameters. The overall results are then added up and classified into five equal intervals again for comparison with the results of the field survey.

Most features are almost equally distributed between classes two to four (65 out of 74), of which class four contains the highest number. Class one contains 8 features and class five only three. Overall, the trust measure suggests an estimation of the dataset as moderately trustworthy regarding its information quality, with only a low number of features reaching the highest level of trustworthiness. Figure 3 shows an overview of the trust assessments for the selected features.

5 Field Survey

In order to evaluate whether trust assessments bear results that realistically reflect feature quality, a reference dataset is required for comparison. This section describes how the reference dataset was collected in a field survey and how the data quality was rated.

[5] See http://wiki.openstreetmap.org/wiki/Tags for an overview of the tagging conventions the OpenStreetMap community has developed.

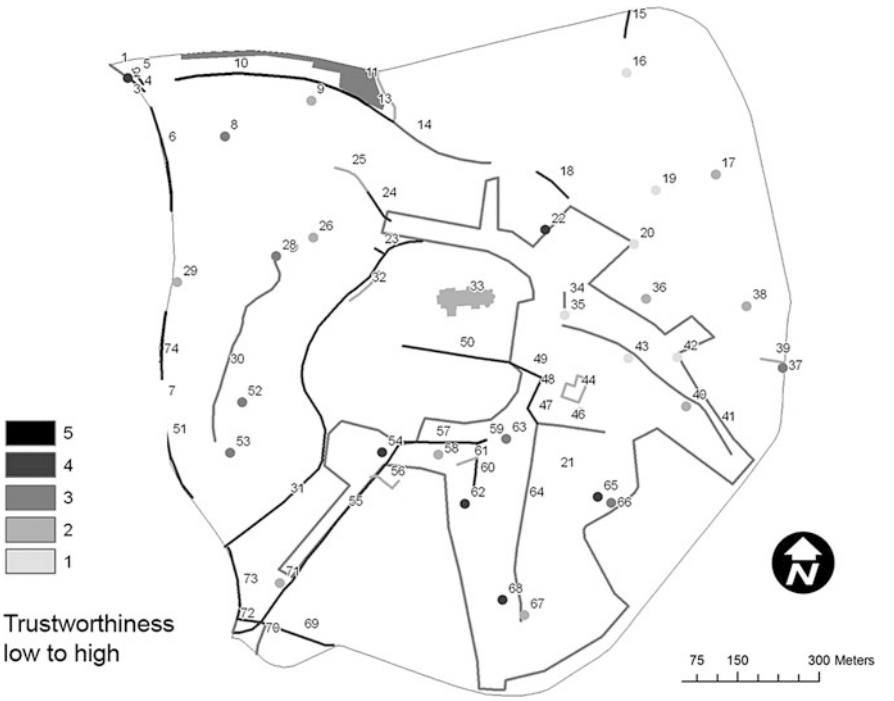


Fig. 3 Assessed trustworthiness of selected features based on numbers of versions, users, confirmations, corrections, and rollbacks

5.1 Thematic Accuracy and Topological Consistency

During the field survey, all selected features have been inspected on site, pictures were taken, and, if possible, people were asked to confirm the feature type in consideration to minimize bias. The following components were examined during this survey and are listed here in the order of higher to lesser importance regarding its influence on the quality of the thematic accuracy:

1. The correctness of the main tag: e.g. is this place a restaurant or a café?
2. The correctness of other tags that are described: for example, is the house number stated in OSM correct, or is the number of lanes of a street correct?
3. Is there any confusion or doubt about whether the description in OSM represents the feature in the right way:
   - Unclearness about the nature of the feature; lack of description when it is not obvious what the function is (e.g., `information=guidepost`: is the guidepost about street information or historical information?)
   - Doubt about the type within a main feature type (e.g., is it `highway=primary` or `highway=secondary`?).

- The feature pointed to could be part of the whole feature instead of the feature itself (e.g., the entrance of parking lot could be marked as the parking lot feature)

Based on these criteria, the features were divided into four classes. Class 1 represents features of which the main tag does not correspond to what has been found in the field. Class 2 is assigned to features where other tags are incorrect. Class 3 contains features that have a shortcoming as described in the third point. And lastly, class 4 is assigned to those features of which the available information is fully correct. Out of the 74 features, 6 features (∼8 %) were assigned class 1 (lowest thematic accuracy), 2 features (3 %) were assigned class 2, and 9 features (∼12 %) were assigned class 3. For the remaining 57 features (∼77 %), the available information is fully correct (class 4).

Positional accuracy in OpenStreetMap has already been extensively studied, as discussed in Sect. 2. For this reason, and for the lack of a straightforward method to create accurate positional information in our reference dataset, the selected features were tested for *topological* consistency. For all 74 features, it was checked whether they are qualitatively correctly positioned relative to the surrounding features. The results were positive throughout, except for one feature representing an information panel that was located on the wrong side of a street.

5.2 Information Completeness

The tagging guidelines documented in the OSM wiki provide recommendations which tags should be used for certain types of features, and describe how additional information such as opening hours can be added so that they can be automatically reused by applications building on the OSM data. However, due to the large number of potential feature types, the wiki does not provide lists of tags that should be provided for a certain type in order to provide complete information. We therefore had to come up with a method that determines importance of a specific tag for a given feature type. Based on the *term frequency—inverse distance frequency* measure (tf-idf) (Salton et al. 1975) that measures how characteristic a specific term is for a given document, we have defined an *inverse feature type frequency* (iff). It measures the general importance of a tag by dividing the total number of features in the dataset by the number of features with which this tag is associated. Taking the logarithm of this quotient generates values that indicate a higher relevance of the tag the closer it is to 0:

$$\mathrm{iff}(t) = \frac{\log|\mathrm{F}|}{|\{\mathrm{f}: \mathrm{t} \in \mathrm{f}\}|}$$

where |F| is he total number of features in the whole dataset and $|\{\mathrm{f}: \mathrm{t} \in \mathrm{f}\}|$ is the number of features in that set containing tag t. The *tag frequency—inverse feature type frequency* is then determined by:

$$\text{tf - iff}(t, f) = \text{tf}(t, f) * \text{iff}(t)$$

where tf is the tag frequency in the set of features that belong to the same feature type. The output of this calculation shows the relevance of a tag within a set of features of a certain feature type, also considering all the features in the whole dataset. Tags that are relatively unique to the feature type in consideration receive higher importance values than tags that are also commonly used to describe other feature types.

This approach turned out to be problematic for tags such as name that are relevant for a broad range of feature types. As they are not type-specific, their tf−iff measure is generally comparably low and would remove them from the list of important tags for a feature type. To solve this problem, we use the iff measure again, but now the denominators have a different meaning; $|F|$ is the total number of features within a feature type and $|\{f: t \in f\}|$ is the number of features with a certain tag within a feature type set. It turned out that this method filters out the most important and obvious tags per feature, even if they occur less frequently than in half of the features of a particular type. In a feature type like pub the tag smoking should be important. It occurred in less than half of all the pub features, but it is still incorporated in the set of 'obligatory tags' when determining its importance with the measure of general importance (iff). The values for both ways of importance determination were then normalized. We applied a threshold of 0.5 to identify the 'obligatory tags' for a given feature type.

In order to determine the information completeness for a specific feature, we checked each of the obligatory tags for presence in the feature and counted omitted tags. In order to stress the importance of tags that characterize a feature type, omissions of such tags (e.g., religion for churches) receive double weights.

5.3 Overall Quality

After preparing the individual quality element tests, linking them to the map IDs and classifying them into five classes (equal intervals), the outcomes of the three quality tests have been summed up and the summed values were reclassified again into five classes. 42 out of the 74 features (∼57 %) fully meet the quality requirements defined above. Generally, their theme is correctly described and their topological relations with respect to their surroundings are correct. The quality concerning information completeness is mixed; however, since each parameter of the quality assessment receives the same weight, the overall outcome is that many features are of high quality: 60 out of 74 features (∼81 %) are classified into the two highest quality classes. Figure 4 gives an overview of the observed data quality of the sample dataset.

6 Comparison and Results

This section compares the results of the trust assessment to those of the data quality observations collected in the field survey. We discuss the differences and provide a statistical correlation analysis.

6.1 General Observations

A visual comparison of the trust assessment shown in Fig. 3 and the observed data quality already shows that the features are generally deemed less trustworthy than indicated by the actual observed data quality shown in Fig. 4. A comparison including class difference is shown in Fig. 5. The visualization shows that the trust assessments generally underestimate the actual data quality: the mean quality of the features is at $\sim$4.2, whereas the mean trust value is at $\sim$2.8.

The trend across all features shown in Fig. 6, however, shows that the trust assessments generally follow the same pattern as the quality observations.

The salient high and low points of both measures revealed that low trust values were caused mostly by a low number of confirmations and the presence of tag

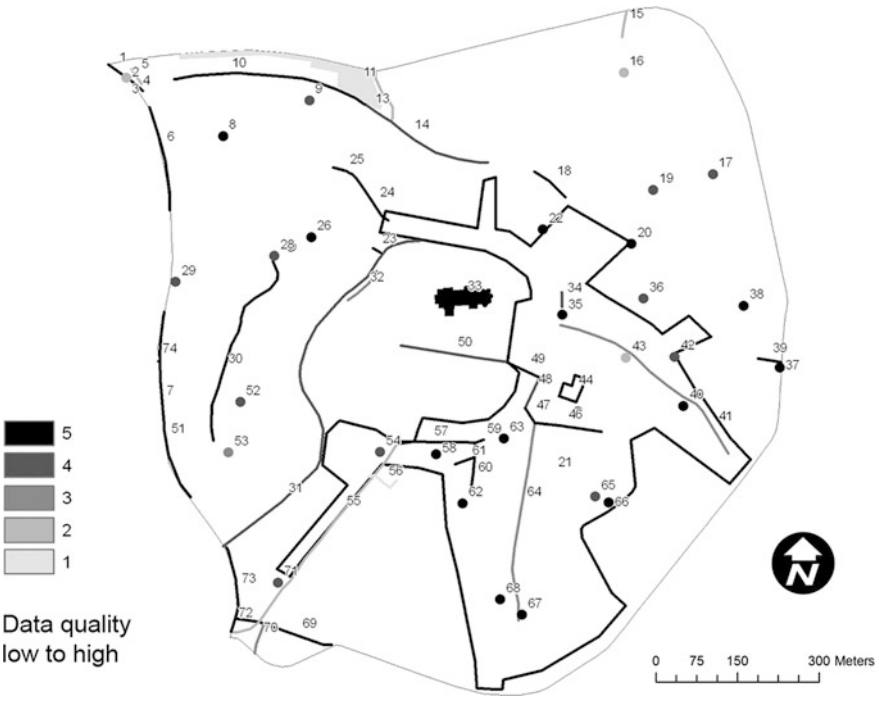


Fig. 4 Observed data quality of the sample dataset

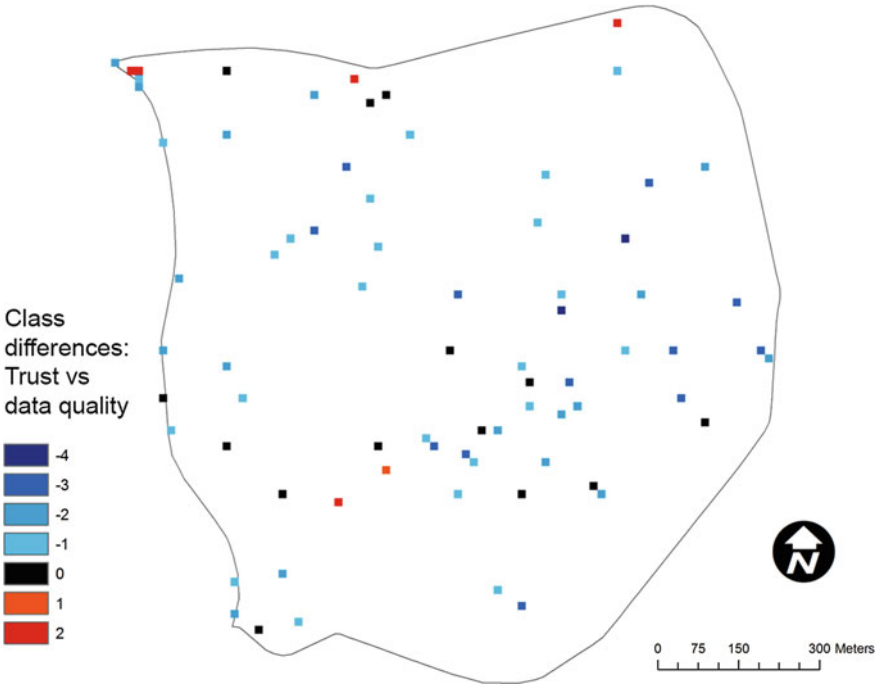


Fig. 5 Comparison of trustworthiness and data quality. *Blue colors* indicate that the trust assessments bear lower values than the observed data quality; vice versa for *red colors*

corrections or rollbacks, while the low quality values were often caused by a wrong feature type, in one case in combination with a high number of missing mandatory tags. This suggests that confirmations and rollbacks are the most distinctive indicators in our quality assessment.

Taking a closer look at the parameter values of the 20 features for which the classes of the two tests show no correspondence but rather opposition, reveals that a low number of versions, a low number of confirmations and the occurrence of tag corrections cause a low trust value for these features. In the rare case that the differences are in the opposite direction (high quality, but low trust), the theme of the feature was incorrect. There are only a few of these cases. Most of the big mismatches concern the first observation.

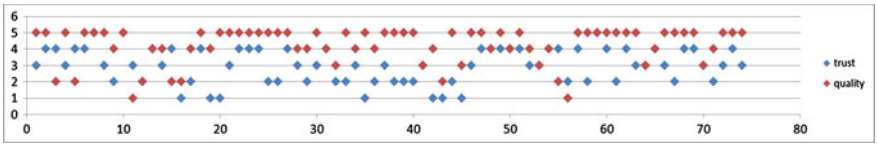


Fig. 6 Class value differences between the trust and quality measures; *the red dots* represent the quality measure results, *the blue* represent the trust measure results

The features with a high negative class difference generally have a higher number of tag corrections and slightly lower numbers of confirmations and users. Less emphasis on the tag corrections would increase the trust value of these features. Trying this out, however, resulted in more deviation from the trust value. This supports the model as is, but it also shows that one cannot always rely on counts of numbers of variables designated as trust parameters in order to get an indication of data quality.

6.2 Statistical Analysis

The general trend in Fig. 6 suggests a correlation between the trust assessment and the observed data quality. A statistical correlation analysis has been carried out to determine if there is an association between the two classifications, i.e., whether we can falsify our null hypothesis that there is no correlation between a trust assessment based on feature history, and the feature's observed data quality. Because the trust and quality measures both consist of a number of reclassifications and the quality measure includes a ranking, the whole set of trust and quality results does not have a clear numerical basis and is rather ordinal. We therefore applied Kendall's τ, a non-parametric rank correlation measure (Kendall 1938).

The total set of features was cleaned from outliers before the correlation measure was applied. The rationale here is that the 20 features not taken into account all show atypical behavior as discussed above. In this study, however, we are interested in whether using trust assessments as a proxy for data quality works at all, i.e., whether the trust assessments are able to predict the data quality in the common case. We thus leave a refinement of the trust assessment, including weights for the different parameters and the addition of new parameters such as user reputation, for future work.

Calculating Kendall's τ for the 54 features following the trend of the quality measure bears a positive correlation of $\tau \approx 0.52$, indicating a moderate correlation between the two measures. The p value for this correlation is close to zero (~ 0.00002), i.e., the correlation is highly significant. We can therefore reject the null hypothesis and accept the hypothesis that trust assessments and quality measures are correlated in a linear fashion. The statistical analysis thus suggests that provenance-based trust assessments do indeed work as proxies for data quality.

6.3 Discussion

Putting the result of the statistical analysis into context, several facts have to be taken into account. First, the model for trust assessment used in this chapter is very simple and does not take any of the finer details of trust models for VGI discussed in the existing literature into account. Specifically, the model does not take user

reputation into account, which could not be computed in a meaningful way on the sample dataset. A realistic model of user reputation would have to take all edits a user has made into account and evaluate how the edited features have been treated by the community afterwards (e.g., by looking at revisions), potentially weighted by their respective local knowledge (van Exel et al. 2010). This would enable us to assess that user's reputation through an assessment of the quality of her edits. The required data for this, however, are not efficiently to retrieve through the OpenStreetMap API for a large number of users, so that we would have to bring the OSM history dump into a format that supports such queries first. As mentioned before, this is a research question of its own and we wanted to make sure that the idea of trust as proxy for data quality makes sense before we address user reputation in the next step.

Second, we have been using a straightforward combination of the five different parameters in this chapter by assigning all parameters the same weight. Similar to the user reputation, a detailed analysis is required in order to find out which of these parameters have a bigger influence on data quality. This could also be done in a study similar to the one presented in this chapter; however, a larger sample dataset would be required to be able to see how different weights affect the correlation of trustworthiness and data quality. The weighting of the features could then also address the fact that the trust measure in its current form generally underestimates the data quality.

Third, a more thorough study of the systematics behind the outliers that were not taken into account for the statistical analysis is required. A closer look at the outliers reveals that some of them were deemed too trustworthy with respect to their actual data quality because the trust assessment did not take information completeness into account, which could be included in future versions of the trust assessment in the same way we have included it in the quality measurement, i.e., by identifying the critical tags per feature type. This method obviously fails, however, if the feature type is not correctly assigned, which was also the case for some of the outliers.

Despite the small sample dataset used in our study, the statistical analysis shows that assessments of trustworthiness as a proxy for data quality are a topic that is worth pursuing in future research, which was our main motivation for the research presented in this chapter. In the following section, we will conclude the chapter and discuss directions for future work.

7 Conclusions

In this chapter, we have investigated the question whether trust assessments based on the provenance of features in OpenStreetMap can act as a proxy measure for data quality. To test this hypothesis, we have calculated trust assessments for 74 features in Münster's old town district based on a simple trust model. For comparison, a ground truth dataset has been collected in a field survey, for which we

have defined a novel measure for tag importance in order to be able to measure information completeness. Based on a classification into equal intervals, the trust assessments and quality measurements per feature were tested for statistical correlation, which shows moderate, yet significant support for the hypothesis.

As discussed in the previous Sect. 6.3, this chapter is a first step towards meaningful quality assessments for Volunteered Geographic Information based on trust measures derived from feature provenance. This approach is novel in that it does not require ground truth data (except for the evaluation of the method itself) and can hence provide guidance for data consumers on an ad-hoc basis. Moreover, it bears the potential for new tools that support the community in spotting potentially problematic features that might need to be revised. In order to make trust as a proxy for VGI data quality operational, the weighting of the different parameters that influence trust needs to be analyzed, and user reputation needs to be taken into account. Moreover, the trust assessment needs to be extended to address the relatively large number of outliers that we have observed in our study.

The next steps in this research are therefore to collect a larger ground truth dataset to test different versions of the trust measure. The field surveys required to collect the ground truth data are very time-consuming and limited to a small area if, as in our case, performed by a single person. Crowd sourcing the quality assessment could hence be a useful approach to obtain a more diverse ground truth dataset, potentially through a smartphone app that lends ideas from gamification to motivate the users to participate. A larger ground truth dataset will allow us to address several open questions, including the actual importance and influence (negative or positive) of the parameters taken into account. This especially applies to the influence of tag corrections and rollbacks, which are currently treated as negative indicators. Moreover, additional and alternative approaches could be tested, such as using the number of change sets a feature is part of, instead of the number of versions. A systematic comparison of the test data against the best practices documented on the OpenStreetMap wiki could improve our understanding and rating of thematic accuracy and tag completeness. Putting these ideas into practice would also require an automation of the analysis, which was done largely manually for this chapter.

On the computational side, the user reputation needs to be addressed. If user reputation is handled as a function of a the quality of the respective user's edits, which in turn can only be assessed by the eventual development of the edited features, the computation becomes exponential and does not scale for large numbers of users. The user reputation assessment therefore needs to be driven by heuristics. User reputation is also an important aspect to address the quality assessments of features with a sparse editing history where analysis as discussed in this chapter is unlikely to bear meaningful results. In these cases, it is potentially more accurate to base the quality assessment solely on the reputation of the involved users, and eventually the social network relations between them (Mooney and Corcoran 2012a). User reputation is hence the most pressing issue to address in our future work.

References

Artz D, Gil Y (2007) A survey of trust in computer science and the semantic web. Web Semant 5:58–71

Bishr M, Janowicz K (2010) Can we trust information?—The case of volunteered geographic information. In Devaraju A, Llaves A, Maué P, Keßler C (eds) Towards digital earth: search, discover and share geospatial data 2010. Workshop at future internet symposium, Sep 2010

Bishr M, Kuhn W (2007) Geospatial information bottom-up: a matter of trust and semantics. In Fabrikant SI, Wachowicz M (eds) The European information society—leading the way with geo-information. Lecture Notes in Geoinformation and Cartography. Springer, Berlin Heidelberg, pp 365–387

Bishr M, Mantelas L (2008) A trust and reputation model for filtering and classifying knowledge about urban growth. GeoJournal 72(3–4):229–237

Flanagin AJ, Metzger MJ (2008) The credibility of volunteered geographic information. GeoJournal 72(3):137–148

Golbeck JA (2005) Computing and applying trust in web-based social networks. Ph.D. thesis, University of Maryland. Available from http://drum.lib.umd.edu/bitstream/1903/2384/1/umi-umd-2244.pdf

Goodchild MF (2007) Citizens as sensors: the world of volunteered geography. GeoJournal 69(4):211–221

Haklay M (2010) How good is volunteered geographical information? a comparative study of OpenStreetMap and ordnance survey datasets. Environ Plann B, Plann Des 37(4):682–703

Haklay M, Basiouka S, Antoniou V, Ather A (2010) How many volunteers does it take to map an area well? The validity of Linus' law to volunteered geographic information. Cartographic J 47(4):315–322

Helbich M, Amelunxen C, Neis P, Zipf A (2010) Investigations on locational accuracy of volunteered geographic information using OpenStreetMap data. GIScience 2010 Workshop on the role of volunteered geographic information in advancing science. Zurich, Switzerland. Available from http://ornl.org/sci/gist/workshops/2010/papers/Helbich.pdf

International Organization for Standardization (2002) ISO Standard 19113:2002: Geographic information—quality principles

Jakobsson A, Giversen J (2009) Guidelines for implementing the ISO 19100 geographic information quality standards in national mapping and cadastral agencies. EuroGraphics 2009

Kendall M (1938) A new measure of rank correlation. Biometrika 30(1–2):81–89

Keßler C, Trame J, Kauppinen T (2011a) Provenance and trust in volunteered geographic information: the case of OpenStreetMap. Poster presentation, conference on spatial information theory, 12–16 Sep 2011, Belfast, Maine, USA

Keßler C, Trame J, Kauppinen T (2011b) Tracking editing processes in volunteered geographic information: the case of OpenStreetMap. In Duckham M, Galton A, Worboys M (eds) Identifying objects, processes and events in spatio-temporally distributed data (IOPE), workshop at conference on spatial information theory 2011 (COSIT'11), 12 Sep 2011, Belfast, Maine, USA

Koukoletsos T, Haklay M, Ellul C (2012) Assessing data completeness of VGI through an automated matching procedure for linear data. Trans GIS 16(4):477–498

Mezzetti N (2004) A socially inspired reputation model. In: Proceedings of 1st European PKI Workshop. Lecture Notes in Computer Science, vol 3093. Springer, pp 191–204

Mooney P, Corcoran P (2012a) Who are the contributors to OpenStreetMap and what do they do?. In: Proceedings of 20th annual GIS research UK (GISRUK), Lancaster University, Apr 2012

Mooney P, Corcoran P (2012b) Characteristics of heavily edited objects in OpenStreetMap. Future Internet 4(1):285–305

Mooney P, Corcoran P (2012c) The annotation process in OpenStreetMap. Trans GIS 16(4):561–579

Neis P, Zielstra D, Zipf A (2012) The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011. Future Internet 4(1):1–21

Salton G, Wong A, Yang CS (1975) A vector space model for automatic indexing. Commun ACM 18(11):613–620

Simmhan YL, Plale B, Gannon D (2005) A survey of data provenance in e-science. ACM Sigmod Rec 34(3):31–36

Sztompka P (1999) Trust: a sociological theory. Cambridge University Press, Cambridge

van Exel M, Dias E, Fruijtier S (2010) The impact of crowdsourcing on spatial data quality indicators. GIScience 2010 workshop on the role of volunteered geographic information in advancing science. Zurich, Switzerland. Available from http://www.giscience2010.org/pdfs/paper_213.pdf

Zielstra D, Zipf A (2010) A comparative study of proprietary geodata and volunteered geographic information for Germany. AGILE 2010. In: The 13th AGILE international conference on geographic information science. Guimarães, Portugal

A Thematic Approach to User Similarity Built on Geosocial Check-ins

Grant McKenzie, Benjamin Adams and Krzysztof Janowicz

Abstract Computing user similarity is key for personalized location-based recommender systems and geographic information retrieval. So far, most existing work has focused on structured or semi-structured data to establish such measures. In this work, we propose topic modeling to exploit sparse, unstructured data, e.g., tips and reviews, as an additional feature to compute user similarity. Our model employs diagnosticity weighting based on the entropy of topics in order to assess the role of commonalities and variabilities between similar users. Finally, we offer a validation technique and results using data from the location-based social network Foursquare.

1 Introduction

Online social networking (OSN) offers new sources of rich geosocial data that can be exploited to improve geographic information retrieval and recommender systems. OSN platforms such as *Foursquare*, *Twitter*, and *Facebook* have taken advantage of the popularity of GPS-enabled mobile devices, allowing users to geotag their contributions, thus adding spatiotemporal context to their social interactions.

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

K. Janowicz
e-mail: jano@geog.ucsb.edu

B. Adams
National Center for Ecological Analysis and Synthesis (NCEAS),
University of California, Santa Barbara, USA
e-mail: adams@nceas.ucsb.edu

D. Vandenbroucke et al. (eds.), *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography,
DOI: 10.1007/978-3-319-00615-4_3,

This increase in social networking through portable devices has resulted in a shift from location-static updates to location-dynamic interactions, freeing online communication from the clutches of the desktop and immersing it in our mobile lives. Social network users post updates on the go from anywhere in the world, be it from a restaurant, mountain top, or airplane. These data are having a profound impact in the study areas of human mobility behavior, recommendation engines, and location-based similarity measurements.

The abundance of data published through online sources provides an exceptional foundation from which to investigate user similarity. To many users of these OSNs, the benefits of allowing access to this personal information is worth the cost of privacy. From a research perspective, these data offer an unprecedented opportunity to observe human behavior and design new methods for exploring the similarity between individuals. Studying similarity is important for several reasons. First, it can be used to suggest new contacts and thus, enrich the social network of a user. Second, as similar users are more likely to share similar interests, user similarities play a key role in recommender systems (Matyas and Schlieder 2009) and geographic information retrieval (Jones and Purves 2008). For instance, the *Last.fm* music platform offers social networking functions by which users can explore their *musical compatibility* with others and listen to their personalized radio stations. Third, and of most importance for our work, the information available about users, their locations, and activities is still sparse. User similarities can be exploited to predict *types* of activities and places preferred by a user based on those of users with similar preferences.

So far, most work on user similarity has mainly focused on structured, e.g., geographic coordinates, or semi-structured, e.g., tags and place categories, data. Unfortunately, these data are often unable to uncover nuanced differences and similarities. For instance, two users may frequently visit places tagged as *bar* and rated with a *Yelp* price range of $$. However, unstructured, textual descriptions reveal that only one of these users constantly visits places that offer pub quizzes. In this chapter we suggest exploring location-based social networking (LBSN) data to enhance current user similarity measures by focusing on unstructured data, namely *tips* provided by users. This approach explicitly focuses on the non-spatial components of user-contributed data, utilizing *topic modeling* together with *diagnosticity weights* determined by the entropy of different topics. The temporal properties of a user's trajectory are also included when calculating user similarity. Our initial results show that the similarity between individuals is not uniform throughout the day. Thus, instead of generalizing similarity simply to the user level, we propose a method for assessing similarity on an activity-by-activity basis, exploiting the temporal as well as the spatial attributes of a user's trajectory.

The remainder of the chapter is organized as follows. In Sect. 2, we discuss related work on user similarity and location-based social networks. Section 3 focuses on data mining and the methods used for defining user similarity. In Sect. 4, we present results based on actual user data. Section 5 discusses a few of the limitations we faced in conducting this research and Sect. 6 presents our conclusions and points out directions for future work.

2 Related Work

Assessing user similarity has become an important topic in information retrieval and recommender systems over the past few years. The motivations for developing user similarity measures range considerably, from recommendation systems (Guy et al. 2009; Horozov et al. 2006) and dating sites (Hitsch et al. 2010) to location and activity prediction (Lima and Musolesi 2012; Noulas et al. 2012).

A number of recent studies have focused on measuring user similarity through trajectory comparison (Lee et al. 2007; Li et al. 2008; Ying et al. 2010). In Lee et al. (2007), explore a geometric approach to trajectory similarity by exploiting three types of distance measures in order to group trajectories. While their *Partition-and-Group* framework is unique, it is limited to the geospatial realm, overlooking the types of activities and social information related to the activity locations. Similarly, Li et al. (2008) focused on the spatial components of user trajectories. Their method employs hierarchical trajectory sequence matching to determine similar users. Making use of GPS tracks, Li et al. extract *stay points* at which a user's activity is determined based on the affordances of a specific location.

While the above methods measure user similarity based on geospatial aspects of user trajectories, we argue that an understanding of the semantics of an activity space are essential. Ye et al. (2011) investigate the concept of semantic annotations for venue categorization. In developing a semantic signature for a categorized place based on *check-in* behavior, similar, uncategorized places could be discovered. This concept of semantic signatures may also be applied to assessing user similarity through semantic trajectories. In this vein, Ying et al. (2010), measured semantic similarity between user trajectories in order to developed a *friend recommendation system*. This work focuses on the type of activities completed by each user and the sequence in which these activities take place. Akin to the *stay point* work presented by Li et al. (2008), the authors focus on *stay cells* and obtaining a semantic understanding of the types of activities conducted within the cells. From there, a semantic trajectory is formed and patterns are assessed and compared between users.

Activity prediction research can also benefit from exploring user similarity. Based on check-in data gathered through *Foursquare*, Noulas et al. (2012) exploit factors such as transition between types of places, mobility flows between venues and spatial-temporal characteristics of user check-in patterns to build a supervised model for predicting a user's next check-in. This method, while exploring previous check-ins across users, does not assess similarity between users in predicting future locations, an aspect that our research suggests is beneficial. Traditional work in collaborative filtering (e.g., Amazon recommendations) has also focused on measuring user similarity, but typically concentrates on "structured" data such as numerical (star) ratings (Linden 2003; Herlocker et al. 2004).

Recently, Lee and Chung (2011) presented a method for determining user similarity based on LBSN data. While the authors also made use of check-in information, they concentrated on the hierarchy location categories supplied by

Foursquare in conjunction with the frequency of check-ins to determine a measure of similarity. By comparison, our approach is novel in that it makes use of an abundance of unstructured descriptive text (tips) provided by visitors of specific venues rather than a single categorical value.

3 Methodology

In this section, we describe the data collection, topic extraction, and methodology used for developing our user similarity measures.

3.1 Data Source

The location-based social networking platform, *Foursquare*, was used as our primary source of modeling data based on the shear number of crowdsourced venues as well as its ubiquity as a location-based application. As the application defines it, a venue is a user-contributed "physical location, such as a place of business or personal residence".[1] and as of publication, *Foursquare* boasts over 9 million venues in the continental United States alone. This platform allows users to *check in* to a specific venue, sharing their location with anyone they have authorized as well as other OSNs such as *Facebook* or *Twitter*. Built with a gamification strategy, users are rewarded for checking in to locations with badges, in-game points, and discounts from advertisers. This game-play encourages users to revisit the application, compete against their friends and contribute *check-ins, photos* and *tips*.

Venue Tips An additional feature of *Foursquare*, is the ability for a user to contribute text-based *tips* to a venue. *Tips* consist of user input on a specific venue and can range from a restaurant review to a hiking recommendation. Lacking any official descriptive text for venues on *Foursquare*, these unstructured tips describe and define the venue and location. As with most crowdsourced data, the length, content, and number of tips vary significantly throughout the *Foursquare* venue data set. Of the 9 million *Foursquare* venues available in the continental United States, approximately 22.8 % included at least one tip. Taking only venues that have had more than ten unique user check-ins, this value jumps to 54.0 %. Of the venues to which our sample population checked in, 77.0 % include at least one tip with the mean length of a single tip being 74 characters (stdev = 49.3). Table 1 shows a few examples of tips left at different venues.

[1] https://foursquare.com/

Table 1 Example tips

Order your tacos with flour tortilla and use their amazing green salsa!
Free wifi & power outlets outside work. Let's support and make sure they'll be there a long time
I just bought some leather chairs and I love them, great quality furniture

3.2 Data Collection

Publicly geotagged *Foursquare* check-ins were accessed via the *Twitter API* for 6,000 users over a period of 128 days. Check-ins to venues with less than ten tips were removed as well as users with an overall check-in count less than 16. This resulted in a dataset totaling 24,788 check-ins over 11,915 venues for 797 users (mean of 31.1 check-ins per user). From a geosocial perspective, we define an individual's activity identity as an amalgamation of the venues to which she checks in.

3.3 Themes

In this work, we use a Latent Dirichlet Allocation (LDA) topic model to extract a finite number of descriptive themes (topics) from the user-generated tips assigned to venues in our Foursquare dataset. While numerous topic models are discussed in the literature, LDA is a state-of-the art generative probabilistic topic model that can be used to infer the latent topics in a large textual corpus in an unsupervised manner (Blei et al. 2003). A topic is a multinomial distribution over terms, where the distribution describes the probabilities that a topic will generate a specific word. LDA models each document as a mixture of these topics based on a Dirichlet distribution. Several mature implementations of LDA with improvements exist; for this work we employ the implementation in the MALLET toolkit (Kachites and McCallum 2002).

A topic model is run across all *Foursquare* venues in the continental United States containing ten or more *tips* (approximately 125,265 venues). Tips are grouped by unique venue ID and all stop-words, symbols, and punctuation are removed as well as the 30 most common words.

Venue Themes Using this model we are able to express each venue as a mixture of a given number of topics. The model was tested with 40 topics at 2,000 iterations. Future work could involve running similarity models with a varied number of topics. A few of the topics are concerned with a specific type of food, while others are focused on tourism and even baseball. Table 2 shows four examples of topics, based on top terms, extracted using LDA.

Temporal Themes The daily trajectories for each of the 797 users in our dataset are grouped by user and aggregated to a single day. Given the limited number of

Table 2 Sample topics derived from Tip text represented as word clouds, where larger words are higher probability words for the topic

High Entropy	Low Entropy
	

check-ins, aggregating user activities to a single day was deemed appropriate. Over the 128 days of data collection, this produced a sparse average of 31.1 check-ins per user. This would not be sufficient for any prediction and additionally highlights the need to select similar users as proxies. Selecting one user as our base-line or *focal user*, each check-in in her trajectory is buffered by 1.5 h. This so-called 3 h *time window* is used as the temporal bounds from which all additional users' activities are collected. From there we calculate the topic signature for all users within this same time window. This produces an aggregate venue topic distribution for every user over a 3-h time window around each of the *focal user's* check-ins; 1.5 h before and 1.5 h after the check-in. Given these distinctive topic signatures, it is feasible to compare users temporally, across these topics in order to produce a user similarity measure.

A topic *signature* is computed for each of the collections via Eq. 1 where T_i is one topic in the collective topic distribution, n is the number of venues in the collection, $\#V_j$ is the number of times the same venue appears in the collection and t_i^{Vj} is a single topic probability of Venue j. It is important to note that this method takes the frequency of check-ins to a unique venue into consideration. This ensures that multiple check-ins to a single location do not over-influence the topic distribution.

$$T_i = \sum_{j=1}^{n} (\log_{10} \#V_j + 1) t_i^{Vj} \tag{1}$$

3.4 Variability Versus Commonality Weighting

This approach to calculating the topic signature for a collection of venues puts an equal amount of emphasis on all topics. This is not ideal when measuring the similarity between signatures as some topics are more prevalent across all venues than others. In order to augment the similarity model, we compute the entropy for each topic across all venues. In Table 2, two of the word clouds are examples of topics showing high entropy while the other two represent topics with low entropy.

Let t_i be the weight of topic t for venue i. A new discrete variable is defined for topics over venues by normalizing each t_i to t'_i by setting $t'_i = \frac{t_i}{\sum_{j=1}^{N} t_j}$, where n is the number of venues. The topic's entropy over all venues, E_T, is defined in Eq. 2.

$$E_T = -\sum_{j=1}^{n} t'_j \log_2 t'_j. \tag{2}$$

Given this set of entropy values, a method for incorporating them as weights in a user similarity model must be assessed. This leads to questioning the role of topic prevalence in constructing a model for assessing user similarity. The approaches we present in the following subsections are influenced by literature in the cognitive sciences that examined the role of context (or framing) in human similarity assessments. Tversky (1977) found that when two objects are compared for similarity, the set of objects from which the two objects are selected has the effect of making some properties more or less salient in the similarity judgment. The properties that are more salient are termed to be more 'diagnostic'. Tversky argued that two factors contribute to the *diagnosticity* of a property. The first is *variability*, which finds that the properties that vary across the elements of the context set are used more to determine the similarity (or dissimilarity) of two objects. The second factor *commonality*, is the opposite, that properties that are shared by most elements of the context set are the important properties, because they help explain what is important in the domain of discourse.

Although this context effect is well-studied in the cognitive sciences most computer science similarity measurements are without context in this sense. A notable exception is the *Matching-Distance Similarity Measure* (MDSM), created to compare similarity of spatial entity classes (Rodriguez and Egenhofer 2004). MDSM defines commonality and variability metrics for feature-based classes. In the following sections we adopt these notions to the venue topic signatures.

Variability One approach postulates that though the commonality topics remain critical in defining the venue (or user), they are less valuable in determining the similarities between two users. For example, if all venues in a dataset are high in a topic related to coffee, this topic does little in determining which two users are most similar. It is the less ubiquitous topics which are more *diagnostic* in the similarity model. Based on the literature on similarity (Tversky 1977), we call this type of diagnosticity, the *variability* weight.

In order to add weight to these more diagnostic topics, we build our similarity model based on a subset of ten topics with the highest entropy. Given the reduction in the number of topics, the collective topic distribution must then be normalized ($n = 10$) to sum to 1 in order to compare distributions.

Commonality It may be argued that the inverse effect of variability, *commonality* is more applicable. A *commonality* weight implies that more prevalent topics should be more influential in measuring user similarity. In essence, the more coffee shops one visits, the more similar they are to other coffee shop visitors.

The influence of entropy on topics using this commonality method involves taking the top ten topics with the lowest entropy and building our similarity model based purely on those topics. Again, the collective topic distribution is normalized in order to sum to 1.

3.5 Comparing Users

Since each aggregate venue signatures consist of a distribution over an equal number of topics, a divergence metric may be used to measure the similarity between our *focal user* and all other users at at any given activity. Using the *Jensen-Shannon divergence* (*JSD*) (Eq. 3), we compute a dissimilarity metric between each user's topic distribution and the *focal user's* respective topic signature. $U1$ and $U2$ represent the topic signatures for User 1 and User 2 respectively, $M = \frac{1}{2}(U1 + U2)$ and $KLD(U1 \parallel M)$ and $KLD(U2 \parallel M)$ are *Kullback-Leibler divergences* as shown in Eq. 4.

$$JSD(U1 \parallel U2) = \frac{1}{2}KLD(U1 \parallel M) + \frac{1}{2}KLD(U2 \parallel M) \tag{3}$$

$$KLD(P \parallel Q) = \sum_i P(i) \log_2 \frac{P(i)}{Q(i)} \tag{4}$$

The *JSD metric* is calculated by taking the square root of the value resulting from the equation. Given the inclusion of the logarithm base 2, the resulting metric is bound between 0 and 1 with 0 indicating that the two users' topic signatures are identical and 1 representing complete dissimilarity.

4 Results and Discussion

Selecting a *focal user* at random from the 797 users, we first run the basic *JSD* dissimilarity model without including an entropy weight. In order to keep the number of topics uniform across all models, a set of ten topics are randomly

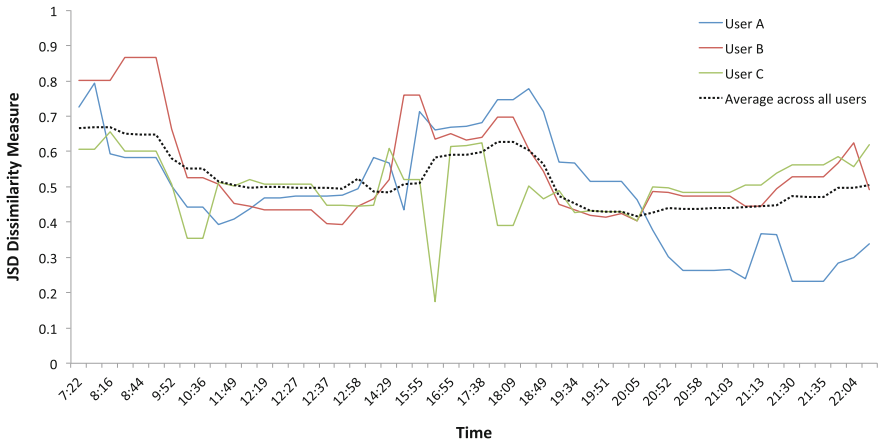


Fig. 1 Similarity of *User A, B & C* to Focal User (randomly selected topics)

selected for comparison. Figure 1 shows the dissimilarity metrics at activity level resolution for 3 individuals compared to the *focal user*. As one can see, *User A's* similarity to the *focal user* generally decreases as the day progresses, with late evening proving to be the most similar time of day, *User B* is similar around lunchtime and quite dissimilar in the morning. Lastly, *User C* mirrors the average for most of the day with a small bump in the morning and a sharp peak of similarity at around 16:30.

In comparison, Fig. 2 shows the effect of including the entropy measure with the purpose of emphasizing more diagnostic topics within the venue distributions. The same three users are compared to our *focal user*, but this time the venue distribution is composed of topics high in variability. The most visible outcome of the variability weight inclusion is an increase in range of similarity measures

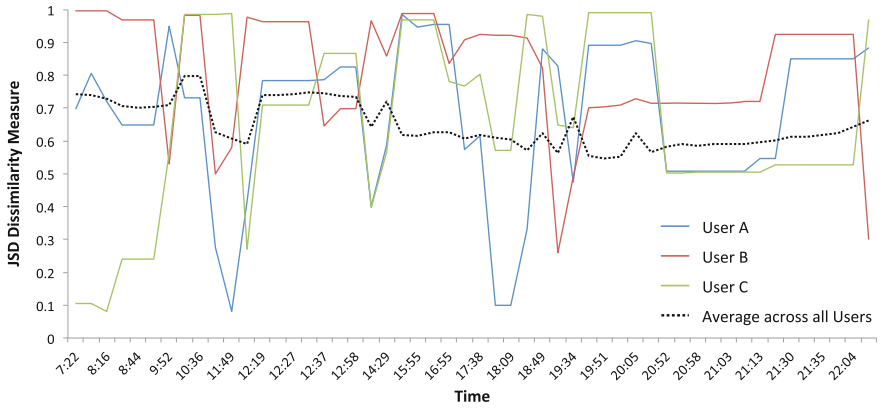


Fig. 2 Similarity of *User A, B & C* to Focal User (topics with highest entropy)

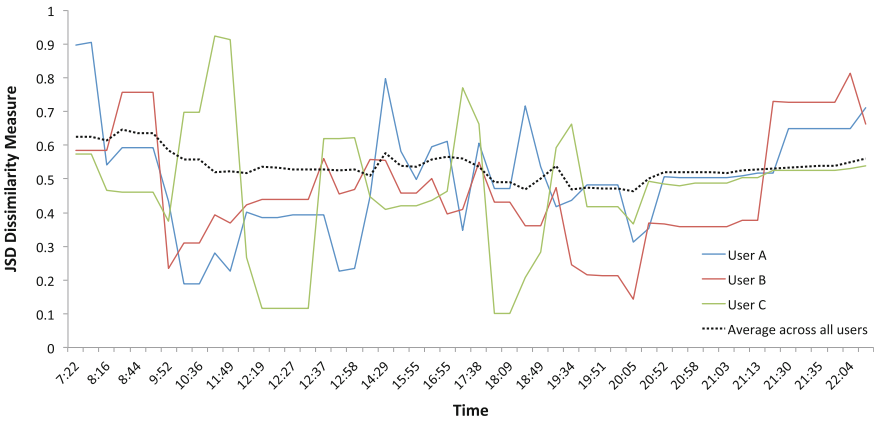


Fig. 3 Similarity of *User A, B & C* to Focal User (topics with lowest entropy)

across users. Each of the three users is completely dissimilar to our *focal user* at some point during the day and the average dissimilarity across all users has increased.

Interestingly enough, each of the sampled users increased their similarity to the *focal user* at least once throughout the day. Given that these topics offer the largest variability within the dataset, it is not surprising that a measure of similarity between users based purely on these topics will decrease overall in comparison to the non-entropy selection. This variability model will return specific peaks of similarity between users given that it is emphasizing the topics not as common across all venues. *User A* and *User B* show dramatic increases in similarity in the morning, with *User C* peaking around dinnertime. As this figure makes apparent, the change in user similarity is not uniform across all activities or users, it is dependent on the prevalence of a given topic (or combination of topics) within the aggregated distribution of an activity venue.

The *commonality* model offers a very different perspective. Figure 3 shows that on average, the similarity between all users and the *focal user* increased. While some semblance of the random-topics figure still exists, the users appear more uniform in their similarity to our *focal user*.

4.1 Validating the Model

This section presents the methods used to validate the similarity model as well as the results of the validation. Both of the entropy-based similarity models are evaluated along with the non-entropy model. The methods below are applied to each model.

To start, the topic distributions for the top-k most similar users for each check-in are combined using Eq. 5. The influence of each user on the combined topic

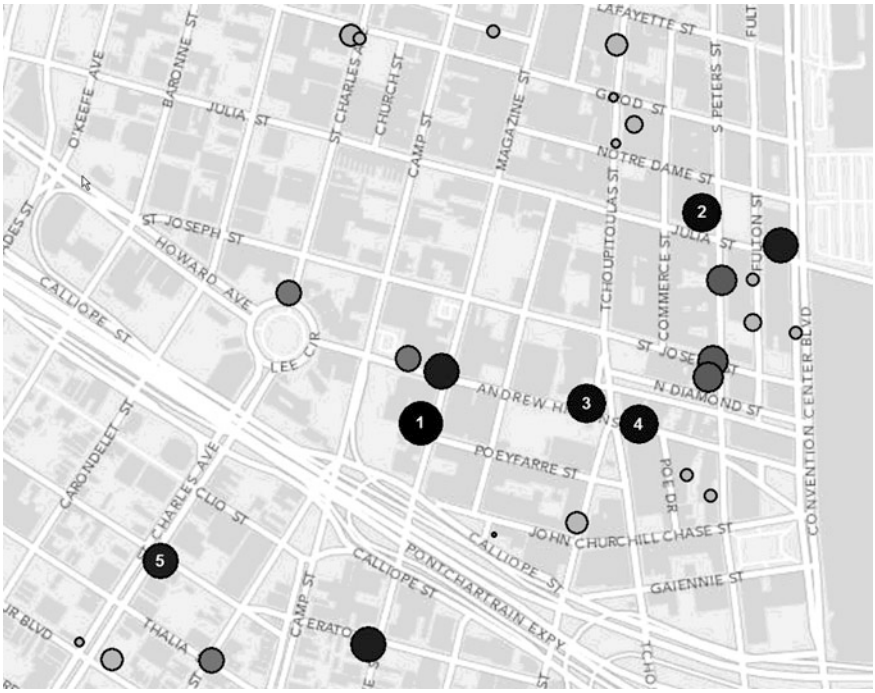


Fig. 4 Map fragment showing graduated symbols for the 30 nearest venues

distribution (*HV*) is calculated by multiplying the topic by the similarity value *sim* where $sim = 1 - dissimilarity$. This ensures that more similar users have a larger impact on the overall topic signature.

$$HV_{T_i} = \sum_{j=1}^{n}((sim_j) * T_i^j) / \sum_{i=1}^{m} T_i \tag{5}$$

The resulting topic distribution represents a *hypothetical venue (HV)* that is the most similar to the *focal user's* check-in location as possible based on the model. In order to evaluate this *hypothetical venue*, we extract the 29 nearest (physically) venues (along with their topic distributions) for each of the *focal user's* check-ins. This collection of venues, along with the actual check-in venue, form the test set from which the similarity model is assessed.

The 30 sample venues are ranked in order of similarity to the *hypothetical venue* and the position of the real check-in venue within this ranked set is recorded. Figure 4 shows an example with graduated symbol markers representing the dissimilarity of each venue (large dark color = low dissimilarity). In this example, the top 5 most similar venues are labeled with the actual check-in venue resulting in 1 (the most similar venue to the *hypothetical venue*). This process is run across all check-ins for all users with the three levels of weighting. Table 3 shows an

Table 3 Placement of actual venue based on similarity to *Hypothetical Venue*

Placement	Commonality (%)	Variability (%)	Random (%)
1	77.02	45.04	65.85
2	14.16	17.17	18.05
3	4.30	9.38	7.22
4	1.98	5.65	4.02
5	1.16	2.86	2.23
6	0.53	2.17	1.04
7	0.28	1.88	0.56
8	0.16	1.22	0.25
9	0.09	1.16	0.25
10	0.03	0.97	0.16
11	0.03	0.50	0.03
12	0.06	0.88	0.06
13	0.00	0.53	0.03
14	0.03	0.35	0.00
15	0.03	0.16	0.00
16	0.00	0.22	0.06
17	0.00	0.97	0.03
18	0.00	0.63	0.00
19	0.00	0.44	0.03
20	0.00	0.50	0.00
21	0.00	0.09	0.03
22	0.06	0.31	0.03
23	0.00	0.19	0.00
24	0.03	0.22	0.00
25	0.00	0.53	0.03
26	0.00	0.82	0.03
27	0.00	1.10	0.00
28	0.00	1.29	0.00
29	0.03	1.69	0.00
30	0.00	1.07	0.00

ordered-position table based on 3,188 sampled check-ins over the 797 users in our dataset (4 randomly sampled check-ins per user). Both the 40 Topic model and the 30 Topic model are present in this table, showing the results for the *Variability*, *Commonality* and *No weight* models.

The *Commonality* weighted model produced the best results with over 77 % of the *hypothetical venues* contributing to a correct estimation of the actual venue. In fact, the *Commonality* weighted model placed the actual check-in venue within the first 3 most similar venues 95 % of the time. By comparison, the *Variability* weighted model was significantly less accurate, correctly estimating the actual check-in venue 45 % of the time. While this performance is not as strong as the *commonality* weighted model, it is to be expected as the purpose of exploiting the *variability* topics within the topic distribution is to find the nuanced differences between venues rather than the overall commonality between them. Lastly, the

results of the *non-weighted* , randomly-sampled topic model are presented. As a base-line, we see that even without the inclusion of entropy weighting, this similarity model produces excellent results with 65 % of actual venues being correctly estimated. In all cases, these results suggest that the model performs quite well in estimating an actual check-in based purely on the check-ins of similar users.

5 Limitations

While the methods presented in this chapter offer a promising approach to assessing user similarity through unstructured data, there are a number of limitations. Since the topic models are built on crowdsourced data (*tips*) from users of the application, the standard bias and errors of crowdsourcing are present. There is no way to ensure that a user submitting a tip has ever been to the venue or is offering a truthful tip. Additionally, since all tips for a single venue are combined in order to run the LDA model, those tips with more content have a large impact on the overall generation of topics. While there has been an increase in the number of people using LBSN applications, it should be noted that one's *Foursquare* check-in history does not account for every single activity that the user conducts throughout her day; the average user does not *check in* to every venue that she visits. It is more likely that a user checks in to locations that are unique or different from those to which she normally checks in. To some users, one venue might offer more social capital (Pultar et al. 2010) than another (e.g., nightclub vs. hospital) and user's opinions range on what is *unique*. However, the limitations discussed here also hold for most other methods designed based on volunteered geographic information and are a research challenge.

6 Conclusion and Future Work

The work presented in this chapter offers an overview of an innovative approach to assessing user similarity across sparse, unstructured geosocial check-ins. In this chapter, we explicitly extract the non-spatial components from the spatial data by focusing purely on the textual descriptions of locations. Given the amorphous nature of online social networking data, topic modeling has allowed us to extract themes from crowdsourced social data. These themes are merged across venues to produce a unique signature that defines an individual's geosocial activities at any given point in time. Through exploration of *variability* and *commonality* measures, based on the entropy calculated across these themes, we have shown two opposing methods for evaluating user similarity through publicly available *check-in* data. A model based on *Commonality* within the data produces the best results when estimating real check-ins from a set of nearby locations. The *Variability* within the venue topics allows us to explore the nuanced similarities between users and the

venues they frequent. In all, these methods demonstrate value in their ability to enhance existing user similarity models.

Future work in this area will flow in a number of directions. With an increase in the amount of user check-ins, the data will allow for further temporal factoring to reflect day of the week and month. It is expected that a user's activity patterns are not limited to hours within a day, but also reflects days of the week. The addition of temporal components will further enhance the ability of the model to discover similar users. Exploring the factors that contribute to this measure of user similarity will be a next step in this area of research as well. Analysis involving the correlation between location types and similarity measurements should be examined as well as outside factors that may contribute to similarity between users (e.g., demographic data, climate, etc).

Additional sources of unstructured geosocial content will be explored with the goal of enhancing the extraction of topics for venues. An incredible amount of unstructured geo-tagged content is available online and the addition of this data to our model will dramatically increase its accuracy. Lastly, while the sparsity of the data and the results gathered from such data is a novelty of this research, more precise activity information for a population of individuals (through a GPS enabled mobile device for example) will be tested order to assess the robustness of the model.

References

Blei David M, Ng Andrew Y, Jordan Michael I (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022

Guy I, Ronen I, Wilcox E (2009) Do you know? recommending people to invite into your social network. In: Proceedings of the 14th international conference on Intelligent user, interfaces, pp 77–86

Herlocker JL, Konstan JA, Terveen LG, Riedl JT (2004) Evaluating collaborative filtering recommender systems. ACM Trans Inf Syst (TOIS) 22(1):5–53

Hitsch Günter J, Hortaçsu Ali, Ariely Dan (2010) Matching and sorting in online dating. Am Econ Rev 100(1):130–163

Horozov T, Narasimhan N, Vasudevan V (2006) Using location for personalized poi recommendations in mobile environments. SAINT, page 124G129

Jones CB, Purves RS (2008) Geographical information retrieval. Int J Geograph Inf Sci 22(3):219–228

Lee JG, Han J, Whang K-Y (2007) Trajectory clustering : a partition-and-group framework G. In: International conference on management of data, pp 593–604

Lee M, Chung C (2011) A user similarity calculation based on the location for social network services. DASFAA, pp 38–52

Li Q, Zheng Y, Xie X, Chen Y, Liu W, Ma WY (2008) Mining user similarity based on location history. In: Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems - GIS '08, p 34

Lima A, Musolesi M (2012) Spatial dissemination metrics for location-based social networks. In: UbiComp 2012

Linden G, Smith B, York J (2003) Amazon. com recommendations: Item-to-item collaborative filtering. Internet Comput, IEEE 7(1):76–80

Matyas C, Schlieder C (2009) A spatial user similarity measure for geographic recommender systems. GeoSpatial Semantics, pp 122–139

McCallum AK (2002) Mallet: a machine learning for language toolkit. http://mallet.cs.umass.edu

Noulas A, Scellato S, Lathia N, Mascolo C (2012) Mining user mobility features for next place prediction in location-based services. In: International conference on data mining

Pultar E, Winter S, Raubal M (2010) Location-based social network capital. In GIScience, Extended Abstracts

Andrea Rodriguez M, Egenhofer MJ (2004) Comparing geospatial entity classes: an asymmetric and context-dependent similarity measure. Int J Geograph Inf Sci 18(3):229–256

Tversky A (1977) Features of similarity. Psychol Rev 84(4):327–352

Ye M, Shou D, Lee WC, Yin P, Janowicz K (2011) On the semantic annotation of places in location-based social networks. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge discovery and data mining, pp 520–528

Ying JJC, Lu EHC, Lee WC, Weng TC, Tseng VS (2010) Mining user similarity from semantic trajectories. In: Proceedings of the 2nd ACM SIGSPATIAL international workshop on location based social networks - LBSN ’10, pp 19–26

Using Data from Location Based Social Networks for Urban Activity Clustering

Roberto Rösler and Thomas Liebig

Abstract Understanding the spatial and temporal aspects of activities in urban regions is one of the key challenges for the emerging fields of urban computing and emergency management as it provides indispensable insights on the quality of services in urban environments and helps to describe the socio-dynamics of urban districts. This work presents a novel approach to obtain this highly valuable knowledge. We hereby propose a segmentation of a city into clusters based on activity profiles using data from a Location Based Social Network (LBSN). In our approach, a segment is represented by different locations sharing the same temporal distribution of check-ins. We reveal how to describe the topic of the determined segments by modelling the difference to the overall temporal distribution of check-ins of the region. Furthermore, a technique from multidimensional scaling is adopted to compute a classification of all segments and visualize the results. The proposed method was successfully applied to Foursquare data recorded from May to October 2012 in the region of Cologne (Germany) and returns clear patterns separating areas known for different activities like nightlife or daily work. Finally, we discuss different aspects related to the use of data from LBSNs.

1 Introduction

Residents do not use urban space homogeneously. Whereas some areas consist of residential quarters, others represent nightlife or industrial districts. Thus, the "usage pattern" of a city centre differs from an industrial region or a trendy

R. Rösler (✉)
Fraunhofer IAIS, Schloss Birlinghoven, 53757 Sankt Augustin, Germany
e-mail: roberto.roesler@iais.fraunhofer.de

T. Liebig
Department of Computer Science LS8, TU Dortmund University,
44221 Dortmund, Germany
e-mail: thomas.liebig@tu-dortmund.de

D. Vandenbroucke et al. (eds.), *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography,
DOI: 10.1007/978-3-319-00615-4_4,

neighbourhood; the shops in the city centre are visited during regular opening times whereas the bars attract people especially in the evening hours.

The identification of places with similar usage is an interesting topic for authorities, urban analysts and residents, as it provides valuable insights. For example it can be used for assessing the quality of services in urban environments and helps to describe and understand the socio-dynamics of these areas or to support town development planning. Furthermore it can be used to explore the genuine use-related urban structure on an up-to-date and micro-geographic level, which could be contrasted with official planning data (may be out dated, on a much higher level or incomplete) and hence reveal important information about planning deviation. Another application could be, to improve official databases. In this context it has also a second function, as the underlying empirical, open and fine-grained data capturing the socio-dynamics might provide some unbiased knowledge, which is generally unavailable to individuals without local knowledge and access to special data sets. Furthermore, the identification of similar regions is crucial for evaluating (scoring or ranking) the performance of places under an economic view e.g., shopping facilities or nightlife areas. Shop planners would benefit from this information, when making a selection of possible sites for daytime dependent business. In the field of disaster management, local activity profiles are of high value for planning preventive actions by responsible agencies because the profiles provide knowledge of typical spatio-temporal activities in different districts. Hence, it enables the planning authorities to facilitate effective and forward-looking action plans and gives them the opportunity for an optimized resource management.

This chapter addresses the question of how to identify spatial regions with similar temporal activities using widely available up-to-date data about the interaction of people and places. Our approach utilizes location based social network data as input for a spectral clustering.

The data contains so-called check-ins for spatial locations combined with person identifiers and a feature type (i.e., bar, restaurant, etc.) for the location (referred as venues). The temporal aggregation of the check-in frequencies per hour results in a vector per location containing 24 integers. These vectors and therefore the associated spatial objects are used for clustering. This is in contrast to Andrienko et al. (2012), where the spatial situations (i.e., the presence or flow aggregates among all locations at a particular time-stamp) are subject to clustering. The method we propose utilizes spectral clustering and thus provides the capability to find arbitrarily shaped clusters without posing any constraints. The intuitive colouring of the resulting clusters helps to understand the activity profiles of the considered regions. To achieve this, we apply Sammon's projection (compare Sect. 3).

The proposed algorithm has been successfully applied to Foursquare check-ins recorded from May 2012 to October 2012. The novel contributions of this chapter are threefold: (1) The clustering based on constructed activity profiles provides a deeper understanding of the spatio-temporal structure of a city. (2) It is shown, how a new combination and extension of existing approaches allows a more natural way to cope with typical real-world clustering problems like the estimation

of parameter values, the choice of a useful similarity metric or assumptions about the cluster type. (3) Moreover, to our best knowledge, this is the first approach using LBSN data for micro-geographic modelling outside the "densely monitored regions" (foremost in the US), which are used by most of the studies analysing LBSNs. The hereby-studied dataset is sparser. With this dataset, we still achieved reasonable results, however, a systematic analysis of the impact and requirements on sampling density are not subject of this chapter.

A systematic literature survey and outlook on future extensions completes this work.

The remainder of the chapter proceeds as follows. Section 2 highlights other related approaches for analysing or using data from location based social networks. In Sect. 3, we introduce our novel approach to model local spatio-temporal activities. Afterwards, in Sect. 4, we conduct our experiments with a subsample of the Foursquare data. We close with a discussion and an outlook on future work in Sect. 5.

2 Related Work

Area of Application

The field of Urban Computing analyses how modern ICT infer and integrates with urban life and is one of the emerging research fields. There is also an obvious interest in the relation between urban dynamics and data gathered from various sources depicting human activity to improve and fasten the understanding of social processes and interactions (Kindberg et al. 2007). One approach is about using cell phone data to model time-dependent behaviour of a city by clustering and, similar to our approach, interpreting the resulting patterns (Reades et al. 2007). However, obtaining cell phone data is much more complicated (access, size, preparation and legal aspects) and often restricted to certain purposes (not to mention the public concern about the violation of privacy by analysing mobile-device data).[1] In contrast to that, parts of the data from LBSNs are public to everyone and feature a considerably less complicated structure.

The same holds when it comes to data from surveys. The research in Jiang et al. (2012) shows an example where data from a large-scale survey is used to model the urban spatio-temporal structure in the Chicago metropolitan region. Even though the data is publicly available and forms a good representation of the total population, it only covers one region. In addition, the question arises if the individual design makes surveys from different regions comparable to each other. Nevertheless, the fusion of data from various sources like cell phone data, surveys and LBSNs seems to offer a major advantage in understanding urban life.

[1] http://www.telecompaper.com/news/german-govt-to-limit-telefonica-plans-to-sell-customer-data–905518 (last visited: 14.11.2012).

Data from the new field of LBSNs stimulates different researchers to analyse urban life. In (Aubrecht et al. 2011) the aim is to model the spatio-temporal characteristics of urban land use based on information from Foursquare whereas Noulas et al. (2011) uses the feature type of venues to identify user communities and urban neighbourhoods. Similar to that, the 'Livehoods Project' (Cranshaw et al. 2012) takes a more natural approach to characterize and distinguish different social areas. For that, the authors used check-in data from Foursquare giving them a more realistic picture than using official municipal organizational units. Hence, in contrast to our approach the 'Livehoods Project' is more focused on the delimitation of social than functional areas.

Besides the understanding of urban life also in emergency management, data from location-based social networks and microblogging services seem to form a valuable source to get some early information of potential crisis events. Different approaches therefore use data from Twitter (De Longueville et al. 2009; Noulas et al. 2011; Thom et al. 2012) or the combined data from different LBSNs to detect disasters (Chen et al. 2011) and present information to officials and emergency personnel.

General Aspects Using Data from LBSNs

LBSNs form a completely new phenomenon for the research community and therefore require some basic understanding of the motivation why, where and how people share information about their location and mobility behaviour and how they deal with the aspect of privacy (Lindqvist et al. 2011). Other researchers analyse characteristics of human mobility through data from social networks like in Noulas et al. (2011). For example, Cheng et al. (2011) explore some general aspects of movement patterns, returning probability and economic and geographic constraints using a dataset of 22 million check-ins worldwide. They also analyse the textual content given by short messages or announcements from the check-ins to identify significant terms and sentiments about the locations visited by the users. The authors in Cho et al. (2011) analysed users movement in relation to their social relationships represented by the structure of their social network. Their findings show that the social relationship explains a significant amount of human movement even though most of it is explained already through periodic behaviour. With the possibility to explain historical movements, it seems logical that other researchers also look at the task of making predictions of future movements on the basis of the user's history and the structure of the social network (Cho et al. 2011; Gao et al. 2012).

Spatial Topic Modelling

One important field that brings together the data from Location Based Services and Urban Computing is topic modelling. Here different authors analyse the existence of local geographic topics which essentially are locations connected by user movements and share some common theme—like 'sports' or 'business trip' (Joseph et al. 2012; Long et al. 2012). While Foursquare data seems to be

dominating most of the research papers mentioned here, the authors in Ye et al. (2011) explore Whrrl,[2] a different source which is not active anymore. They discover temporal patterns related to venues and their categories. They show that even if the feature type (e.g., college, restaurant) is not consistently used across all venues, the distribution of check-ins in time reveals distinguishable patterns between categories. Besides using data from LBSN other researchers show how to extract topics, make location prediction (this is an important step because still a lot of user generated content comes without coordinates) (Hong et al. 2012) and use identified sentiments to improve services for tourists (Shimada et al. 2011), all with data from micro-blogging services like Twitter.

Privacy

The last discussed aspect here (but not less important) is the dimension of privacy for both, users and analysts, when dealing with data or topics related to LBSNs. A Characterization of non-private information and status messages of users and venues which are available from Foursquare (e.g., tips, mayorship status) is provided in Pontes et al. (2012). The authors furthermore estimate the home city of a person using only publicly accessible information. In Jin et al. (2012) a new framework for preserving residential privacy for users from Foursquare is proposed.

3 Modelling Local Spatio-Temporal Activities

Location based social networks (LBSNs) allow people to share location based information (e.g., position, time, location description, etc.) with other users. In return they get incentives from the LBSN provider for being an active user in the community or they benefits from local shops for visiting them. While the user 'checks in' at a place (called venue) his location will be shown on his mobile phone and also be depicted to his friends. In addition, it is possible to rate venues and attach notes, pictures or other information to check-ins. Use cases for this new type of service cover a broad spectrum from exploration over recommendation to location-based gaming (Bawa-Cavia 2011).

Throughout this chapter we use data collected from one of the largest LBSNs, called Foursquare, with a community of more than 25 million users worldwide who produce over one million check-ins per day (October 2012).[3] This rich data source gives us the opportunity to analyse the spatio-temporal interaction between individuals and places and to project the results onto a crisp classification of local clusters, which are described by their activity profiles.

Before explaining our methodology, we hereby state our notation: V is a set of n_V Foursquare venues i, U is a set of n_U Foursquare users u, C is set of check-ins

[2] http://en.wikipedia.org/wiki/Whrrl (last visited: 14.11.2012).

[3] https://foursquare.com/about/ (last visited: 14.11.2012).

where each check-in c^t_{vu} consists of a venue v, a user u and a timestamp t. For two venues we can calculate the geographic distance $d(i, j)$ for all $i, j \in V$ using the given coordinates from i and j.

According to Andrienko et al. (2012) we address the uncertainties in spatio-temporal data by aggregation. Thus, in a first step aggregates are estimated for every venue i containing the count of all check-ins c_i per hour to get a vector of the hourly distribution of check-ins v_i, where the tth Element (written as v^t_i) represents the number of check-ins at the venue i at hour t with $t = 0 \ldots 23$—we call this the *activity profile* of a venue. The next preparation step is data filtering. Thus, we remove all venues that have less than two check-ins or two unique users.

Affinity Measure

For clustering the venues according to their activity profile, we define a similarity measure in the following way: Interpreting the profile of a venue as a short time series, we calculate a distance between two venues based on a comparison of the shapes of their profiles; see Todorovski et al. (2002). The idea is, that if v_i and v_j are similar, the change from t to $t + m$ should be similar for both venues for all possible values of t and m. For a chosen t and m this means a comparison of the shift from v^t_i to v^{t+m}_i with the shift from v^t_j to v^{t+m}_j. The possible values of the shift q among t and $t + m$ are qualitative, therefore $q\left(v^t_i, v^{t+m}_i\right)$ gets the label 'increase' if $v^t_i < v^{t+m}_i$; 'decrease' if $v^t_i > v^{t+m}_i$ and 'no-change' if both are the same. The similarity between two venues v_i and v_j with respect to a shift from t to $t + m$ is denoted as as $sim(q(v^t_i, v^{(t+m)}_i), q(v^t_j, v^{(t+m)}_j))$ and is defined in the following Table 1.

At last we define an affinity measure between two venues as

$$Aff\left(v_i, v_j\right) = \frac{\sum_{t<t'} sim\left(q\left(v^t_i, v^{t'}_i\right), q\left(v^t_j, v^{t'}_j\right)\right)}{NC}$$

where NC is the number of all possible comparisons (253 in our case).

Clustering Algorithm

For our analysis, we apply the spectral clustering as in Ng et al. (2001) because it finds arbitrarily shaped clusters and does not pose any constraints on them (in contrast to the k-means, for example, which assumes cluster to be convex). For

Table 1 Similarity of the shifts $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$ between the venues $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$

$sim(q_1, q_2)$		q_1		
		Increase	No-change	Decrease
q_2	Increase	1	0.5	0
	No-change	0.5	1	0.5
	Decrease	0	0.5	1

this, we follow the preparations given in Cranshaw et al. (2012) to create first the $n_V \times n_V$ affinity matrix

$$A = \left(a_{i,j}\right)_{i,j=1,\ldots,n_V} \text{ where:}$$

$$a_{i,j} = \begin{cases} Aff\left(v_i, v_j\right) + \alpha & \text{if } j \in N_m(v_i) \text{ or } i \in N_m\left(v_j\right) \\ 0 & otherwise \end{cases}.$$

Here $N_m(v_i)$ refers to the m nearest venues to v_i with respect to their distance d. The small constant α assures that every venue is connected to its neighbours (α was set to 0.01). The affinity matrix A together with the number of desired partitions k is given as input to the spectral clustering algorithm described in Ng et al. (2001). The result is a partition of all venues into k disjoint clusters $C_1, \ldots, C_k$ where every cluster C_i could be mapped on a subgraph $G(A_i)$ of graph $G(A)$ where A_i is a set of the corresponding vertices to C_i. As in Cranshaw et al. (2012), we apply a post-processing step to get spatially contiguous partitions. Here we replace every subgraph $G(A_i)$ set of clusters produced by splitting it into its connected components.

Evidence Accumulation Clustering (EAC)

There exist many different clustering algorithms each having its advantages like simplicity, a small number of parameters, the ability to identify arbitrary shaped clusters or the capability to handle large data sets (Jain et al. 1999). The difficulty is that every clustering algorithm and even any set of parameters will produce a somewhat different solution. This makes it hard to decide, which result should be kept. In our case, there is no prior knowledge about the number of clusters in the region under study. Consequently, it is not easy to estimate the right value for the parameter k. An approach to overcome this problem is called evidence accumulation clustering (EAC) and was proposed in Fred and Jain (2002). The notion behind this method is to build clusters with different algorithms and parameterizations and then to aggregate all solutions into one final partition using every single partition as a voting if instances should be placed together. If two venues will be placed together in most solutions, it is reasonable to assign them to the same cluster in the final partition. In this context, this method could also be understood as a tool to enhance the validity of the resulting partition by reducing the impact resulting from a single non-optimal clustering (like not well-separated clusters).

We adopt the idea of evidence accumulation clustering, to combine different runs of the spectral clustering, where k is sampled from the interval k_{min} to k_{max} in every single run. The algorithm is as follows:

Algorithm 1 Slightly modified version of the EAC Algorithm based on Fred and Jain (2002).

Input:	R—number of runs; k_{min} and k_{max}—interval for the possible values of k n_V—the number of venues
Output:	Final partition P^{final}
1.	Initialize the $n_V \times n_V$ association matrix $S = \left(s_{i,j}\right)_{i,j=1,\ldots,n_V}$, which will contain the "votes" of every iteration, with zero
2.	Do R times: 2.1 Randomly select k from the interval k_{min} to k_{max} 2.2 Run the spectral clustering with the chosen k and the pre-computed affinity matrix A to produce a partition P 2.3 Update the association matrix S according to the following rule: for every pair of venues (i, j) in the same cluster in P set $s_{i,j} = s_{i,j} + \frac{1}{R}$
3.	Transform the association matrix S into a distance matrix S' by $1 - S$
4.	Extract the final partition P^{final} with complete linkage clustering using the distance matrix S'. On the resulting dendrogram the final cutpoint is obtained by the highest 'cluster lifetime' which is the longest 'gap' between two successive merges

4 Experiment

4.1 Data Collection

The data used here contains Foursquare check-ins from May 2012 to October 2012. We collected all public check-ins found by searching the Twitter Timeline[4] for Foursquare tags by restricting them to the German language (for this we used the search options from Twitter). Then the information about all check-ins was gathered by querying the Foursquare API.[5] Afterwards we verified the country attribute of the check-ins and the extracted coordinates and discarded all check-ins not belonging to venues in Germany. In addition, venues were enriched with information using the Foursquare API again. Essentially, this information contains the hierarchy of the feature type used in Foursquare (e.g., Arts and Entertainment, Travel and Transport). Finally, the check-in ID, the user ID, the venue ID, the venue name, the check-in date and time, the main feature type of the venue and the coordinates for every check-in are known.

Our analysis focuses on the city of Cologne as an example for a major German city with more than 1 million inhabitants. After removing a few inconsistent items, the final dataset for Cologne consists of 11,890 check-ins from 2,093 users over more than 1,008 venues. The size of our dataset is considerably smaller than in

[4] https://dev.twitter.com/docs/using-search (last visited: 14.11.2012).

[5] https://developer.foursquare.com/ (last visited: 14.11.2012).

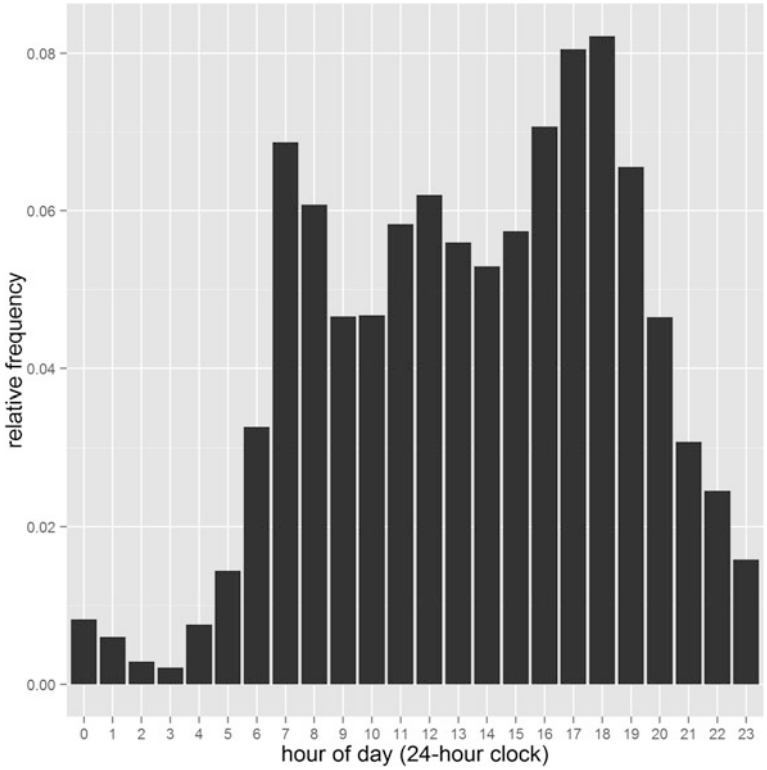


Fig. 1 Histogram of the check-ins per hour

other approaches [for example Cranshaw et al. (2012) and Long et al. (2012)] but most of them focus on major regions in the US with an disproportionate number of active users compared to the majority of the regions in Germany. So the question is raised if such micro-geographic approaches based on public data from LBSNs are also applicable in urban regions in Germany.

Figures 1 and 2 depict some general characteristics of the data. In Fig. 1 the hourly check-in frequency averaged over all days is plotted; clearly showing three peaks: commuter traffic in the morning, lunch time and the evening rush hour. This graphic could also be considered as the Foursquare perception of the activity profile of Cologne. Figure 2 shows the relative frequency for all main feature types where the check-ins are dominated by the categories 'Travel and Transport' followed by 'Food' and 'Professional and Other Places' (work).

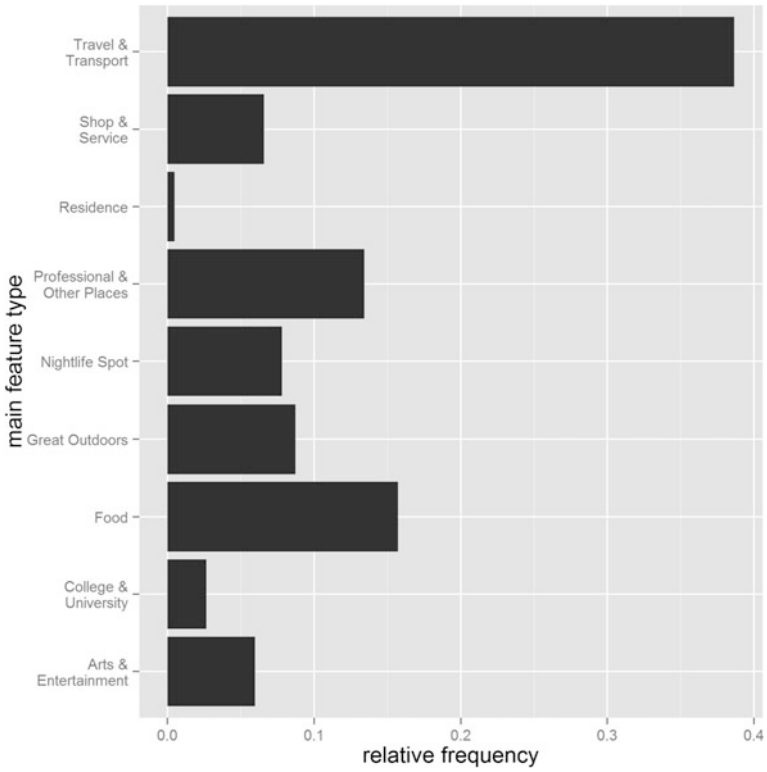


Fig. 2 Histogram of the check-ins per main feature type

4.2 Analysis and Results

Computing the Local Clusters

For carrying out our analysis, we used the following parameters: To build our affinity matrix we choose $m = 5$ (five nearest neighbours) which gave us mostly contiguous clusters. The EAC was carried out with $R = 2,000$. The proposed value ensured convergence for our analysis. We set $k_{min} = 5$ and $k_{max} = 60$ and thus accounted for both, small and big clusters, targeting the inherent uncertainty about the correct value for k.

Our first step after data preparation was to build up our affinity matrix A according to the given parameters. Secondly we ran the proposed EAC algorithm to compute the association matrix S. Figure 3 shows the results. Due to the maximal cluster lifetime of 0.057 we decided to take 31 as our value for the number of regions (this assured that the venues to be placed in the same final cluster must have been placed together in at least 82 % of all runs).

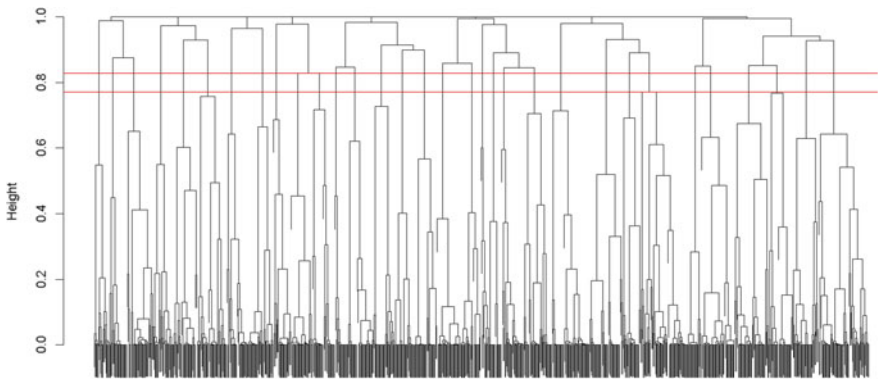


Fig. 3 Determining the number of clusters in the final partition through exploration of the resulting dendrogram (the maximal cluster lifetime of 0.057 is between the *two red lines*—see Algorithm 1 for the definition)

General Cluster Structure

The final clustering is displayed in Fig. 4. In favour of a clear overview, we focus the shown map extent to the inner city of Cologne (every cluster is represented by a different colour and the convex hull). It is clearly visible that the city centre consists of some small clusters while the surrounding districts are represented by large partitions. This is plausible since often city centres are functionally much more heterogeneous, while suburban areas are more homogenous (like large residential or industrial areas). While the clustering seems reasonable, we now want to show how each of the computed partitions describe some special topic depending on the activity profile.

Activity profiles

Subsequently, we compute the activity profile of a cluster, which is defined as the sum over all activity profiles of the corresponding venues. The next step is to subtract it from the overall activity profile after transforming all absolute frequencies into relative frequencies. This yields the graphic shown in Fig. 5 in which the difference to the regional activity profile is displayed for every cluster. If a bar has a value near zero, the activity in this cluster (which means check-ins) is very similar to the activity at a regional level with respect to that time slot.

There are some typical profiles like the one displayed for cluster 5, 15 and 18 which show the main hotspots of the Cologne's nightlife (Township 'Ehrenfeld', 'Belgisches Viertel' and 'Zülpicher Strasse'). While the number of check-ins is significantly less during the day, there is a lot more activity in the evening and at night. It is also typical that the last visitors of bars and clubs go home long time after midnight. It is interesting to see that cluster 2 (building the southern part of the region named 'Belgisches Viertel') also seems to have an attractive nightlife but contrary to the other nightlife hotspots, there are many individuals nearby

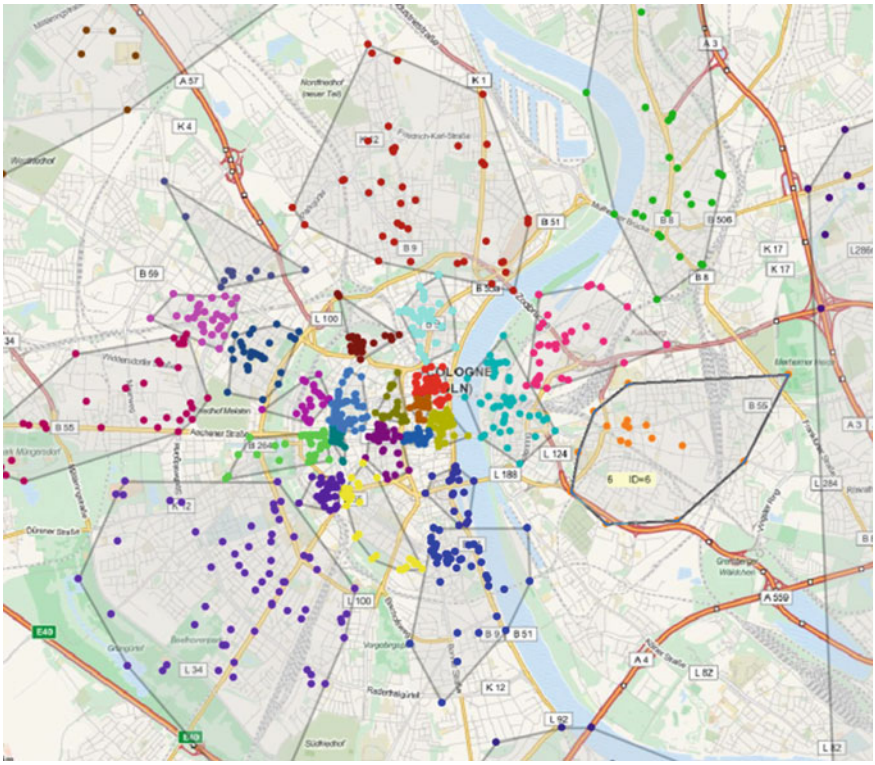


Fig. 4 Results from the clustering of Foursquare check-ins in Cologne city centre and surroundings

around lunchtime. On the other hand, there are also clusters showing completely different aspects of daily life. For example, cluster 1 is putting up an area around Cologne main station. It matches almost perfectly the overall activity profile (so the differences are near zero) because all of the mentioned peaks from Fig. 1 are in some kind related to aspects of public transport. For example, the clusters with IDs 19, 23, 24 and 30 are typical instances for work-related areas. There the main activity takes place during typical working hours between 10 and 15 o'clock.

To demonstrate the plausibility of our findings, we create a different plot (Fig. 6) showing the difference between the relative frequency of the main feature types between cluster and regional level. In favour of readability, we left out the label of every second feature type on the y-axis, which are the same like the corresponding one in Fig. 2. Here it is conspicuous that the clusters forming the nightlife also consist of an outstanding number of venues categorized as nightlife (compared to the average). For the cluster representing the area around Cologne main station venues from the categories 'Food' and 'Travel and Transport' are dominating. While the second category seems to be instantly intuitive, also the first one is plausible because most of the small shops at the main station

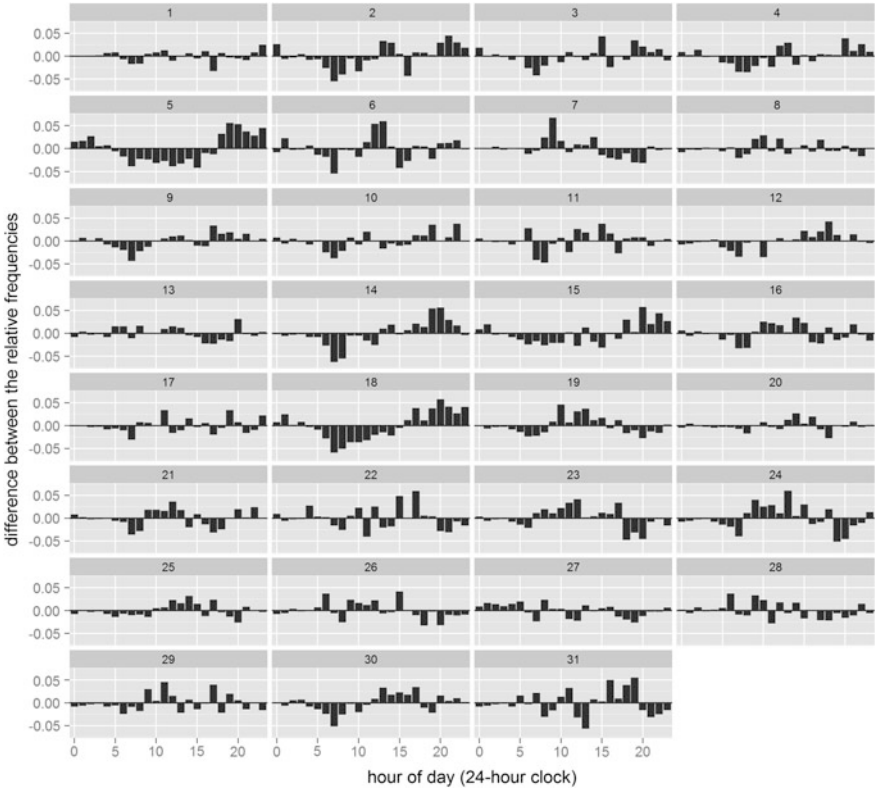


Fig. 5 Per cluster difference in the activity profiles between cluster and regional level—*the ordinate axis* displays the difference between the activity in the cluster and at the regional level for every hour

sell food and beverages. The description for the class of the work-related clusters is somewhat more heterogeneous—those venues often have the main feature type 'Shop and Service', 'Great Outdoors' (botanic garden and places to rest along the river Rhine), 'Professional and Other Places' and 'Food'—most places restricted to opening hours which explain the typical shape of the activity profile.

Classifying activity profiles

So far, we can create a description for every constructed cluster. Still an overall description of the entire region is missing, which relates all clusters according to their activity profile. This could also be considered as a clustering of the constructed partitions where partitions featuring the same profile should be kept together. To get an intuitive and visualizable solution we therefore choose Sammon's mapping (Sammon 1969), a method from multidimensional scaling which does a projection from the 24-dimensional space (the profile) into a space of lower dimensionality (we apply here a two-dimensional colour plane). Here the

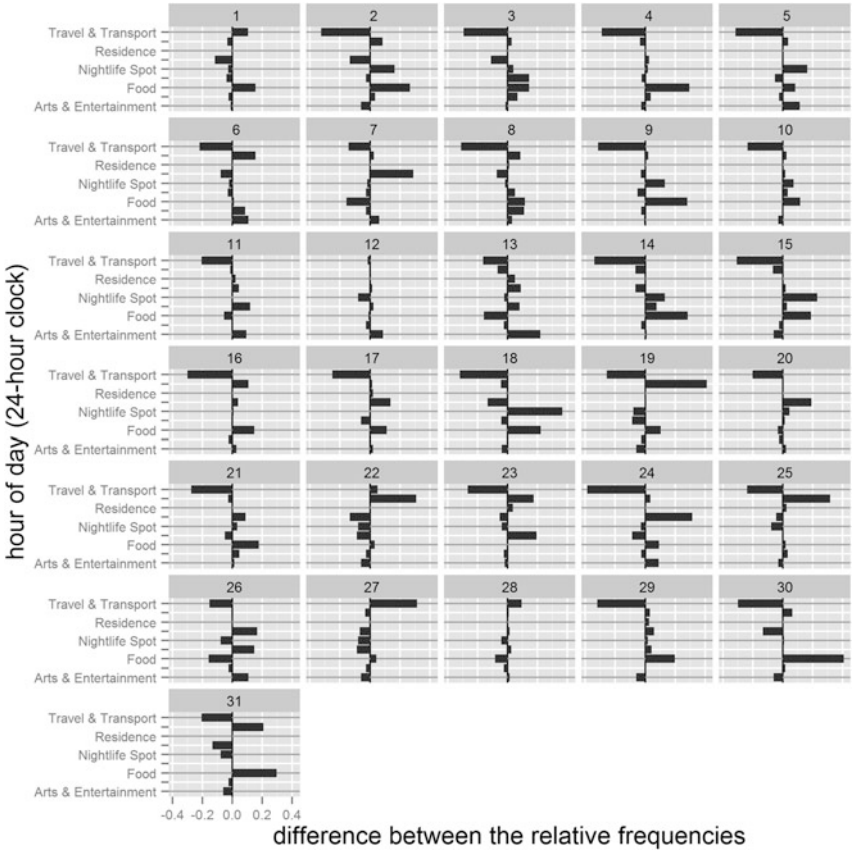


Fig. 6 Individual deviation of the distribution of the check-in category for every cluster from the average distribution of the check-in categories

similarity between two clusters with respect to their profiles is expressed through closeness. A good way to visualize this in terms of our solution is to choose the colouring according to the mapping, so that similar clusters get similar colours. For the discovered nightlife hotspots we depict this colouring in Fig. 7. The small graphic in the upper right corner shows a part of the results from the Sammon's projection focussing on the identified nightlife areas 5, 15 and 18 (coloured in green). As expected, they are all arranged close together. The adjacent clusters 4 and 14 also seem to have a significant number of venues belonging to nightlife. However, looking at the activity profile indicates that closing time is earlier and the nightlife aspect is less pronounced.

In summary the general classification can be described by the following rules of thumb: while green indicates an active nightlife (or activity in the evening), blue clusters are more often characterized by 'daylight activities' and red stands for partitions not differing too much from the average regional profile shown in Fig. 1.

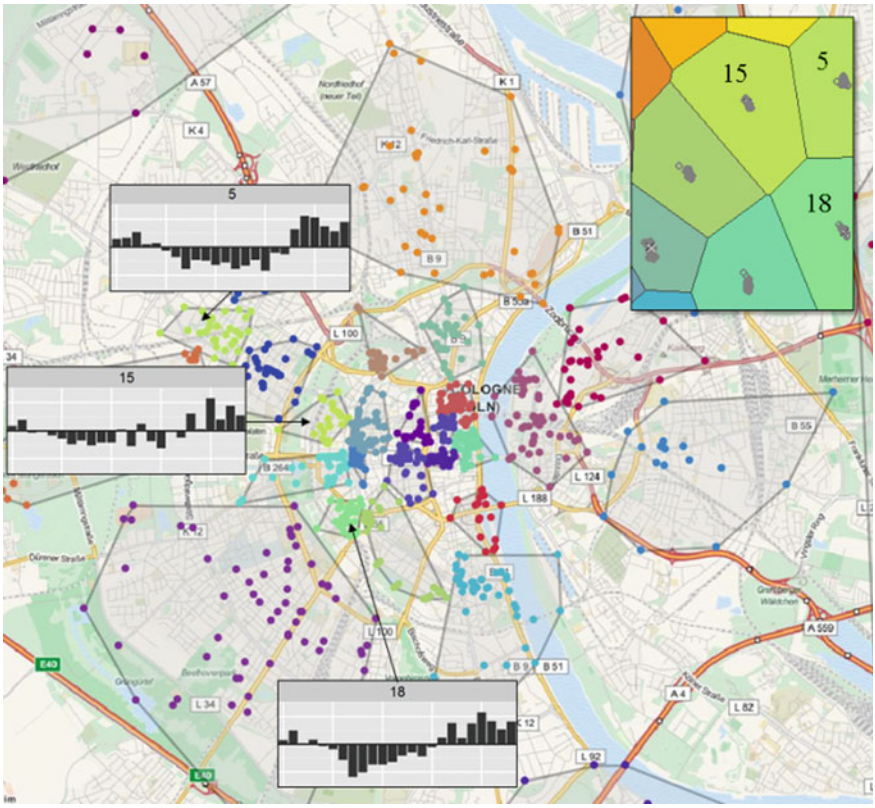


Fig. 7 Results from the Sammon's mapping for the inner city of Cologne

5 Discussion/Future Work

In this work, we presented a novel approach to identify, describe and classify areas according to the temporal distribution of individual human activities. Therefore, we utilized a new source of data obtained from LBSNs. A major advantage of using this type of information is its general availability. It consists of events (called check-ins) triggered by users and bundle spatial (the location), temporal (the time of the event) and personal (the user ID) information. An important difference to the usage of coordinates is, that the spatial information here is connected with a physical location (a venue) and attached with a human readable description. Thereby people visiting the same physical location (e.g., a restaurant or a shopping area) could be matched on the same coordinates.

We used this information to build an activity profile for every venue in the area around Cologne. The profile is based on a similarity measure, which is particularly suitable for short time series.

We then applied spectral clustering to obtain partitions that contain venues with similar activity profiles. To overcome the problems arising from the "instability" of clustering methods we used evidence accumulation clustering to deduce the parameters needed directly from the data (in this case the number of clusters).

Then our method proposed for classifying the partitions is straightforward. We computed the difference between the activity profile of the partition and the activity profile of the regional level for every cluster and visualized all differences. Especially partitions with the emphasis on nightlife and workplace could easily be identified. Furthermore, we classified every cluster using a technique from the field of multidimensional scaling to obtain an intuitive visualization. Therein the similarity between clusters is expressed through similar colours.

Based on this we showed how to explore regions of similar activity and how to characterize them by colour. Because of the promising results, we think that the overall approach could be a starting point for a better understanding of urban dynamics.

There are three main findings resulting from this work: (1) The construction of activity profiles for every location allows a clustering based on the temporal distribution of the venues. It thereby features a description of the spatio-temporal structure of a city. (2) We show how a new combination of existing approaches allows the creation of local and contiguous clusters without suffering from problems like the uncertainty about the right parameter values or assumptions about the cluster type. (3) To the best of our knowledge, this is the first approach using LBSN data for micro geographic modelling in the largest country in Europe. We conclude that while the size of the obtained data is still small compared to studies in the US, it is nevertheless possible to feature a much better understanding of social processes and interactions in urban regions. The outlook is even better because there is an on-going increase in the use of mobile devices and LBSNs.

In terms of future work, we intend to focus on three main aspects from which we expect the most benefits.

The first one is the integration of different sources of data coming from LBSNs and microblogging services, which will be a challenging task. Especially matching venues from the different 'ecosystems' is clearly non-trivial. This is because there is no standardization concerning the names, feature types or localization. On the other hand, it will directly increase the sample size and thus likely improve results. Additional data from microblogging services like Twitter could fill gaps especially in regions where the usage of LBSNs is not sufficient and vice versa.

The second direction for future work is the exploration of general limitations by using Volunteered Geographic Information (VGI) (Goodchild 2007) for urban analysis. For example, it is well known that this data does not provide a representative sample from the whole population. This can be important when interpreting the results from our method (and of course the work from others when using VGI data). Mentioned in the part of related work, a possibility to overcome this situation could be an approach that uses different sources of mobility data. The fusion of data from cell phones, surveys and VGI will provide valuable information for analysing activity patterns and possibly enhances the transformation

from urban regions to 'smart cities'. In this case the public discussion about privacy should be taken into account as well.

The third and last direction will focus on the algorithmic parts, particularly the extension of the classification. For example, the method could support the interpretability by providing an intuitive and automatic description for every cluster and interactive tools to let the user dive into the results. In addition, the computation should be made 'big-data-ready' to cope with the massive amount of data produced in LBSNs and micro-blogging services.

References

Andrienko N, Andrienko G, Stange H, Liebig T, Hecker D (2012) Visual analytics for understanding spatial situations from episodic movement data. KI - Künstliche Intelligenz 26:241–251 Springer

Aubrecht C, Ungar J, Freire S (2011) Exploring the potential of volunteered geographic information for modeling spatio-temporal characteristics of urban population. In: Proceedings of the 7th international conference on virtual cities and territorie. 7VCT '11, Lisbon, pp 57–60

Bawa-Cavia A (2011) Sensing the urban: using location-based social network data in urban analysis. In: The 1st workshop on pervasive urban applications. PURBA '11, San Francisco

Chen LJ, Li CW, Huang YT, Shih CS (2011) A rapid method for detecting geographically disconnected areas after disasters. In: IEEE international conference on technologies for homeland security. HST '11, Greater Boston, pp 501–506

Cheng Z, Caverlee J, Lee K, Sui DZ (2011) Exploring millions of footprints in location sharing services. In: The social mobile web. ICWSM '11, Barcelona

Cho E, Myers SA, Leskovec J (2011) Friendship and mobility: user movement in location-based social networks. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining. KDD '11, New York, NY, USA, ACM (2011), pp 1082–1090

Cranshaw J, Schwartz R, Hong JI, Sadeh N (2012) The livehoods project: utilizing social media to understand the dynamics of a city. In: To appear in the 6th international AAAI conference on weblogs and social media. Dublin, Ireland

De Longueville B, Smith RS, Luraschi G (2009) Omg, from here, i can see the flames!: a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In: Proceedings of the 2009 international workshop on location based social networks. LBSN '09, New York, NY, USA, ACM (2009), pp 73–80

Fred ALN, Jain AK (2002) Evidence accumulation clustering based on the K-Means algorithm. In: Proceedings of the Joint IAPR international workshop on structural, syntactic, and statistical pattern recognition, London, UK, Springer, pp 442–451

Gao H, Tang J, Liu H (2012) Exploring social-historical ties on location-based social networks. In: Breslin JG, Ellison NB, Shanahan JG, Tufekci Z (eds) ICWSM. The AAAI Press, California

Goodchild MF (2007) Citizens as sensors: the world of volunteered geography. GeoJournal 69(4):211–221 Springer

Hong L, Ahmed A, Gurumurthy S, Smola AJ, Tsioutsiouliklis K (2012) Discovering geographical topics in the twitter stream. In: Proceedings of the 21st international conference on World Wide Web. WWW '12, New York, NY, USA, ACM (2012), pp 769–778

Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 31(3):264–323

Jiang S, Ferreira Jr J, Gonzalez MC (2012) Discovering urban spatial-temporal structure from human activity patterns. In: Proceedings of the ACM SIGKDD international workshop on urban computing. UrbComp '12, New York, NY, USA, ACM (2012), pp 95–102

Jin L, Long X, Joshi JB (2012) Towards understanding residential privacy by analyzing users' activities in foursquare. In: Proceedings of the 2012 ACM workshop on building analysis datasets and gathering experience returns for security. BADGERS '12, New York, NY, USA, ACM (2012), pp 25–32

Joseph K, Tan CH, Carley KM (2012) Beyond "Local", "Categories" and "Friends": Clustering foursquare users with latent "Topics". In: Proceedings of the 2012 ACM conference on ubiquitous computing. UbiComp '12, New York, NY, USA, ACM (2012), pp 919–926

Kindberg T, Chalmers M, Paulos E (2007) Guest editors' introduction: urban computing. Pervasive Comput IEEE 6(3):18–20

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. CHI '11, New York, NY, USA, ACM (2011), pp 2409–2418

Long X, Jin L, Joshi J (2012) Exploring trajectory-driven local geographic topics in Foursquare. In: Proceedings of the 2012 ACM conference on ubiquitous computing. UbiComp '12, New York, NY, USA, ACM (2012), pp 927–934

Ng AY, Jordan MI, Weiss Y (2001) On spectral clustering: analysis and an algorithm. Adv Neural Inf Process Syst 2:849–856 MIT press

Noulas A, Scellato S, Mascolo C, Pontil M (2011) An empirical study of geographic user activity patterns in Foursquare. In: Proceedings of the 5th International AAAI Conference on weblogs and social media. ICWSM '11, Barcelona, pp 570–573

Noulas A, Scellato S, Mascolo C, Pontil M (2011) Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. In: The social mobile web. ICWSM '11, Barcelona

Pontes T, Vasconcelos M, Almeida J, Kumaraguru P, Almeida V (2012) We know where you live: privacy characterization of Foursquare behavior. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing. UbiComp '12, New York, NY, USA, ACM, pp 898–905

Reades J, Calabrese F, Sevtsuk A, Ratti C (2007) Cellular census: Explorations in urban data collection. Pervasive Comput IEEE 6(3):30–38

Sammon JW (1969) A nonlinear mapping for data structure analysis. IEEE Trans Comput 18(5):401–409

Shimada K, Inoue S, Maeda H, Endo T (2011) Analyzing tourism information on twitter for a local city. In: 1st ACIS international symposium on software and network engineering. SSNE '11, pp 61–66

Thom D, Bosch H, Koch S, Worner M, Ertl T (2012) Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages. In: Proceedings of the Pacific visualization symposium. PacificVis'12, IEEE Press, pp 41–48

Todorovski L, Cestnik B, Kline M, Lavrac N, Dzeroski S (2002) Qualitative clustering of short time-series: a case study of firms reputation data. Helsinki University Printing House, Helsinki, pp 141–149

Ye M, Janowicz K, Mülligann C, Lee WC (2011) What you are is when you are: the temporal dimension of feature types in location-based social networks. In: Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems. GIS '11, New York, NY, USA, ACM (2011), pp 102–111

Part II
Remote Sensing

Automatic Extraction of Complex Objects from Land Cover Maps

Eliseo Clementini and Enrico Ippoliti

Abstract The ESA Support to Topology (STO) project addressed the problem of extracting so-called complex objects, intended as particular land use elements (urban fabric, industrial units…) from land cover maps, by means of topological relations among the different land cover objects. We developed an approach to give a semantic characterization to complex objects. Based on that, we developed a functional strategy to identify complex objects in an image and to build a visual representation compatible with the scale and resolution of the original map. The spatial operators, not only topological but directional and metric as well, were either taken from already existing systems or specifically implemented for the study. The developed approach and prototype web-GIS system, named Topology Software System (TSS), have been validated through several use cases, run by specialized end-users, in order to verify that the expected operations could be performed.

1 Introduction

Starting from 2000, object based image analysis (OBIA–or GEOBIA for geospatial object based image analysis) had a big development (Malinverni et al. 2010; Novack et al. 2010; Thunig et al. 2010). The objective of the OBIA approach is to develop a methodology for automated or semi-automated classification of geographical elements or complex physical features of Earth land cover

E. Clementini (✉) · E. Ippoliti
Department of Industrial and Information Engineering and Economics, University of L'Aquila, Via G. Gronchi 18, 67100 L'Aquila, Italy
e-mail: eliseo.clementini@univaq.it

E. Ippoliti
e-mail: enrico.ippoliti@bluedeep.it

D. Vandenbroucke et al. (eds.), *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography,
DOI: 10.1007/978-3-319-00615-4_5,

(Baltsavias 2004; Barnsley et al. 2001; Hussain et al. 2007; Liu et al. 2008; Wijnant and Steenberghen 2004). This includes principles using multi-resolution object-oriented approaches like segmentation, object parameterization and classification that make use of combined spectral, textural, shape and contextual object features (Debeir et al. 2002; Friedl and Brodley 1997). Typical software used in GEOBIA are Trimble eCognition (Trimble 2013), Feature analyst (Overwatch 2013), ENVI Feature Extraction Module (Exelis 2013). The GEOBIA field has been recognized as a bridge between classical remote sensing image analysis and the Geographical Information Systems (GIS) field. It seems that to-date this integration did not fully take place, since typical spatial analysis GIS methods are not always used in Earth Observation (EO) image analysis.

Nowadays, many public and private agencies, like the Environmental Protection Agency of Austria in the LISA (Land Information System Austria) project (Grillmayer et al. 2010; Land Information System Austria 2012; Prüller et al. 2011; Weichselbaum et al. 2009), use a set of comprehensive automated and manual approaches, based on expert rules using geospatial data from various themes and classic photo-interpretation techniques, to derive land use information from land cover maps. They use ancillary data as well, that is, data coming from different sources from EO images. These methodologies are expensive, time consuming and subjective. In other projects, semi-automatic procedures are applied: for instance, to produce GMES Urban Atlas maps (Atlas 2012), image analysis packages such as eCognition are utilized. Automatic processing techniques may reduce the time employed for manual interpretation, satisfying current demands for continuous and precise data that accurately describes the territory.

A semantic gap exists between the features resulting from typical classification methods and real complex objects. The latter ones have a meaning that can be represented by a network of semantic relations, expressing both the spatial and thematic component. Our aim is to make explicit such knowledge and come up with a complex object definition (COD) that can be used to automatically identify the object in a land cover map and find a visual representation at the same scale. The proposed methodology takes advantage of a taxonomy of spatial operators (some of them are already available in current GIS analysis tools and some needed to be implemented from scratch). We took as input data classified images (land cover), coming from existing databases, such as those of various national agencies, and defined a hopefully automatic procedure for the identification of complex land use objects based on contextual rules. Preliminary results of this methodology were presented in Ippoliti et al. (2012a, b); Natali et al. (2012).

Advantages of the developed approach with respect to existing methods can be summarized as follows:

- once the ontological part (spatial rules) is defined, the process is automatic;
- the process can be carried out from land cover data without a costly integration with other data sources;
- direct use of vector data in standard OGC format (OGC 2011), which facilitates the integration with other systems;

- capability of modelling complex objects with a rich internal structure, made of parts and subparts;
- independence from the graphical representation: the same complex object can have different graphical representations, depending on context and scale.

The remainder of the chapter is structured as follows. In Sect. 2, we briefly illustrate the use cases that were adopted in our project. In Sect. 3, we illustrate the methodology for complex object definition referring to the specific use case of urban settlements. In Sect. 4, we discuss the set of spatial operators that are at the core of the methodology, distinguishing between operators for geometry characterisation and operators for geometry transformation. In Sect. 5, we discuss the adopted web-GIS architecture and evaluate the results of use cases from a statistical point of view. Section 6 provides short conclusions.

2 Description of Use Cases

Remote sensing imagery needs to be converted into tangible information which can be utilised in conjunction with other data sets, often within widely used Geographic Information Systems (GIS). Land cover is the observed (bio)physical cover of the earth's surface. Land use is characterized by the arrangements, activities and inputs that people undertake in a certain land cover type to produce, change or maintain it (Di Gregorio and Jansen 2000).

The Land Information System of Austria (LISA) project aims at modelling the Austrian environment, offering an "Object-Oriented" (OO) data model to represent Land Cover and Land Use objects. The LISA data model provides 14 classes to represent Land Cover categories (e.g., buildings, built-up areas, rocks…); Land Cover is directly derived from EO data (Fig. 1). It provides 25 classes to represent Land Use categories (e.g., settlement, traffic, agriculture, forestry…); Land Use is derived using additional spatial data, such as:

- Spatial planning (land use zoning plans);
- Street maps;
- Agricultural information system;
- Water information system.

We considered other use cases about Urban Atlas generation (Fig. 2). In Urban Atlas as well, land use maps are obtained in two steps: from the source EO image, the first step is the obtainment of a pre-classified map with land cover surfaces or biophysical parameters (MMU $\approx$ 25 m^2). In the following step, the resulting Urban Atlas database (MMU = 0.25 ha) is generated.

Some land use objects that need to be recognized are residential urban settlements, industrial or commercial urban settlements, roads, airports, agricultural farmlands, and river basins (Fig. 3).

Fig. 1 Data from the land information system of Austria: orthophoto, land cover, and land use

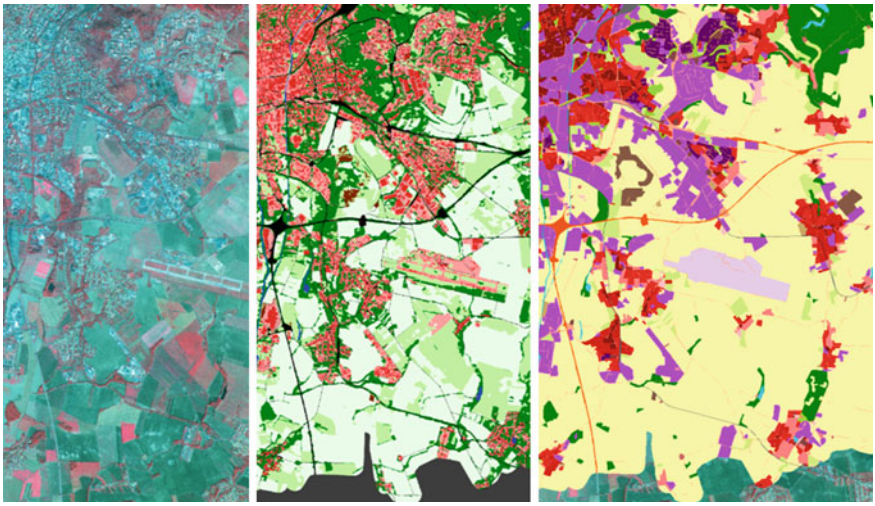


Fig. 2 Data from urban atlas

3 Complex Objects Definition

Complex objects can be recognized by observing their spatial structure. For example, airports are characterized by runways, which geometrically are of elongated shape, wider than a normal road and truncated at the ends. Nuclear

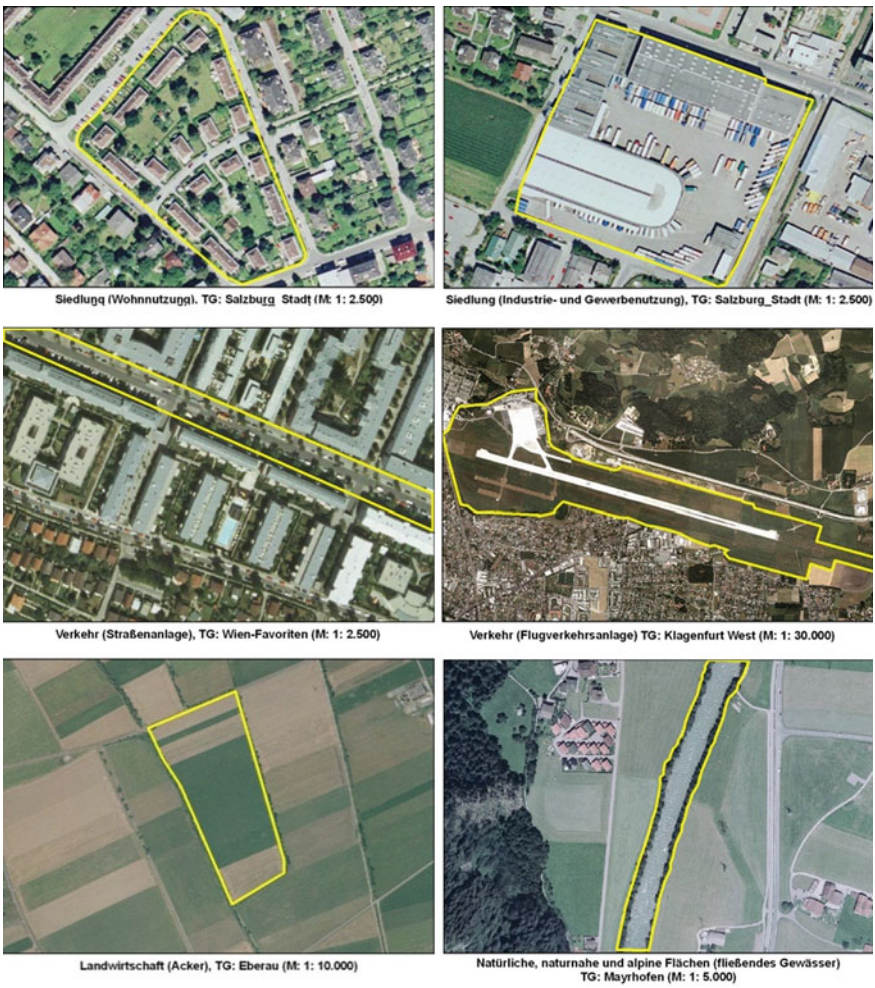


Fig. 3 Complex objects: residential area, industrial or commercial areas, roads, airports, agricultural farmlands, and river basins

plants are characterized by the presence of water basins and round towers. Artificial channels can be distinguished from rivers by the presence of straight boundaries versus round-shaped boundaries.

Let us concentrate on residential urban settlements. How an urban settlement is defined? From user requirements (cartography experts) we could define a set of rules that identify the object. So, an urban settlement of residential type is defined as:

- A group of buildings;
- Each building should be smaller than a certain size (otherwise the use of the building would be most likely non-residential: commercial or industrial use);
- Small parts of different land cover (high and low vegetation, water, bare soil) connected to the buildings should be part of an urban settlement;
- Narrow segments of roads passing through the group of houses should be aggregated to the complex object, and parking and cul-de-sac as well;
- Main roads should separate the urban settlements;
- Other larger areas (woods, bare soil, and so on) should delimit the complex object as well.

The definition of a complex object is essentially a combination of constraints, both thematic and geometric. To identify the simple objects that are part of the complex object, we apply various spatial operators of Boolean result, e.g., an operator to check whether two objects are touching each other. To build a representation of the complex object, we apply a series of spatial operators to transform the geometries, e.g., a merge operator to combine two simple objects and a split operator to take a piece of a larger object. Let us consider the following procedure to build urban settlements:

1. Start from a seed object (a given building) (Fig. 4a);
2. Finding the neighbouring objects (of given land cover classes) (Fig. 4b);
3. Repeat the previous step with other neighbouring objects;

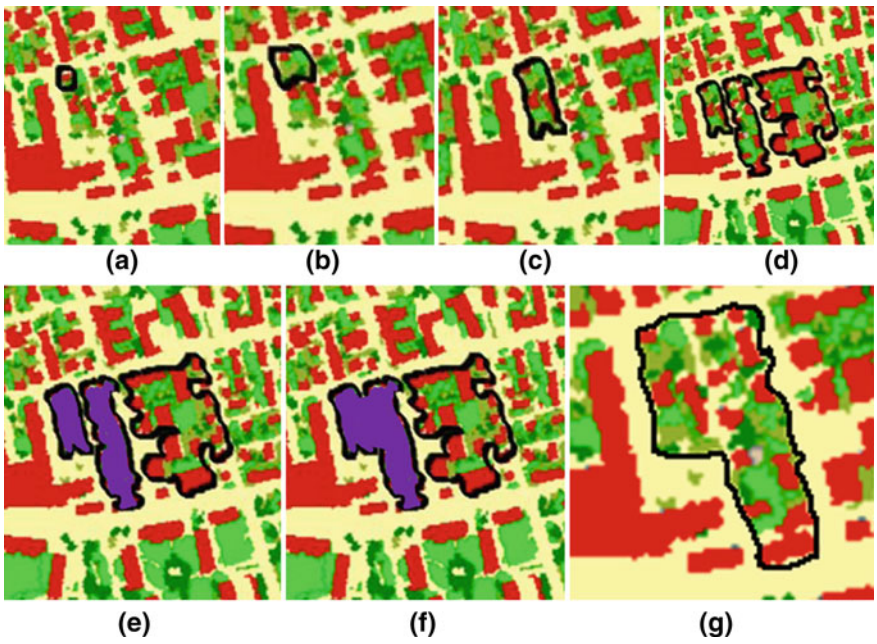


Fig. 4 Procedure to build urban settlements

4. Stop when the aggregate is entirely surrounded by other constructed areas (roads, parking,…) (Fig. 4c);
5. Repeat previous steps with other buildings not previously considered (Fig. 4d);
6. Group the objects found till now in such a way there exist pairs of neighbouring objects that are at a distance less than a given threshold (this means that they are separated by a secondary road) (Fig. 4e);
7. Connect the groups of objects previously identified by some corridors (Fig. 4f);
8. Filter the result to remove small holes and concavities (internal roads and parking) (Fig. 4g);
9. From the set of results, eliminate objects that have a size below a given minimum mapping unit.

Procedures as the one illustrated above can be considered as sequences of functions (see Fig. 5). The functions that can be identified are four: "aggregate", "group", "refine", "validate". The function "aggregate" is used to build an aggregate of simple objects that satisfy precise topological relations. The function "group" is used to group together objects that satisfy specific distance criteria. The function "refine" is used to filter out small parts, such as separations, concavities, and holes, obtaining a smooth shape. The "validate" function is used to exclude from the results the complex object candidates that do not satisfy the minimum mapping unit.

Such functions can be reused in other contexts as well. For this reason, we designed them to accept several parameters. For instance, the function "aggregate" may be invoked by changing the seed land cover class, the size and other geometric properties of simple objects to be aggregated, the spatial relations to be satisfied by simple objects with seed object. In this way, the same function can be reused to perform various kinds of aggregations.

4 Taxonomy of Spatial Operators

The spatial operators that we adopted to provide an operational framework are divided into two groups: those related to *geometry characterization* and those related to *geometry transformations*. In the first group, mainly Boolean operators are considered: they are used to check various spatial properties of objects to find the ones that obey the complex object definitions. The second group relates to various geometric construction operators that are used to obtain an appropriate visual representation of a complex object. Such a visual representation depends on scale and context. Once a complex object has been identified, we can envisage various representations at different levels of resolution emphasizing different aspects depending on context. In essence, the visual appeal of the resulting map showing complex objects can be improved by the right choice of geometric operators (Fig. 6).

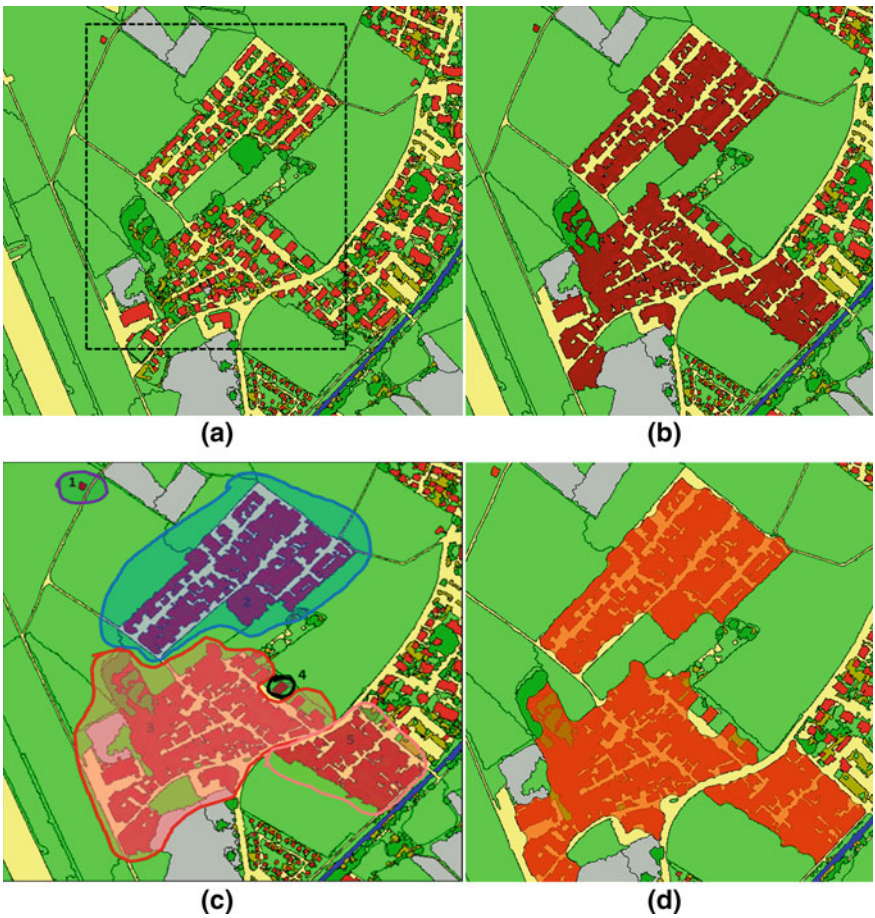


Fig. 5 Original data (**a**), and the result of the application of functions "aggregate" (**b**), "group" (**c**), and "refine" (**d**)

The set of Boolean operators are based on an ontology of spatial relations (Bucher et al. 2012; Clementini and Laurini 2008). According to it, spatial relations can be categorized following six orthogonal axes: the level of representation, the geometrical properties of space, the cardinality of relations, the granularity, the type and size of objects, and the dimension of the embedding space (Fig. 7).

Regarding the levels of representation, spatial relations can be categorized according to three levels: the geometric level, the computational level, and the application level. The geometric level is an abstract representation in terms of mathematical objects, where the spatial relations between objects are defined by specific geometric properties: for example, in the model of four intersections (4IM) (Egenhofer and Franzosa 1991), the topological relations are defined by the empty and non-empty values of the intersections of boundaries and interiors of the two

GO.01.	Geometry characterization	
	GO.PI.01.	Size and Shape (of single object)
	GO.PI.02.	Binary spatial relations
	GO.PI.03.	N-ary spatial relations
	GO.PI.04.	Network analysis
	GO.PI.05.	Validate planar subdivisions
GO.02.	Geometry transformation	
	GO.GT.01.	Skeleton
	GO.GT.02.	Buffer
	GO.GT.03.	Simplification
	GO.GT.04.	Generalization
	GO.GT.05.	Container
	GO.GT.06.	Network transformation
	GO.GT.07.	Planar subdivision enforcement
	GO.GT.08.	Set operations

Fig. 6 Taxonomy of spatial operators

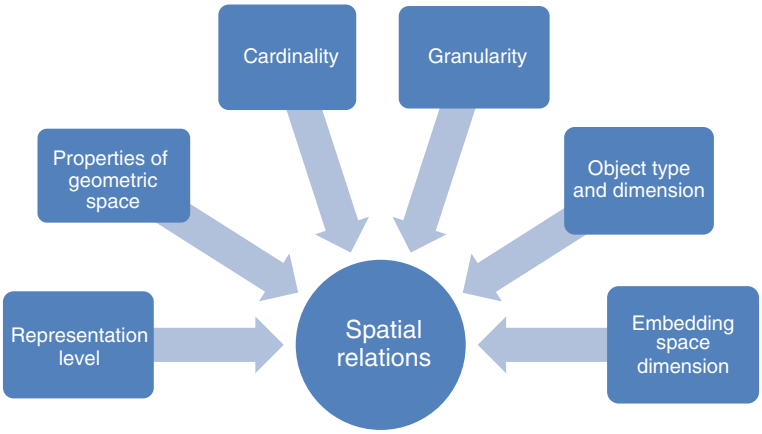


Fig. 7 Categorization of spatial relations

objects. The geometric level can be considered as the most primitive level for the study of spatial relations, since it allows finding formal definitions. The other two levels always relate to the definition of spatial relations at the geometric level.

At the computational level, spatial objects are represented as spatial data types and spatial relations between objects correspond to spatial operators. In essence, it is the level of spatial relations as supported by a database system. Defining relations at the application level may require defining what kind of user will perceive the relations. At this level, relations can be seen as semantic descriptions of underlying spatial properties [see also Clementini (2010); Klien and Lutz (2005); Tarquini and Clementini (2008)].

Regarding the properties of geometric space, we will refer to a commonly recognized categorization of spatial relations in three geometric kinds, topological, projective, and metric, that are based on the properties of topological space,

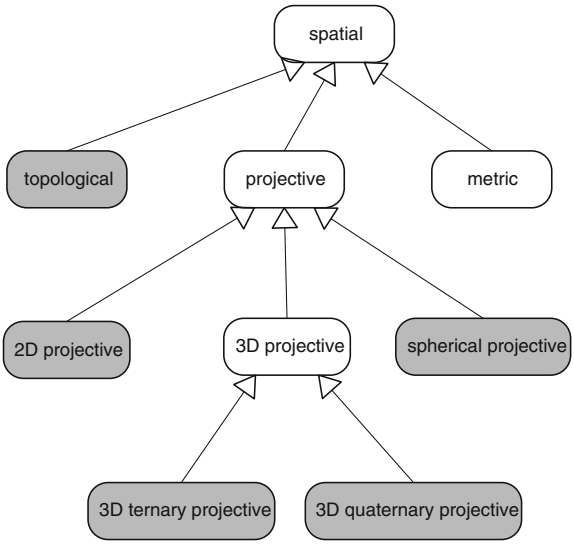


Fig. 8 Classification with respect to geometric space and dimension

projective space, and Euclidean space, respectively (Clementini and Di Felice 2000) (see also Fig. 8). Topological relations have been widely discussed in the literature [e.g., (Clementini et al. 1993; Cohn et al. 1997; Egenhofer and Herring 1991)] and implemented in spatial standards (ISO 2010; OGC 1999) (see also Fig. 9), while the other two kinds are the object of more recent research.

Projective relations are a category of spatial relations that can be described by projective properties of the space without resorting to metric properties (Billen and Clementini 2004) (Fig. 10). Like topological relations, projective relations are qualitative in nature because they not need exact measures to be explained (Egenhofer and Mark 1995). Also, projective relations are more specific than topological relations and can serve as a basis for describing relations that are not captured by topology. Standing at an intermediate step between metric and topology, projective relations are as much varied as "right of", "before", "between", "along", surrounded by", "in front of", "back", "north of", "east of", and so on. While specific models have been developed for particular sets of projective relations, such as cardinal directions (Frank 1992), orientation relations (Hernández 1993), cardinal directions for extended objects (Goyal and Egenhofer 1997), there is the need of a unifying model that is able to represent all variations of projective relations (Clementini 2012). Regarding metric relations, such as the distance between two points, they are normally intended as quantitative relations, though in our approach we see them mainly as qualitative relations (Clementini et al. 1997).

The other main group of spatial operators is the one labelled as "Geometry Transformation". We took advantage of several operators already available in various spatial analysis packages, such as JTS or GeoTools, though several useful

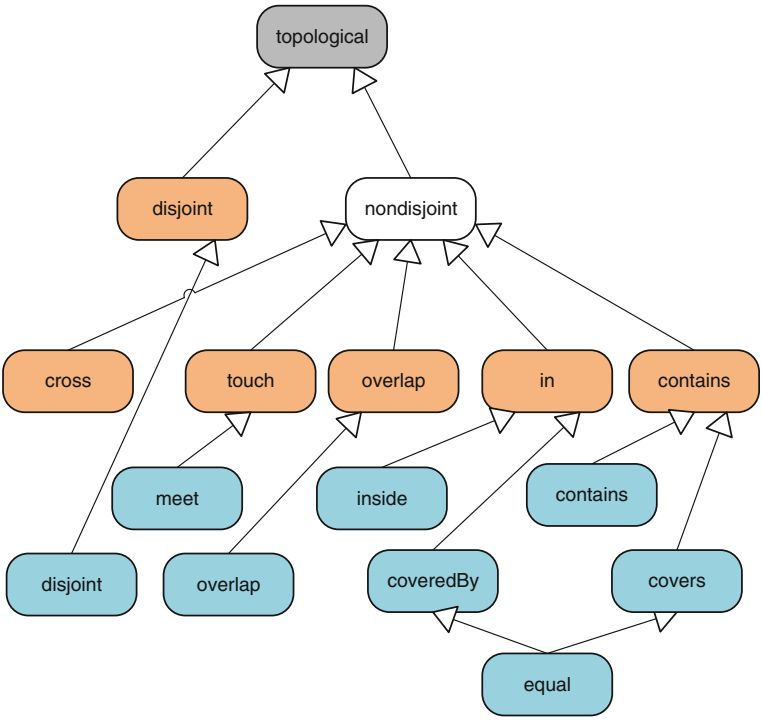


Fig. 9 Classification of topological relations

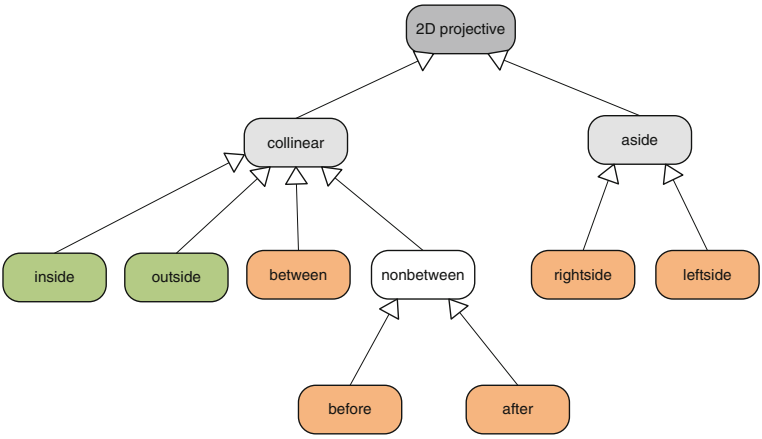


Fig. 10 Classification of 2D projective relations

operators are not included in those tools. For reasons of space, we will name just a few of the newly implemented operators: the "elongated", the "sameShapeOrientation", and the "fusion" operator.

boolean **isElongatedInShape**(Geometry *inputGeometry*, Double *tolerance*)
1 - Compute the *inputGeometry*'s MBR
2 - Compute the ratio between the perpendicular edges of the MBR
2.1 - if ratio >= *tolerance*→ return *true*
else return *false*

Fig. 11 The "elongated" operator

The operator "elongated" is able to evaluate the qualitative elongatedness of a shape (Fig. 11). The operator "sameShapeOrientation" is able to evaluate whether two elongated shapes have the same qualitative orientation (Fig. 12). Specifically, the two input geometries are retained to have the same orientation if their angular difference is less than $\pi/8$ (Fig. 13).

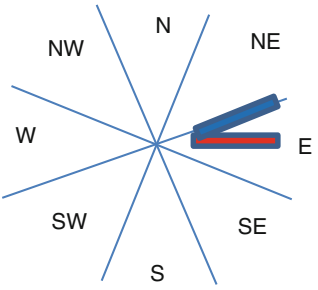
The operator fusion is applied to a group of disconnected objects in which each component has at least one nearest neighbor at a distance of less than a given threshold. The operator produces an aggregated object where components are joined together by adding an amount of outer space (see Fig. 14). We envisaged several techniques for fusion, by varying the amount of outer space that was attached to the resulting object: for option 0, components are joined by corridors obtained by computing the convex hull of the neighboring parts of boundaries that

boolean **sameShapeOrientation**(Geometry *firstInputGeometry*, Geometry *secondInputGeometry*, Double *elongatedTolerance*, Double *tolerance*)

1. Check whether the *firstInputGeometry* and the *secondInputGeometry* are "elongated in shape"
2. Compute the *firstInputGeometry*'s and *secondInputGeometry*'s MBRs (MBR1 and MBR2)
3. Compute the angles α1 and α2 between the longest edges of MBRs and x-axis
4. If difference(α1, α2) <= π/8 (+ *tolerance*) OR 7π/8 (- *tolerance*)<= difference(α1, α2) <= π
   →return *true*
   else return *false*

Fig. 12 The "sameShapeOrientation" operator

Fig. 13 At most $\pi/8$ difference corresponds to same orientation



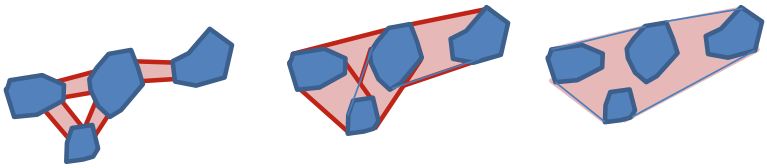


Fig. 14 Application of different versions of the fusion operator

are below the distance thresholds. For option 1, we joined pairs of nearest neighbor objects by their convex hulls. For option 2, we considered the convex hull of the entire group. This sequence of options allows us to obtain various graphical representations for the resulting complex objects (see Fig. 15).

5 Experiments

A prototype Web-GIS system, named Topology Software System (TSS), is implemented to allow users to define combinations of existing spatial functions (aggregate, group, refine, and validate), in order to identify specific land use features.

Fig. 15 Test of the application of the fusion operator from initial image with option 0, 1, and 2, respectively

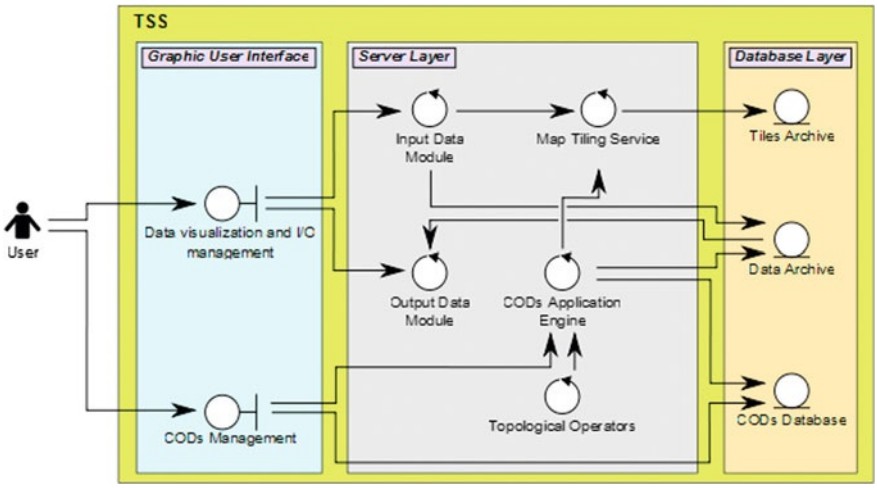


Fig. 16 The three schematic layers of TSS, including the main modules that compose each layer

The architecture of TSS is organized in three main layers: the Graphic User Interface (GUI) layer, the Server layer and the Database layer. Specific modules have been identified for each layer, in order to satisfy all user and system requirements (see Fig. 16).

The GUI layer represents the front-end of TSS to the user. Two main interfaces are available: a Visualization interface (for the input land cover maps and for the visualization of the identified land use objects) and the Complex Object Definitions (CODs) management interface for CODs management (creation, retrieval of existing CODs, manipulation and storage) and the application of a COD to a subset or to the entire input dataset. The COD specifies the land use object name and the combination of spatial functions that need to be applied in sequence to identify it. The server layer is the core of TSS: it contains all the modules for I/O processes, and for the application of CODs to the input dataset; the topological operators module contains all the developed topological operators and functions organized as a library: each module is called by the so-called CODs application engine, that manages the input data and the resulting complex object layers, as well as the successive application of the different functions. The database layer contains three main databases: the data archive, which hosts the input datasets and the created complex object layers, the CODs database, which stores the already developed complex objects definitions, and the tiles database, which contains all the tiles for each input dataset as well as those for the output layers. We decided to use a standard web map tiling service to make the input and output data visualization fast and multi-resolution.

The developed approach and implemented TSS tool have been validated through four use cases, run by specialized end-users, in order to verify that the expected operations can be performed, and that these operations are sufficient to

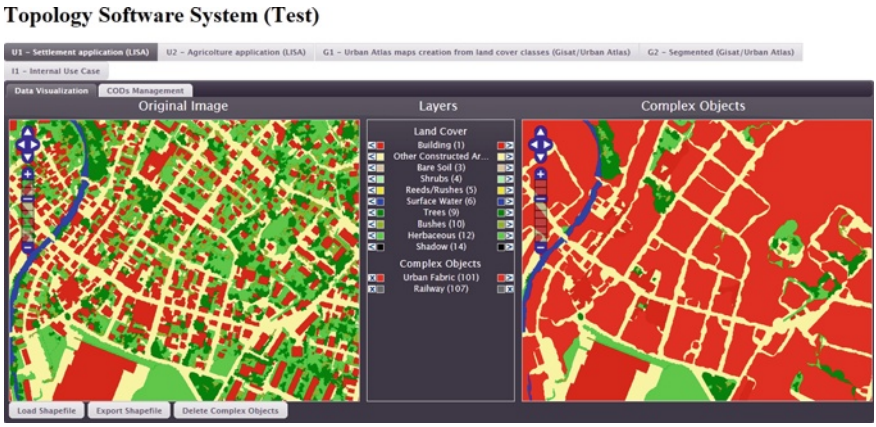


Fig. 17 Test on residential urban settlements

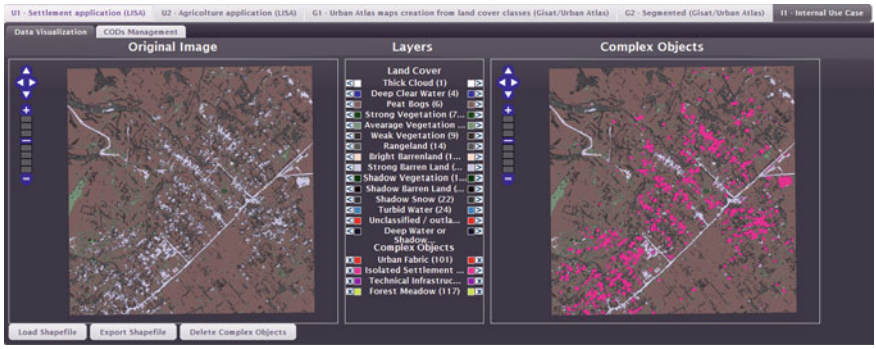


Fig. 18 Test on destroyed buildings

create land use maps from land cover maps. In Fig. 17, the TSS system shows the results after having applied a procedure to find residential urban settlements. In Fig. 18, the results are about destroyed buildings in a war context: destroyed buildings were identified as having internal holes of shadow resulting from collapsed roofs.

End-users reports were quite satisfying. Results were evaluated by comparing our identified complex objects with reference data of the LISA land use dataset and Urban Atlas, respectively. Error matrices have been calculated for various use cases (e.g., see Fig. 19). The most valid results had an overall accuracy of 87 %.

Single buildings aggregation		Reference data			
		not aggregated	aggregated	total	User's accuracy
Classification	not aggregated	54	1	55	**98,2**
	aggregated	12	33	45	**73,3**
	total	66	34	100	
Producer's accuracy		**81,8**	**97,1**		
		Overal accuracy		**87,0**	
		Kappa index		0,731	

Urban fabric for UA		Reference data			
		not aggregated	aggregated	total	User's accuracy
Classification	not aggregated	42	1	43	97,7
	aggregated	15	42	57	73,7
	total	57	43	100	
Producer's accuracy		73,7	97,7		
		Overal accuracy		**84,0**	
		Kappa index		0,686	

Fig. 19 Error matrix for two use cases

6 Conclusions

Agencies (e.g., Environmental Protection Agency of Austria) use a combination of automated and manual approaches, based on expert knowledge, to derive land use information from land cover maps. They use ancillary data as well. These methodologies are expensive, time consuming and subjective. In other projects, semi-automatic procedures are applied: e.g., to produce GMES Urban Atlas maps, image analysis packages such as eCognition are used.

We propose an automatic approach for the recognition of complex objects by a combination of spatial rules and thematic information. In this way, costly integrations with other data sources are avoided. The vector format in standard OGC model allows us to increase interoperability with other systems. We can capture the semantics of complex objects in the rules that define them, keeping objects' structure separated from their visual representation, which can take various forms depending on scale and context.

The proposed approach is based on the application of a complete set of spatial operators for checking spatial rules and construction operators for defining an appropriate representation of complex objects. The experiments performed on test data provided by users showed that the approach is promising. Several kinds of complex objects, such as residential urban settlements, industrial sites, agricultural farmlands, river basins, road networks, could be recognized with an average overall accuracy of more than 85 %.

Test data give the means to evaluate single spatial rules by estimating the number of false positives and negatives. We noticed that, in general, when the spatial rules do not give satisfactory results, it is possible to improve the results by a better tuning of the rules themselves, by adding more refined geometric

properties to be checked. Current implementation (the TSS system) was more a proof of concept than a working prototype. We need to improve it in terms of performance and in terms of flexibility to allow an easier definition of complex objects and use of spatial functions.

Acknowledgments The author would like to thank the European Space Agency (ESA) for granting the Support To Topology (STO) project (http://wiki.services.eoportal.org/tiki-index.php?page=STO+Project), the SISTEMA GmbH, Vienna, Austria, for developing the prototype of the Topology Software System (TSS), the Environmental Protection Agency of Austria (UBA), Department for Biodiversity and Nature Conservation, Vienna, for providing LISA images and test cases, the GISAT, Prague, Czech Republic, for providing Urban Atlas images and test cases.

References

Baltsavias EP (2004) Object extraction and revision by image analysis using existing geodata and knowledge: current status and steps towards operational systems. ISPRS J Photogrammetry Remote Sens 58:129–151

Barnsley MJ, Møller-Jensen L, Barr SL (2001) Inferring urban land use by spatial and structural pattern recognition. In: Donnay J-P, Barnsley MJ, A.Longley P (eds) Remote sensing and urban analysis. Taylor and Francis, pp 102–130

Billen R, Clementini E (2004) Étude des caractéristiques projectives des objets spatiaux et de leurs relations. Revue Internationale de Géomatique 14(2):145–165

Bucher B, Falquet G, Clementini E, Sester M (2012) Towards a typology of spatial relations and properties for urban applications. Paper presented at the 3u3d2012: usage, usability, and utility of 3D city models. European cost action TU801 final conference, Nantes (France), pp 29–31

Clementini E (2010) Ontological impedance in 3d semantic data modeling. In: Kolbe TH, König G, Nagel C (eds) 5th 3D geoinfo conference, vol international archives of the photogrammetry, Remote Sensing and Spatial Information Sciences XXXVIII-4, part W15. ISPRS, 3–4 November 2010, Berlin, pp 97–100

Clementini E (2012) Directional relations and frames of reference. GeoInformatica. doi:10.1007/s10707-011-0147-2.10.1007/s10707-011-0147-2

Clementini E, Di Felice P (2000) Spatial operators. ACM SIGMOD Rec 29(3):31–38

Clementini E, Di Felice P, Hernández D (1997) Qualitative representation of positional information. Artif Intell 95(2):317–356

Clementini E, Di Felice P, van Oosterom P (1993) a small set of formal topological relationships suitable for end-user interaction. In: Abel D, Ooi BC (eds) Advances in spatial databases—third international symposium, SSD '93, vol 692., LNCSSpringer, Berlin, pp 277–295

Clementini E, Laurini R (2008) Un cadre conceptuel pour modéliser les relations spatiales. Revue des Nouvelles Technol de l'Inf (RNTI) RNTI-E 14:1–17

Cohn AG, Bennett B, Gooday J, Gotts N (1997) RCC: a calculus for region based qualitative spatial reasoning. GeoInformatica 1(1):275–316

Debeir O, Van den Steen I, Latinne P, Van Ham P, Wolff E (2002) Textural and contextual land-cover classification using single and multiple classifier systems. Photogrammetric Eng Remote Sens 68(6):597–605

Di Gregorio A, Jansen LJM (2000) Land cover classification system (LCCS): classification concepts and user manual

Egenhofer MJ, Franzosa RD (1991) Point-set topological spatial relations. Int J Geograph Inf Syst 5(2):161–174

Egenhofer MJ, Herring JR (1991) Categorizing binary topological relationships between regions, lines, and points in geographic databases. Department of Surveying Engineering, University of Maine, Orono

Egenhofer MJ, Mark DM (1995) Naive geography. In: Frank AU, Kuhn W (eds) Spatial information theory: a theoretical basis for GIS—international conference, COSIT'95, vol 988., LNCSSpringer, Berlin, pp 1–15

Exelis (2013) ENVI feature extraction module. www.exelisvis.com/

Frank AU (1992) qualitative reasoning about distances and directions in geographic space. J Vis Lang Comput 3(4):343–371

Friedl MA, Brodley CE (1997) Decision tree classification of land cover from remotely sensed data. Remote Sens Environ 61(3):399–409

GMES Urban atlas (2012) www.eea.europa.eu/data-and-maps/data/urban-atlas

Goyal R, Egenhofer MJ (1997) The direction-relation matrix: a representation of direction relations for extended spatial objects. In: UCGIS Annual Assembly and Summer Retreat, Bar Harbor

Grillmayer R, Banko G, Scholz J, Perger C, Steinnocher K, Walli A, Weichselbaum J (2010) Land information system Austria (LISA)—Objektorientiertes Datenmodell zur Abbildung der Landbeckung und Landnutzung. In: Strobl J, Blaschke T, Griesebner G (eds) Angewandte Geoinformatik 2010—Beiträge zum 22. AGIT-Symposium, Wichmann, pp 616–621

Hernández D (1993) Maintaining qualitative spatial knowledge. In: Frank AU, Campari I (eds) Spatial information theory: a theoretical basis for GIS—european conference, COSIT'93, vol 716., LNCSSpringer, Berlin, pp 36–53

Hussain M, Davies C, Barr R (2007) Classifying buildings automatically: a methodology. In: Paper presented at the GISRUK 2007: proceedings of the geographical information science research UK 15th Annual conference, Maynooth, 11th–13th April 2007

Ippoliti E, Clementini E, Natali S Automatic generation of land use maps from land cover maps. In: Proceedings of the AGILE'2012 international conference on geographic information science, Avignon, 24–27 April 2012

Ippoliti E, Clementini E, Natali S, Banko G (2012) A methodology for the automatic generation of land use maps. In: GI_Forum 2012: geovisualization, society and learning, Salzburg, Austria, 3–6 July, Wichmann Verlag, Berlin, pp 456–465

ISO (2010) ISO/TC 211 Geographic information/Geomatics. http://www.isotc211.org/. http://www.isotc211.org/

Klien E, Lutz M (2005) The role of spatial relations in automating the semantic annotation of geodata. In: Cohn AG, Mark DM (eds) COSIT 2005, vol LNCS 3693, pp 133–148. Spinger, Berlin

Land Information System Austria (2012) www.landinformationsystem.at/

Liu Y, Guo Q, Kelly M (2008) A framework of region-based spatial relations for non-overlapping features and its application in object based image analysis. ISPRS J Photogrammetry Remote Sens 63:461–475

Malinverni ES, Tassetti AN, Bernardini A (2010) Automatic land use/land cover classification system with rules based both on objects attributes and landscape indicators. In: Paper presented at the Geographic Object-Based Image Analysis GEOBIA 2010, Ghent, 29 June–2 July 2010

Natali S, Clementini E, Ippoliti E, Banko G, Brodsky L (2012) Topology software system: support to the creation of land use maps. In: Paper presented at the ESA-EUSC-JRC 2012. Image Information Mining Conference: Knowledge Discovery from Earth Observation Data, German Aerospace Center (DLR), Oberpfaffenhofen, Germany, 24–26 October

Novack T, Kux HJH, Feitosa RQ, Costa GA (2010) Per block urban land use interpretation using optical VHR data and the knowledge-based system Interimage. Paper presented at the Geographic Object-Based Image Analysis GEOBIA 2010, Ghent, 29 June–2 July

OGC (2011) Geometry object model. OpenGIS implementation specification for geographic information—simple feature access—part 1: common architecture

OGC Open Geospatial Consortium Inc. (1999) OpenGIS simple features implementation specification for SQL. OGC 99–049

Overwatch (2013) Feature analyst. http://www.overwatch.com/

Prüller R, Grillmayer R, Banko G, Mansberger R, Steinnocher K, Stemberger W, Walli A, Weichselbaum J (2011) Nutzen von innovativen Technologien für eine flächendeckende, flexible Landbeobachtung Österreichs. In: Strobl J, Blaschke T, Griesebner G (eds) Angewandte Geoinformatik 2011—Beiträge zum 23. AGIT-Symposium, Wichmann, pp 239–244

Tarquini F, Clementini E (2008) Spatial relations between classes as integrity constraints. Trans GIS 12(s1):45–57

Thunig H, Wolf N, Naumann S, Siegmund A, Jürgens C (2010) Automated LULC classification of VHR optical satellite data in the context of urban planning. In: Paper presented at the geographic object-based image analysis GEOBIA 2010, Ghent, 29 June–2 July

Trimble (2013) eCognition. www.ecognition.com/

Weichselbaum J, Banko G, Hoffmann C, Riedl M, Schardt M, Steinnocher K, Wagner W, Walli A (2009) Land information system Austria (LISA): Bedarfsgerechte Landnutzungsinformationen für die öffentliche Verwaltung. In: Strobl J, Blaschke T, Griesebner G (eds) Angewandte Geoinformatik 2009: Beiträge zum 21. AGIT-Symposium, Wichmann, pp 492–497

Wijnant J, Steenberghen T (2004) Per-parcel classification of ikonos imagery. In: Paper presented at the 7th AGILE conference on geographic information science, Heraklion

Automatic Extraction of Forests from Historical Maps Based on Unsupervised Classification in the CIELab Color Space

P.-A. Herrault, D. Sheeren, M. Fauvel and M. Paegelow

Abstract In this chapter, we describe an automatic procedure to capture features on old maps. Early maps contain specific informations which allow us to reconstruct trajectories over time and space for land use/cover studies or urban area development. The most commonly used approach to extract these elements requires a user intervention for digitizing which widely limits its utilization. Therefore, it is essential to propose automatic methods in order to establish reproducible procedures. Capturing features automatically on scanned paper maps is a major challenge in GIS for many reasons: (1) many planimetric elements can be overlapped, (2) scanning procedure may conduct to a poor image quality, (3) lack of colors complicates the distinction of the elements. Based on a state of art, we propose a method based on color image segmentation and unsupervised classification (K-means algorithm) to extract forest features on the historical 'Map of France'. The first part of the procedure conducts to clean maps and eliminate elevation contour lines with filtering techniques. Then, we perform a color space conversion from RGB to L*a*b color space to improve uniformity of the image. To finish, a post processing step based on morphological operators and contextual rules is applied to clean-up features. Results show a high global accuracy of the proposed scheme for different excerpt of this historical map.

P.-A. Herrault (✉) · D. Sheeren · M. Fauvel
University of Toulouse, INP-ENSAT, UMR 1201 DYNAFOR, Av. de l'Agrobiopôle, BP 32607, Auzeville Tolosane, Castanet Tolosancedex, Toulouse 31326, France
e-mail: pierrealexis.herrault@ensat.fr

D. Sheeren
e-mail: david.sheeren@ensat.fr

M. Fauvel
e-mail: mathieu.fauvel@ensat.fr

P.-A. Herrault · M. Paegelow
University of Toulouse, UTM, UMR 5602 GEODE, 5, allée A. Machado, Toulouse 31058, France
e-mail: paegelow@univ-tlse2.fr

D. Vandenbroucke et al. (eds.), *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography,
DOI: 10.1007/978-3-319-00615-4_6,

1 Introduction

Olds maps contain specific spatial information as location of historical places or historical land cover, elevation contour lines, building footprints and hydrography. Capturing this spatial information interest for various studies as those one about long term changes of landscapes, urban development or coastlines evolution (Cousins 2001; Bender et al. 2005; Gimmi et al. 2011; Smith and Cromley 2012). For few years, a lot of these old maps are available for research thanks to the work of scanning by the national archives in different countries.

The traditional approach to capture the cartographic objects in the old maps is based on a user intervention (for digitizing). This approach is obviously very time-consuming, and difficult to reproduce on large areas. Several works attempted to develop automated data captures techniques in order to establish reproducible procedures (Leyk 2006). However, most of them are specific to only one particular map and generally, cannot be applied on other historical maps.

In this chapter, we propose a new method to extract automatically forest features from the historical 'Map of France' dating from the 19th century. The method is based on image recognition techniques including a color space transformation and an unsupervised classification of the digital map. The chapter is structured as follows. In Sect. 2 we describe existing works developed to the automatic processing of old maps. Next, the method we propose is presented in Sect. 3. Experiments and results obtained on several excerpts of the considered historical map are presented in Sect. 4. Finally, we draw some conclusions in Sect. 5.

2 Previous Works

Capturing features automatically from raster maps is a major challenge in GIS for many reasons. First, there are many planimetric elements that overlap each other such as road lines, elevation contour lines, marks or soil features, and so on. Second, scanning procedure or image compression may conduct to a poor quality of the data that make the recognition more difficult. Last, some old maps may be in black and white (without any other colors) that make the development of an automatic procedure almost impossible (Fig. 1).

Several authors already proposed automatic methods to capture geographical objects in scanned thematic maps (Ansoult et al. 1990). Many examples of maps have been investigated for feature recognition such as the topographic maps of the United States Geological Survey, the military maps of the Polish Geographic Institute (Iwaniowski and Kozak 2012) or the Swiss National historical map (Leyk 2006).

In a general way, automatic extraction procedure can be divided in three steps: (1) a cleaning-up step, (2) a feature recognition/extraction step and (3) a post-processing step. The first step is defined to make the feature extraction easier while the last step is carried out to improve the extraction results. Steps 2 and 3 can be

Fig. 1 **a** Overlapping of planimetric elements on historical 'map of France' (∼1850); **b** review of the historical 'map of France' in *black* and *white* (∼1900)

viewed as a single integrated process according to some authors (Wise 1999). In the next sections, we described the existing approaches for old maps processing by combining the pre and post-processing steps (1 and 3) in the same part because of the similarity of the techniques.

2.1 Existing Pre and Post-processing Methods

Generally, scanned thematic or old maps are complex and contain different kinds of noise. Indeed, many elements that overlap each other (such as road lines, elevation contour lines or text) can be seen as noise when the users want extract specific information. Scanning procedures may also conduct to a shading effect on the maps and alter the quality of the original document. This noise makes difficult the use of the digital maps and in particular the automatic extraction of some features of interest (e.g., the buildings or the forests).

To overcome this problem, digital historical maps are often filtered using image-processing techniques. The filters consist in assigning a new value in each pixel using the pixel values in its neighborhood. They are well adapted to reduce noise in the images while preserving some structural elements like edges or contours. Various filters exist in the image-processing field ranging from convolution filters to morphological filters.

Ansoult and Soille (1990) proposed a morphological-based filtering step in their thematic map processing method. They showed how dilatation operator and image reconstruction process can be used successively to remove text in the map which complicates regions extraction. The proposed approach provides a general

framework to deal with this classical issue. Chiang et al. (2012) also adopted morphological filters in a pre-processing step. He applied successive erosions to remove elevation contour lines before to extract the roads which are very similar. A close approach is followed by Samet et al. (2010) in a post-processing stage. A dilatation operator is applied to reconnect broken contour lines and reduce noise in the map after a color image segmentation.

Iwanowski and Kozak (2012) used an image cleaning run two times in their extraction procedure. One time before segmentation, in order to filter the image and remove small pixel scale variations of colors and another time, after image segmentation, using successively a closing by reconstruction and an opening by reconstruction to improve forest detection results. Arrighi and Soille (1999) in a similar perspective, used some morphological filters to produce a clean mask of the elevation contour lines before a real line detection.

2.2 Existing Features Extraction Methods

Data capture on historical maps may concern various types of features such as (a) text (marks, names of cities or rivers…), (b) regions (land cover, buildings,…), (c) symbol (semantic information or punctual objects like churches, mills…) or (d) lines (elevation contours, edges, roads…). Each of these feature categories may conduct to define specific techniques to extract them automatically.

Text extraction is one of the most complicate tasks to automate in old maps. Indeed, text and graphics are often overlapped that make difficult the separation of the elements and therefore text recognition. Cao and Tan (2002) proposed a method for text recognition based on the observation that the strokes of characters are generally more short segments than those of cartographic elements. The authors used a combination of line continuation with the line width to separate elements and improve text extraction. Text can be also detected using color attributes when the color characters differ from the others objects. Centeno (1998) used for example a Karhunen-Loeve color-space conversion to isolate text and capture it. Last, Myers et al. (1996) performed a character recognition method with a verification-based approach to detect text without requiring pre-segmentation graphical entities.

Concerning regions, existing methods to extract them automatically are mainly based on image segmentation and/or classification techniques. Shaw and Bajcsy (2011) used a combination of a region-growing segmentation and morphological operators to extract the Ontario Lake in several maps. The results shows that this technique is limited when there are numerous overlapped objects. Mahmouda et al. (2011) proposed a similar approach including a multi-resolution segmentation and a (rule-based) classification procedure using color and spatial attributes to extract features areas. Other approaches consist in classifying the image without segmentation step in advance. If the maps contain thematic colors layers, color signature can provide much information to capture objects of interest but means some limitations when processing poor image quality (Leyk 2006).

Chiang et al. (2012) used K-means algorithm to capture road vectors in a raster map. Results show that algorithm is efficient but that the users need to determinate a large number of K clusters to separate different features. Henderson and Linton (2009) carried out a comparison between k-means algorithm, expectation-maximization and graph theoretic to separate semantic classes in a raster maps. Threshold-based classification methods are also used by some authors to separate color or grayscale components related to thematic classes. This way is also used as a first step before to set up a more accurate segmentation method (Ansoult 1990; Chi and Yan 1993).

The extraction of map symbols is complicated by the fact they can form themselves discontinuous chains (Gamba and Mecocci 1999). A majority of works employed knowledge-based approaches (Arias et al. 1993) to perform a symbol extraction based on set of rules. Other approaches identify the symbols in the legend, constructing a training set library to then use it in the classification of the geographic symbols (Samet and Soffer 1998). It is also possible to consider symbols as short lines and then, applying line extraction techniques to symbol recognition (Boesch 1996).

To finish, line extraction in maps may concern contour-objects lines, elevation contour lines, traffic connections lines, drainage networks etc. Kaneko (1992) used directional distances propagations for extracting lines structure in line drawings whereas Mariani et al. (1997) proposed several reconstruction algorithms for the identification of drainage networks. When the maps are printed in colors, this information can provide very reliable results for line extraction. For example, Ansoult et al. (1990) used the mean and variance of the hue channel to extract regions contour in a soil map. Arrighi and Soille (1999) adopted a different approach in extracting extremities of the lines before thinning them with a skeletonization algorithm. In last, Chen et al. (1999) proposed a new method based on local window segmentation approach to overcome thick lines.

All these methods appear as efficient in particular contexts but are not suitable for all supports. Thus, each raster-map contains specific problems of overlapping or quality so an identification step is required to identify the characteristics of the map before choosing the procedure to apply.

3 Proposed Method

In this section, we expose our extraction method of the forest features for the historical 'Map of France' (Fig. 2). Our procedure includes the following steps:

(1) Pre-processing for map filtering
(2) Color space transformation from RGB to CIELab space
(3) Color image classification based on K-means algorithm
(4) Post-processing of forest features.

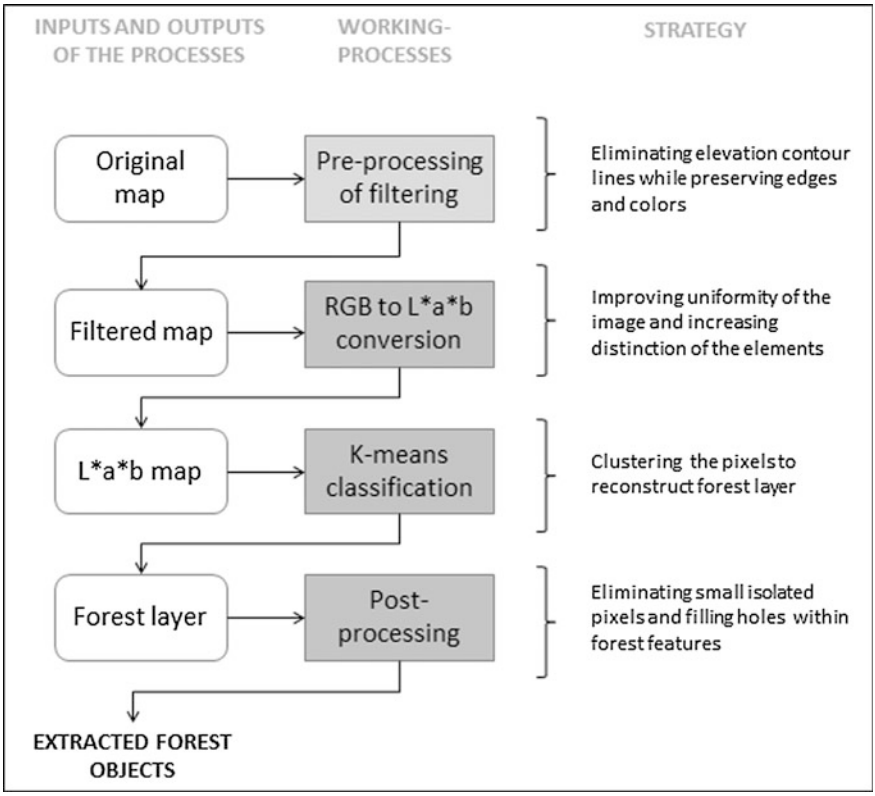


Fig. 2 General methodology (flowchart inspired by Leyk et al. 2006)

3.1 Pre-processing for Map Filtering

Before performing Color Image Classification, a clean-up process is applied to eliminate elevation contour lines which overlap forests. The goal is to remove these contour lines while preserving colors and edges of objects. Three steps are performed: (1) dilatation, (2) median filtering, (3) low-pass filtering (Fig. 3).

A first step was dedicated to fill all holes within forest (created by text, symbols and elevation contour lines) except those one created by roads. We choose to use a morphological dilatation operation which allows filling holes thanks to a structuring element (SE). The best SE we found is a square of 5 × 5 pixels. This operator is applied on each band of the image (R, G, B).

Secondly, we want to simultaneously reduce remaining elevations contour lines pixels while preserving edges and colors. As well- known, two-dimensional median filtering appears as efficient for this objective (Samet et al. 2010). Nevertheless, a neighborhood size ill-suited may conduct to important cost of

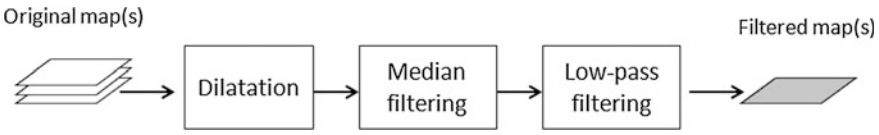


Fig. 3 Flowchart of the pre-processing of filtering

blurring and makes edges invisible in the image (Motwani et al. 2004). We choose to perform a median filter in a window of 5*5 on each dilated band R, G and B. We applied it five times until obtaining a result that we judge satisfactory.

To finish, we have observed a light remaining background noise. Another application of a median filter would conduct to make edges invisible. To overcome this problem and to proceed to a sweet objects extraction, we propose to use a low pass filter in a window 5*5 which allows eliminating background noise while preserving low frequencies.

3.2 Color Space Transformation from RGB to CIELab Space

Previous steps allowed us to filter objects but colors are disturbed during this process: hue is heterogeneous for a same object class and intensity fluctuates according to low slope areas or high slope areas. Globally, color is rarely homogenous in historical maps so there is a need to choose a color space which assures certain uniformity in luminosity.

The choice of a suitable color space is often a crucial step for old color map processing. The traditional color space is the RGB space. However, this color space presents some limitations for the automatic extraction of forests: non-uniformity of the luminosity, lack of human perception (Angulo and Serra 2003). Other color spaces are well-known for computer graphic applications like Hue, Saturation, Value (HSV) or Hue, Luminosity, Value (HLS) but are less suitable for image processing (Hanbury and Serra 2003).

In this method, we propose to transform the image into the L*a*b (or CIELab) color space. This choice is inspired by previous works showing that CIELab is appropriate for color images with various types of noises but not only (Ganesan et al. 2010). The performance of different unsupervised classification (particularly K-means clustering) has been analyzed in few color spaces like RGB (Brunner et al. 1992), HSV or L*a*b (Wiszeki and Stiles 1982). The general consensus implies that L*a*b is more efficient since removing effects of illuminations yields optimal segmentation results.

In this space, the characteristic known as luminosity is reported on an axis L that is perpendicular on a pile of 'ab' planes where each plane contains all the possible colors for a given luminosity (Fig. 4). This axis represents the sensitivity of the human eye to the luminance. Thus, it is possible to consider each variation

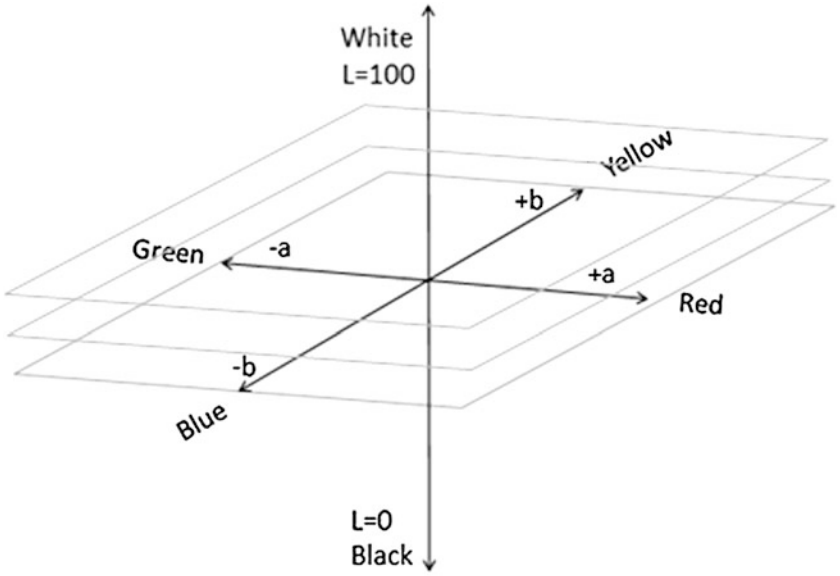


Fig. 4 The axis L perpendicular on a pile of '*ab*' planes

of the green color (e.g., forest features) like a succession of pure colors on axis 'a' since this axis informs us on the degree of green which qualifies the objects. We attempt to search different green colors of the forests for each possible value of L.

This allows us to take into account the green variations within forest areas. For example, in RGB color space, two green pixels with a different luminosity could be separated into two different classes by a clustering algorithm because this space does not take into account luminosity but only the proportion of green within pixel.

The colors are uniformly distributed in an a*b plane, from red to green along the a axis and from blue to yellow along the b axis. The three coordinates of the L*a*b color space is given as following:

L the luminosity layer
a indicates where color falls on the axis from red to green
b indicates where color falls on the axis from yellow to blue (Fig. 5).

Our goal is to find the most representative partition of the forests in the ab chromaticity diagram for each possible value of L. This partition can be found automatically in the 2-dimensional space by clustering.

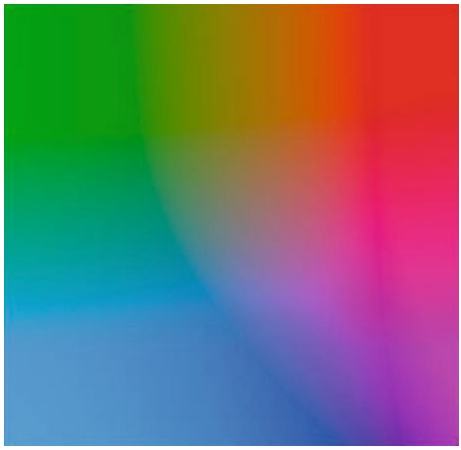


Fig. 5 An *'ab'* plane for only one value of L (from Barbu et al. 2012)

3.3 Color Image Classification Based on K-Means Algorithm

K-means is one of the simplest and most popular unsupervised learning algorithm. This technique is used in various application domains including computer science, ecology, geostatistic or remote sensing (e.g., Lucchese and Mitra 1999; Jigar et al. 2012).

As a reminder, the main idea of the algorithm is to define a partition of N observations into K clusters in which each observation belongs to the nearest cluster (Barbu et al. 2012). K-means algorithm is given as following:

$$J_{K-means} = \sum_{k=1}^{K} \sum_{j \in S_K} d^2(X_j, C_k)$$

- where K is the number of the cluster (class) evaluated in a space defined by S_K
- (X_j, C_K) is the distance between the observation X_j and the class C_k.

The K-means algorithm is performed in four steps:

(1) Place K points into the space represented by the observations that are being clustered
(2) Each observation is assigned to the closest point of K defined as the centroid of one cluster C_k
(3) When all the observations have been assigned to the K clusters, the centroids of the obtained clusters must be recomputed
(4) This procedure (steps 2 and 3) is iterated until the algorithm converges (i.e. until there are no cluster changes for the observations after the re-computation of the centroid positions) or that it respects a fixed stopping criterion.

The difficulty in K-means algorithm resides in the definition of the optimal number of clusters K. Because of a real complexity of old maps, number of colors can widely vary between one scan and another. The difficulty in K-means algorithm resides in the definition of the optimal number of clusters K. Because of a real complexity of old maps, number of colors can widely vary between one scan and another. Here, we decided to compute ten clusters ($K = 10$) in order to extract the forest features. This number enables us to better separate different thematic classes while avoiding mixed clusters. Some clusters are merged after this step in order to obtain a single layer including only forest and non-forest features.

3.4 Post-processing for Features

A post-processing step was carried out to correct some artifacts in the forest layer after the classification step. This post-processing includes a morphological opening in order to remove the small isolated pixels. The opening operator was applied with a structuring element of 3×3. Some holes within forests were also removed using contextual rules: all the non-forest pixels enclosed by forest pixels were classified as forest. This operation created homogeneous forest objects. Finally, the original image including only the forests was reconstructed.

4 Results and Discussion

The method was applied on the historical "Map of France". This map was produced from 1825 to 1866 at 1:40,000 scale and is known for its relatively high planimetric accuracy. The map includes several thematic categories (like forests, buildings, grasslands…) which are represented in color. Experiments were conducted on three different excerpts of the map which differ in terms of slope and relief, quality and colors for the forest features (various green levels) (Fig. 6).

The performance of the method was assessed quantitatively by computing a confusion matrix and some related accuracy index. Here, we give also the intermediate results in order to illustrate the interest of each step. The effect of the filtering step (which enables to eliminate elevation-contour lines) is illustrated in Fig. 7.

In Fig. 7a, we can observe on the histogram of the original excerpt 2 (in gray levels) an important number of pixels in low values (gray level < 100). These pixels are related to the elevation contour lines which overlap to regions. Thanks to the dilatation operator, a great part of these lines were removed while preserving edges (Fig. 7b). However, this filtering was not sufficient because too many pixels with low values remain.

The second step of filtering conducts to further reduce these dark pixels in order to smooth the image. It is interesting to observe on the histogram of the image (c) that

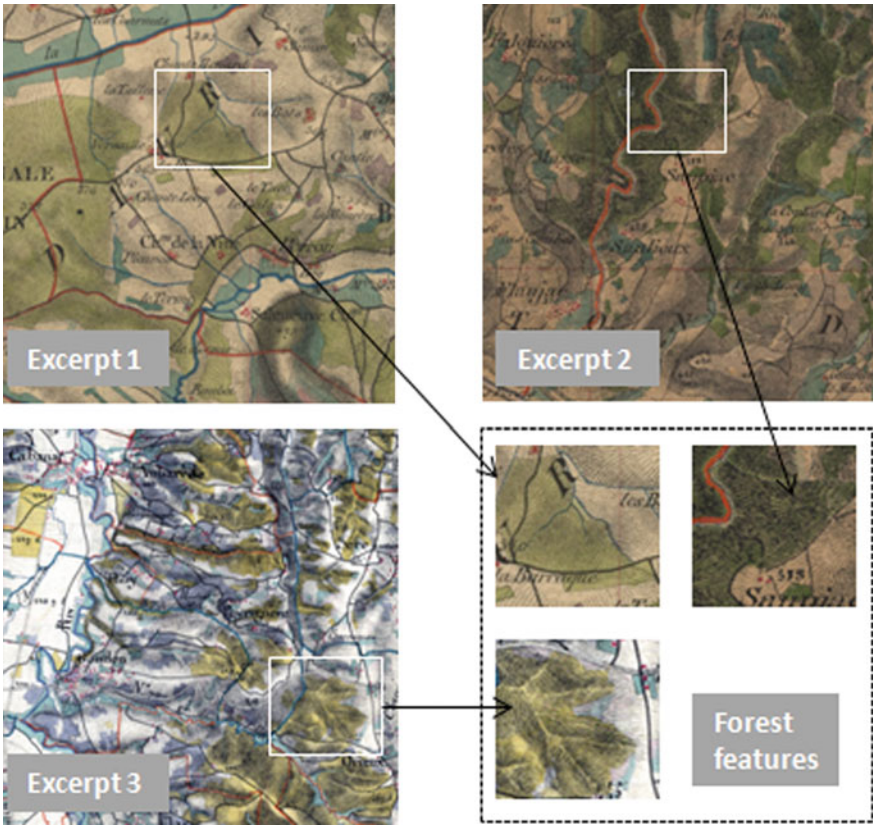


Fig. 6 Three excerpts of historical 'map of France' tested

the median filtering is able to remove dark pixels while transferring to higher frequencies. We can also see that a background noise remains: it is represented by small variations in the minimum values of the histogram (c).

A last step of filtering provides us the way to smooth these small variations: the low pass filtering. Thanks to this convolution filter, we go to reduce noisy pixels while correctly preserving dark pixels which correspond to road lines (Fig. 7d). The final result of the filtering procedure allows us to set up color-image classification. Colors and edges are still visible. This conducts to more uniform regions while facilitating pixel classification.

To show the performance of the L*a*b color space, we conducted a K-means classification on the excerpt 2 in RGB color space in order to compare it with L*a*b (Fig. 8).

The quality of results differs. In the image (b), the extracted features do not represent an identical land use cover. Indeed, after the fusion of the different clusters which contained forest features, the final result is not satisfactory. Forest features are mixed with built areas (pale red) or grasslands (blue). Probably the

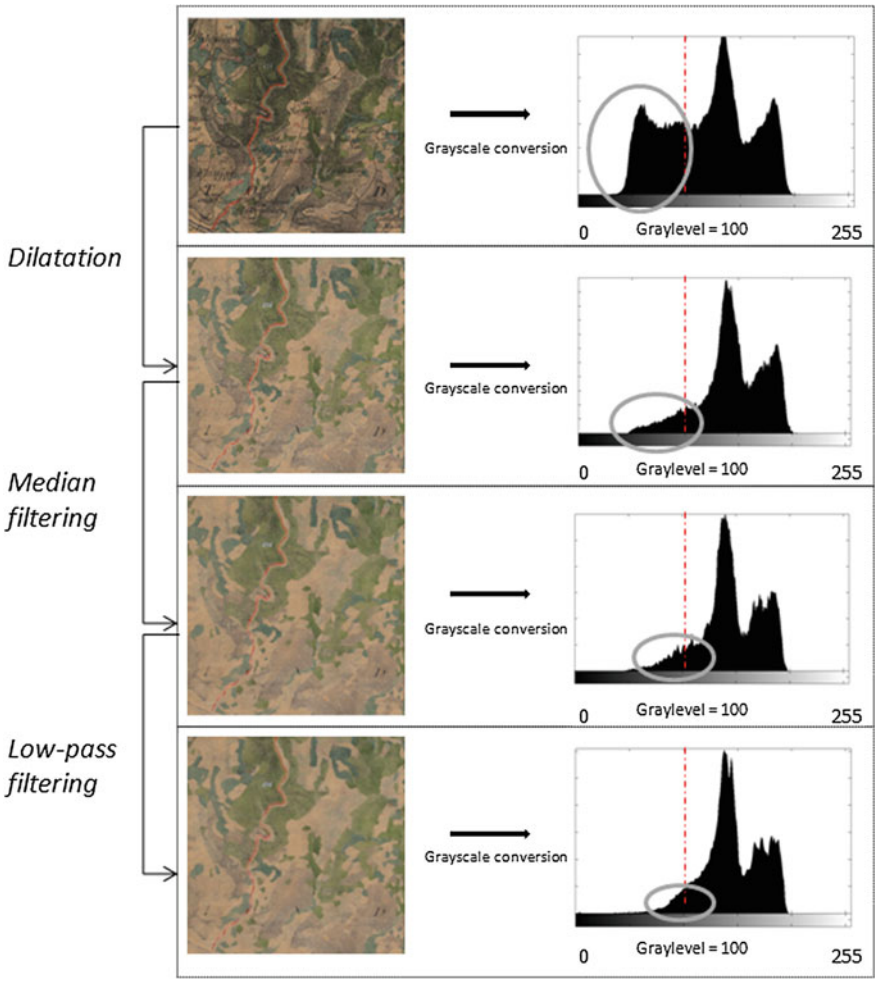


Fig. 7 Successive steps of pre-processing of filtering

lack of uniformity of luminance in RGB color space does not allow separating the color variations of forest features from other land covers. By contrast, image (a) shows that color-image classification in L*a*b color space is the most efficient. Only forest features were extracted and any confusion with other land cover classes persists.

To finish, post processing step based on mathematical morphology and contextual rule was applied to fill holes within forest features and removing small isolated pixels. Figure 9a shows that extracted forest layer is not satisfactory because of too many holes inside. There are also small isolated pixels which do not represent forest elements and contribute to over detect the feature layer. After a binarization step, Fig. 9c means that morphological opening is efficient to remove

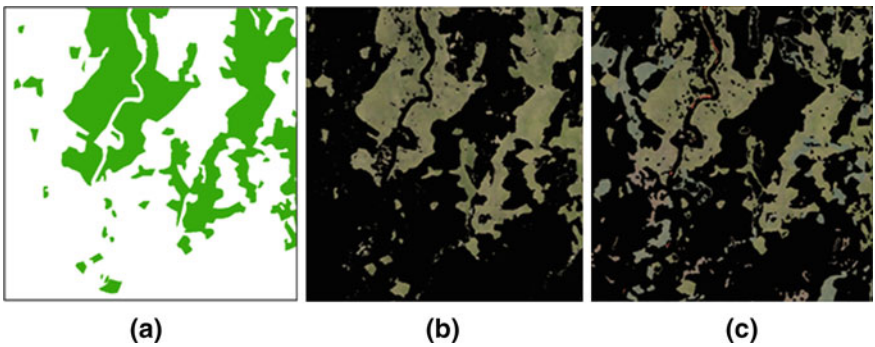


Fig. 8 **a** Manual extraction layer. **b** K-means classification in L*a*b color space and **c** In RGB color space

small isolated pixels which are non-forest pixels. The contextual rule used proves to be also suited to fill holes within forest features. Indeed, we observe that holes in forest in Fig. 9b have been filled and allocated to 'white' class (Fig. 9c). To finish, forest layer of the original image is reconstructed by adding corrected binary image and original image in order to evaluate the performance of the global procedure.

The general procedure has been applied on the three excerpts previously presented (Visual results in Fig. 10). The presented results show a high global accuracy (Kappa $\approx$ 0.90). The processes and the different parameters defined proved to be valid for each excerpt tested (Table 1).

Sensitivity Index and Specificity Index measure conditional probabilities that forest and non-forest are correctly classified. For each excerpt, these two indexes exceed 0.92 which conduct us to think the developed method is efficient to separate forest features and other elements on the image. PCC index is also very high for the three image tested (>0.90). Kappa index is lower for the three examples but is globally satisfactory for the three examples.

Globally, the extraction procedure shows a real robustness because of homogeneous results on the three excerpts tested from the historical 'Map of France'.

Successive steps of the filtering process proved to be efficient to remove contour lines elevation. Each filter used in a specific purpose contribute to eliminate noisy pixels (e.g., contour elevations) while preserving edges and contours. At the end of this procedure, the color is not disturbed and it is always possible to proceed to an unsupervised classification.

Concerning L*a*b color space, we can say that it appears well suited to low quality maps. Its structure offers the possibility to consider the green color variations as a unique color which increases uniformity of the forest regions in one scan and between them. This is a crucial advantage because of the heterogeneous quality of historical maps in general and the different color variations which can represent a same class. Accuracy results highlighted some minor problems.

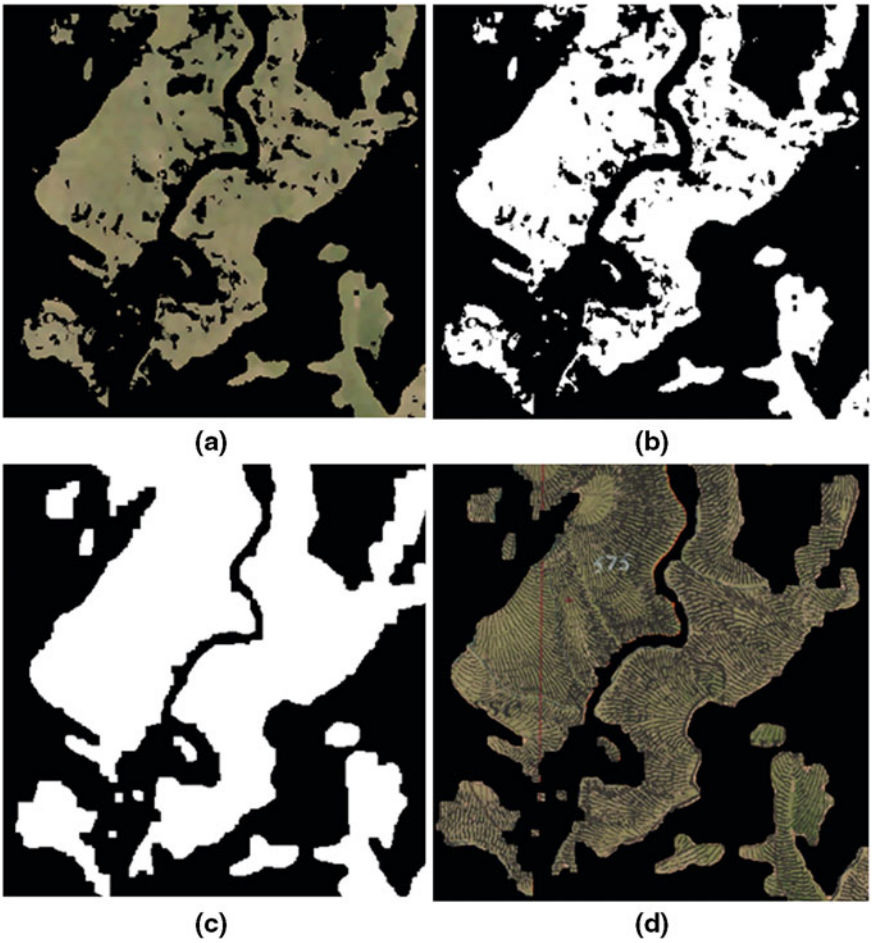


Fig. 9 **a** Extracted forest feature; **b** binary extracted feature; **c** binary corrected feature; **d** reconstructed forest feature

Firstly, there can be a trend to 'under-detect' forest during the unsupervised classification step. Indeed, the pre-processing step is efficient but some high slope areas are still shaded and are not recognized as forest features. This is even truer as the luminance of the forest feature is low: it is the case for the excerpt 2.

The post-processing step may also conduct to some confusion. The definition of the structuring element for morphological opening may change the morphology of several features or connect features to others. Thus, the historical structure of the forest matrix can be slightly modified in a local environment. We can do a similar observation during the step which fills holes within features. Indeed, some areas have been filled while they are not forest features but rather lakes or small

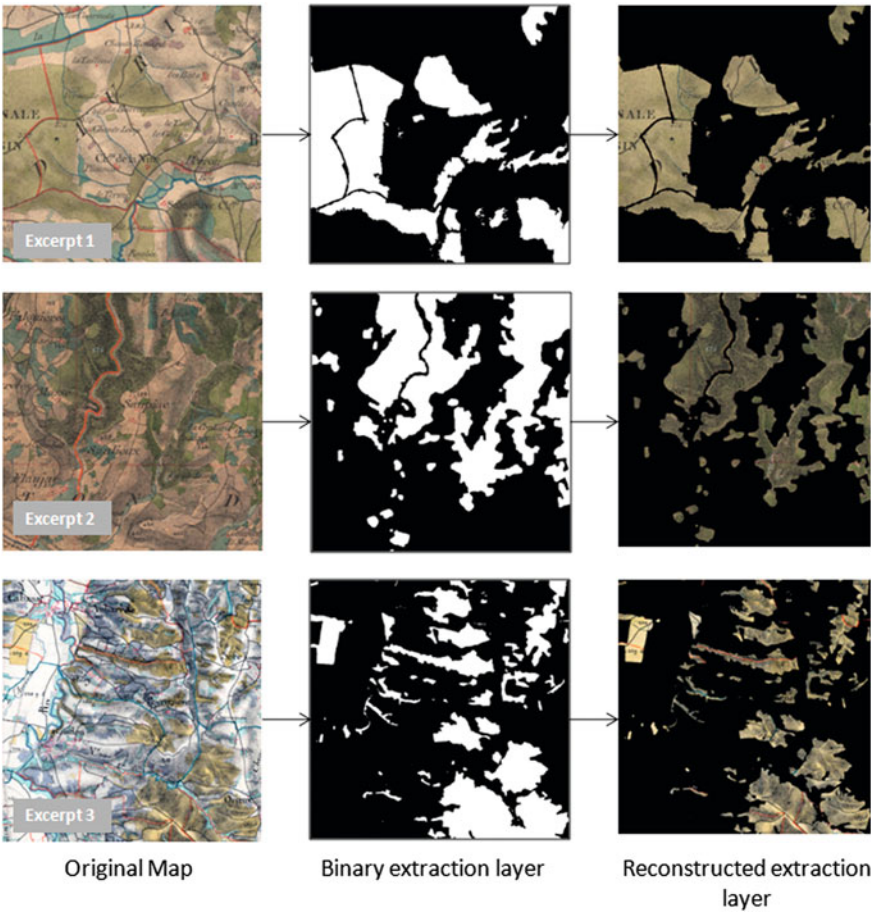


Fig. 10 Visual extraction results on the three excerpts tested

Table 1 Extraction results for three excerpts of 1,500 × 1,500 pixels each

Excerpt tested	Excerpt 1	Excerpt 2	Excerpt 3
Accuracy index			
Sensitivity	0.92	0.92	0.94
Sensitivity	0.97	0.96	0.97
Pcc	0.96	0.95	0.96
Kappa	0.92	0.89	0.9

croplands. This can be explained by the definition of parameters which are not always suited to map specificities.

Last, the quality of the map is crucial in extraction procedure. Some specific areas within one scan are more or less preserved in terms of colors. So, several forest features and croplands (beige color) after pre-processing become confused while creating difficulties in extraction process. Pixels values for these two classes are very close for all the possible values of L. For these specific cases, a user intervention is required.

5 Conclusion

In this chapter, a method for extracting forest features from historical 'Map of France' was presented.

It is based on a pre-processing step followed by an unsupervised classification in a CIELab color space. Last, a post processing step was performed in order to improve extracted forest features.

Results on the three excerpts of the historical map are shown in Table 1 and present a high global accuracy. For three excerpts whose color of forest features, slope level and quality are very different, Kappa index is in average equal to 0.90. These results show a real robustness of the proposed scheme.

The main advantage of this approach is that it takes into account the color variations of forest features. The pre-processing of filtering allowed us to globally eliminate elevation contour lines which widely improve the distinction between features. It should be also noted that L*a*b color space proved to be efficient for increasing uniformity in a low quality map thanks to the taking into account of all possible value of Luminance in the image. Using this color space for other extraction tasks in historical maps could be an interesting alternative.

The accuracy assessment identified some minor problems. Pre-processing and post processing steps include the definition of some parameters which are not always suited to specificities of all sheets of the map. This difficulty may conduct to 'over-detect' or 'under-detect' forest features in these cases. Moreover, this may conduct to modify features shape or connecting between them which may modify global structure of the forest matrix in a local environment. Last, the poor color quality of the map widely varies from one sheet to another. Several classes can become confused while making color image classification difficult.

Future research will be dedicated to develop more advanced contextual rules for the post processing step. Taking into account spatial relationships between features could allow us to improve extraction results without any risks of modifying global structure of the features. This global procedure will be also tested on other raster color maps which suffer of identical quality problems. The main phases of the proposed scheme could be an interesting framework for these extraction problems.

Acknowledgments This research was supported by the French National Research Agency (ANR JCJC MODE-RESPYR 2010 1804 01–01). P.-A. Herrault is also funded through a PRES Toulouse University and Region Midi-Pyrenees grant.

References

Angulo J, Serra J (2003) Mathematical morphology in color spaces applied to the analysis of cartographic images. In: Levachkine S, Serra J, Egenhoger M (eds) Semantic processing of spatial data, in proceedings of the GEOPRO 2003-international workshop semantic processing of spatial data, Mexico City, pp 59–66

Ansoult M, Soille P, Loodts J (1990) Mathematical morphology: a tool for automated GIS data acquisition from scanned thematic maps. Photogrammetric Eng Remote Sens 56(9): 1263–1271

Arias JF, Lai CP, Chandran S, Kasturi R, Chhabra A (1993) Interpretation of telephone system manhole drawings. In: Second international conference on document analysis and recognition, Tsukuba Science City, pp 365–368

Arrighi P, Soille P (1999) From scanned topographic maps to digital elevation models, in the international symposium on imaging applications in geology geovision '99. University of Liege, Belgium

Barbu T, Ciobanu A, Mihaela C (2012) Automatic color-based image recognition technique using LAB features and a robust unsupervised clustering algorithm. In: Proceedings of the 13th WSEAS international conference on automation and information (ICAI '12), Iasi, Romania, 13–15 June 2012

Bender O, Boehmer HJ, Jens D, Schumacher KP (2005) Using GIS to analyse 200 years of cultural landscape change in Southern Germany. Landscape Urban Planning 70:111–125

Boesch R (1996) Detection and extraction of complex map symbols. Int Arch Photogrammetry Remote Sens 31 Part B3, Vienna

Brunner CC, Maristany AG, Butler DA, Vanleuween D, Funck JW (1992) An evaluation of colorspaces for detecting defects in Douglas-fir veneer. Ind Metrol 2(3 and 4):169–184

Cao R, Tan C (Eds) (2002) Text/graphics separation in maps, vol 2390, Springer, Berlin

Centeno JS (1998) Segmentation of thematic maps using color and spatial attributes. Graphics Recognit Algorithms Syst 1389:221–230. Lecture notes in computer science, Springer

Chen L, Liao H, Wang J, Fan K (1999) Automatic data capture for geographic information systems. IEEE Trans Syst 5(2):205–215

Chi Z, Yan H (1993) Map image segmentation based on thresholding and fuzzy rules. Electron lett 29(27):1841–1843

Chiang YY, Leyk S, Knoblock CA (2012) Efficient and robust graphics recognition from historical maps. Graphics Recognit Achievements Challenges Evol Selected Papers of the ninth international workshop on graphics recognition (GREC), Lecture notes in computer science

Cousins S (2001) Analysis of land-cover transitions based on 17th and 18th century cadastral maps and aerial photographs. Landscape Ecol 16:41–54

Gamba P, Mecocci A (1999) Perceptual grouping for symbol chain tracking in digitized topographic maps. Pattern Recogn Lett 20:355–365

Ganesan P, Rajini V, Rajkumar IR (2010) Segmentation and edge detection of color images using CIELAB color space and edge detectors, in emerging trends in robotics and communication technologies (INTERACT), 2010 International conference

Gimmi U, Lachat T, Burgi M (2011) Reconstructing the collapse of wetland networks in the Swiss lowlands 1850–2000. Landscape Ecol 26:1071–1083

Hanbury A, Serra J (2003) Colour image analysis in 3-D polar coordinates, DAGM congress, p 8

Henderson TC, Linton T (2009) Automatic segmentation of semantic classes in raster map images. In: 8th IAPR International workshop on graphics recognition, La Rochelle, France, July 2009, pp 253–262

Iwanowski M, Kozak J (2012) Automatic detection of forest regions on scanned old maps. Electr Rev 88:249–252

Jigar MS, Brijesh S, Satish SK (2012) A new K-mean color image segmentation with cosine distance for satellite images. Int J Eng Adv Technol (IJEAT), 1(5), ISSN: 2249–8958

Kaneko T (1992) Line structure extraction from line- drawing images. Pattern Recogn 25(9):963–973

Leyk S, Boesch R, Weibel R (2006) Saliency and semantic processing: extracting forest cover from historical topographic maps. Pattern Recogn 39(5):953–968. doi:10.1016/j.patcog.2005.10.018

Lucchese L, Mitra SK (1999) Unsupervised segmentation of color images based on k-means clustering in the chromaticity plane content-based access of image and video libraries. (CBAIVL '99), In: Proceedings IEEE workshop on digital object identifier: doi:10.1109/IVL.1999.781127, pp 74–78

Mahmouda A, Elbialya S, Pradhana B, Buchroithner M (2011) Field-based landcover classification using TerraSAR-X texture analysis. Adv Space Res 48(5):799–805

Mariani R, Lecourt F, Deseilligny M, Labiche J, Lecouturier Y (1997) Interprétation de cartes géographiques:algorithmes de reconstruction de réseaux hydrographiques et routiers. Traitement du Signal 14(3):317–334

Motwani MC, Gadiya MC, Motwani RC, Harris FC Jr. (2004) Survey of image denoising techniques. In: Proceedings of global signal processing, Santa Clara, CA

Myers G, Mulgaonkar P, Chen C, De Curtins J, Chen E (1996) Verification-based approach for automated text and feature extraction. In: Kasturi R, Tombre K (eds) First IAPR workshop on graphics recognition, Lecture notes in computer science, vol 1072, Springer, Berlin, pp 190–203

Samet H, Soffer A (1998) Magellan: map acquisition of geographic labels by legend analysis. Int J Doc Anal Recogn 1(2):89–101. doi:10.1007/s10032005

Samet R, Askerbeyli INA, Varol C (2010) An implementation of automatic contour line extraction from scanned digital topographic maps. Appl Comput Math 9(1):116

Shaw T, Bajcsy P (2011) Automation of digital historical map analyses. In: IS&T/SPIE Electronic Imaging, 7869-09, Session 3, Conference 7869: computer vision and image analysis of art II, 23–27 Jan (oral presentation)

Smith MJ, Cromley RG (2012) Measuring historical coastal change using GIS and the change polygon approach. Trans GIS 16(1):3–15

Wise S (1999) Extracting raster GIS data from scanned thematic maps. Trans GIS 3(3):221–237

Wyszecki G, Stiles WS (1982) Color science, 2nd edn. Wiley, London

Part III
Data Quality

Selecting a Representation for Spatial Vagueness: A Decision Making Approach

Mohammed I. Humayun and Angela Schwering

Abstract Representing vague places is a challenge in information systems. There are several approaches, each differing in aspects such as the underlying assumptions they make about space, their data models and reasoning abilities. Despite this there is no general solution and the question of which method to select is a matter of fitness for purpose. So far no methodology exists to support choosing the appropriate representation for a given problem. A formal decision making approach is presented here to select a suitable modelling technique to represent vague places. To do this, the criteria on the basis of which the decision is made are derived first. Commonly used methods to model spatial vagueness and uncertainty are then analyzed on the basis of these criteria. Finally, we describe a methodology that uses the analytic hierarchy process, in order to provide a quantitative ranking of candidate methods in their order of suitability for an application scenario.

1 Introduction

Modelling the real world from observations is imperfect. This may be due to *inaccuracy* (lack of correlation with the actual world due to errors) or *imprecision* (when information is not specific enough) (Worboys and Clementini 2001). One kind of imprecision is *vagueness.* A vague predicate has borderline cases where it is not clear if the predicate applies to it or not as happens in the case of a mountain, forest, lakes or with some named places. The result is that the boundaries for such

M. I. Humayun (✉) · A. Schwering
Institute for Geoinformatics, University of Muenster, Münster, Germany
e-mail: humayun@uni-muenster.de

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

D. Vandenbroucke et al. (eds.), *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography,
DOI: 10.1007/978-3-319-00615-4_7,

places are not well defined, contrary to the demands of digital representations which expect a crisp definition.

Applications such as gazetteers sidestep this complexity by simply choosing a representative geometry signifying the vague object. Where vagueness is to be preserved to some extent, one may resort to alternative methods such as fuzzy model, the egg-yolk representation or rough sets. Though these methods deal with spatial vagueness, their applicability varies from task to task owing to the fact that they are based on different foundational principles and other important factors such as the kinds of data models they operate upon and the reasoning capabilities they offer. Employing one or more of these methods requires a clear understanding of the merits and limitations of each, and a way to match these with requirements of the application. Lacking a proper procedure for selection, a user might end up with a representation method which is sub-optimal for the given situation. We argue for the need to specify user requirements in a consistent manner. We then use a structured decision making process called the analytic hierarchy process (AHP) in order to prioritize and rank possible alternative representations according to the given criteria.

Fitness for purpose of the different methods is contextually dependent on the purpose of representation of the vague place. No single method may claim to be a general solution and the choice of a method is driven by need. If the exact spatial extent is of lower priority and a fuzzy or graduated view of space is sufficient, a different method is needed as opposed to another case, where the nature of the task demands the presence of a crisp boundary (Humayun and Schwering 2012). Differentiating between user needs grows more complex when in addition to representation, reasoning is to be performed. In other words, for a representation method to be acceptable, its characteristics must correlate with the requirements of a user. Selection of a suitable approach presents a couple of important issues, namely (1) what characteristics of a representation are significant where user requirements are concerned, and (2) how these characteristics can be used to decide upon an appropriate method from several choices. Previous efforts in this direction focus on vague *named places* and do not rely on a formal decision making process to suggest potential modelling techniques (Davies et al. 2009). Others suggest what methods could be used for analysis for different types of uncertainty in Geographic Information Systems (GIS) (Leyk et al. 2005).

The multitude of applications and their varying requirements call for an organized manner in which to specify them. The suggested method here is to utilize the characteristics of the representation methods to lay out the requirements concerning representation of spatial vagueness. These characteristics also allow differentiating between the methods. Before doing this, the relevant characteristics need to be identified in the first place. These form the basis for selection of a competent method. The procedure for selection itself should be consistent, independent of the methods themselves and extensible so as to allow for inclusion of newer and improved representations. AHP is chosen for the decision making process here and satisfies these properties.

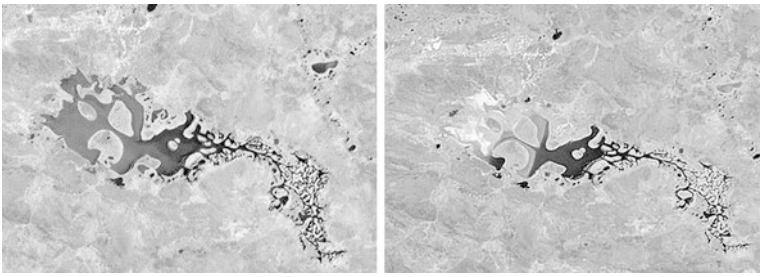


Fig. 1 Lake carnegie, Australia in Apr 2011 (*left*) and Sep 2011 (*right*) (courtesy: U.S. geological survey)

In this chapter we propose a methodology to aid selection of a suitable method for representing vague places. In the Sect. 2 we describe with an example how a vague place may be conceptualized and the need for different representations. Section 3 discusses characteristics in terms of which requirements may be specified. These serve as the criteria for characterizing and analyzing selected representation methods for spatial vagueness in Sect. 4. Comparison between alternative representation methods with respect to criteria and sub-criteria for decision making using AHP is discussed in Sect. 5, followed by a simple demonstration using the application scenarios in Sect. 6. We conclude with a discussion of our approach and identify areas for improvement.

2 Application Scenario

Some lakes are good examples for geographical entities where there is indeterminacy concerning the boundary. Owing to seasonal climatic variations, lakes sometimes undergo extreme variations in their spatial extent. For instance Lake Carnegie (Fig. 1) in Western Australia alternates between a water body and a muddy marsh depending on the amount of precipitation.[1] When there is enough rainfall, the lake is one big contiguous body of water. As the amount of rainfall decreases and the water dries up, the lake is more of a collection of disparate pools of water and is no more continuous. Judging the boundary becomes a problem and brings up questions such as to what extent a muddy marsh is considered a lake, and whether the non-contiguous pools of water are part of the lake. The interpretation regarding the identity of the lake and a suitable representation model which supports this interpretation are dependent on the purpose.

Consider the case of two different users with different requirements for representing the lake:

[1] http://www.nasa.gov/multimedia/imagegallery/image_feature_817.html

1. An analyst intends to model a lake as a crisp body using aerial imagery. The intended use is to individuate the lake and compute how much of it overlaps a certain plot of land.
2. Another model of the lake is intended to gauge the potential of shoreline flooding from a series of water level measurements taken over time. The source of data in this case is measurements obtained from sensors.

Such cases illustrate how variation in requirements means differences in the way the lake is conceptualized, sources of data used and kind of reasoning needed. Before an appropriate method is chosen, clear distinction between possible alternatives is needed. A proposal for possible characteristics followed by an analysis of candidate modelling methods is discussed next.

3 Deriving Representation Characteristics

A representation method is influenced by many factors in addition to the purpose which it is used for. Assumptions made in the way space is conceptualized, sources of data used, the kind of reasoning that is to be performed or the way the place is to be visualized, all have a bearing on the choice of the representational method. To break down the complexity and ensure a consistent way of analyzing requirements, they are abstracted into three different levels (Fig. 2). This is similar to the levels of data modelling in an information system (Longley et al. 2005):

1. The *application domain* level deals with reality. A subset of the real world to be represented is chosen with respect to a particular domain. One identifies here what kind of phenomena are to be represented and what kind of reasoning is to be performed. Of particular interest at this level are the semantics of the place such as 'lake'. Questions such as 'Does a lake require presence of water round the year? Is it a single contiguous body or does it include surrounding pools of water as well?' are asked at this level.

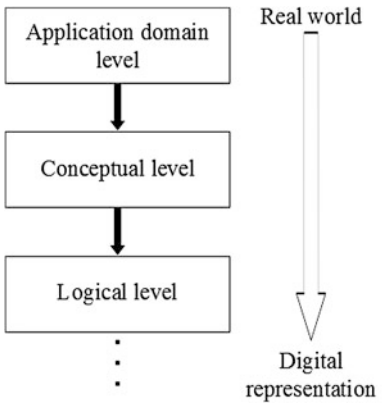


Fig. 2 Levels of abstraction

2. According to Couclelis (Couclelis 1996), the problem of undefined boundaries should be sought neither in the geographic world nor in the GIS representation, but in the cognitive act of transforming geographic entities into database objects. The *conceptual level* attempts to deal with this by conceptualizing the vague place in an implementation independent fashion. Important concerns regarding the place here are—how can it be individuated? (E.g. through use of objectifiable parameters), how it is delineated? and questions regarding its identity (e.g. spatial variation in case of temporal changes).
3. The *logical level* is next and deals with more detailed specifics such as the data model of data sources. Though the logical model itself will be implementation independent, attention must be paid to what kind of underlying formal model will be utilized since this has an influence on how the extents are modelled and what kinds of reasoning can be performed.

One may additionally define lower levels which deal with the data structures and physical storage which actually correspond to the digital representation. Such levels are however too detailed where spatial vagueness is concerned. We limit ourselves to the higher levels of abstraction.

3.1 Criteria for Differentiation

From the levels of abstraction, the following high level characteristics are identified as being decisive in terms of outlining the requirements. They also serve to differentiate between representation techniques:

Conceptualization of space—Varzi (2001) suggests two perspectives on vagueness. *De re* vagueness advocates that indeterminacy of a vague term is ontological and the boundaries of the vague referent are genuinely fuzzy. The *de dicto* reading in contrast treats vagueness as purely semantic. Objects themselves are not vague, but terms used (common nouns such as lake, mountain or proper nouns such as Lake Carnegie or Mt. Everest) vaguely refer to an entity. The adopted perspective of vagueness is domain dependent and affects the choice of the representational and semantic framework (Bennett 2011) because this also has implications on the kind of boundary of the phenomenon (crisp, graduated, indeterminate etc.). Adopting the view that vagueness is intrinsic to an object is better served with a method that preserves the fuzziness, while a semantic view needs a representation that accounts for different possible interpretations of the vague entity (Mallenby 2008). The relation between different representations and the way space is conceptualized arises from the fact that some methods treat space as being *completely indeterminate*, whereas others treat it as being *partially indeterminate* by partitioning it into a certain, uncertain and an excluded region. Another possible refinement is that the space has *multiple crisp interpretations*.

Formal model—Several formal models have been used to manage imperfection in spatial information. Among them are stochastic models, fuzzy set theoretic models, classical logic based and multi-valued logic based models (Duckham M Sharp 2005). A formal model dictates how spatial objects are managed in an information system and influences the kinds of reasoning that can be performed. For instance, methods using classical logic are compatible with the way objects are represented in a spatial database which handles crisp objects (Kulik 2001). This would be applicable in the lake example, if for different interpretations of the lake, each interpretation is treated as crisp which makes it compatible with existing spatial databases (which are founded on classical logic principles). As a consequence, any reasoning which may be performed on crisp data becomes possible. In cases where membership cannot be precisely determined, fuzzy set models and multi-valued logics may be applied. A stochastic model assumes that the indeterminacy is a result of random variations, uses observations with predictable characteristics for modelling (Duckham M Sharp 2005), and is best suited for statistical analysis and reasoning with data. It must also be mentioned here that formal models closely correspond with the way space is conceptualized.

Data model—A data model represents real world features in a form which can be digitally encoded. The term refers here specifically to the spatial data model, which dictates how the geometry of real world features is specified in a logically consistent manner. Commonly employed models include *raster* (data represented as an array of cells with possible attributes for each cell) and *vector* (geometry of discrete objects encoded into points, lines or polygons) (Longley et al. 2005). Alternative data models such as tessellations, networks and object data model also exist. The importance of data models is evident when it comes to employing a method to handle spatial vagueness. Certain methods can only handle regions (which normally use a vector data model), whereas others are well suited for data represented as regularly spaced points or grids (which is the case in rasters). The data model in turn dictates what sources of data can be used with a representation method for vague entities. Additional questions are raised depending on the data models, e.g. how can a crisp boundary be generated in the case of a raster data model?

Reasoning—Reasoning covers processes such as metric, direction and topological operations that are performed on vague places. Since by definition vague places do not have a crisp boundary, a meaningful metric operation will require a crisp instantiation to be performed first. Directional operations cover directional relations (north of, south, north-east etc.) between spatial objects. Topological operations allow reasoning between configurations of two *vague* places or a *vague* and *crisp* place. Reasoning is particularly important from the perspective of a task, since it limits what kind of analysis is done on the vague place. In addition to qualitative reasoning, quantitative reasoning such as statistical reasoning is also possible. Some representations provide a well-defined framework for a particular type of reasoning compared to others.

4 Analyzing Representation Methods

Though a number of methods and specialized improvements upon them exist in order to handle spatial vagueness, we restrict ourselves to the commonly cited methods in literature. An analysis of the methods with respect to the criteria above serves as a starting point with which to distinguish between them.

4.1 Base Representations

This term is coined to refer to those methods which abstract or crisp the vague place. They do not represent vagueness in any form and are included here solely for comparison although they are not used in the decision making process later. They prove adequate in some cases and a precise location sometimes serves to anchor an indeterminate region upon which more complex models to handle vagueness may be added (Galton and Hood 2005). Reasoning is performed by simply treating them as crisp objects.

Following can be classified as base representations:

- A vague place reduced to a simple feature with crisp geometry (point, line or polygon). This can be seen in gazetteers, which use a representative point to list a place of interest. Vague places tagged in volunteered geographic information (VGI), where a real world entity is outlined by contributors (from GPS tracks or tracing from aerial imagery), are another instance where this is seen.
- Entities defined a priori according to some norm. A certain data provider or designer of a system might decide to classify a particular vague place as crisp based on some definition, convention, property or metric (as seen in ordinary maps).
- Use of a minimum bounding rectangle (MBR) which covers the maximal extent of the space where the entity is located.
- Tessellation of space. Entities such as mountains can be outlined from a digital elevation model this way.

4.2 Probabilistic Methods

Probabilistic methods are widely used to handle uncertainty, especially in the case of positional or measurement uncertainty. These methods derive the membership value of an individual in a set through a statistically defined probability function (Erwig and Schneider 1997). They are best suited for continuous features with measurable objective properties such as flow, temperature, water level, or height which can be sampled at certain locations. The advantage of probabilistic methods is that the observed properties relate directly to the entity itself rather than being subjective.

Probabilistic methods have been employed to model city centres, based on probability of sample points computed from trials using participant studies (Montello et al. 2003). However since vagueness occurs despite accurate knowledge of the state of the world, the utility of probabilistic methods to model vagueness is arguable.

4.3 Fuzzy-Set Model

These are based on Fuzzy-set theory and ideal for modelling objects which have graduating or indeterminate boundaries (Fisher et al. 2004; Usery 1996; Wang and Hall 1996). Fuzzy models are based on the subjective assignment of a graded value usually between 0 and 1 indicating the degree of membership α to the set. Only the ordinal properties of the numbers are made use of; fuzzy methods are qualitative rather than quantitative (Worboys and Duckham 2004). It is also possible to obtain a crisp boundary by means of α-cuts which are a way of obtaining crisp sets from a fuzzy set. The threshold value α determines which points are excluded from a region. Fuzzy methods have generally been employed to characterize landscapes which exhibit variation in their attributes such as elevation, to model vague entities such as mountains (Fisher et al. 2004).

4.4 Egg-Yolk Model

The egg-yolk model (Cohn and Gotts 1996) is useful for reasoning when the boundaries of a region are vague, but the location of the region itself is not vague. The vague region becomes a composition of two crisp regions analogous to an egg; the yolk corresponds to the minimal extent, and is surrounded by the concentric indeterminate region akin to the white. However the theory does not make any assertions as to the acceptability of the crisping of these regions. Defined in this fashion, the method allows performing qualitative spatial reasoning between two vague regions or between a vague region and a crisp region under the framework of the region connection calculus (RCC-5). Reasoning is possible on different possible configurations of two regions represented this way. Another comparable approach treats vague regions as having *broad boundaries* and provides an algebraic treatment of topological relations using the 9-intersection model (Clementini and Di Felice 1996).

4.5 Rough Set Model

Similar to the egg-yolk method, rough set model relies on a three valued logic to operate on finitely partitioned space (Pawlak 1982). A vague region is represented

in terms of its *lower approximation* and *upper approximation* corresponding to members that are 'definitely in' or 'possibly in' the region respectively. This is especially useful when datasets at different resolutions are to be reasoned with (Worboys 1998). Topological reasoning can also be performed by defining the approximations in terms of RCC-5 relations (Vögele et al. 2003).

4.6 Supervaluation

Supervaluation (Kulik 2001) is a method useful to handle indeterminacy in the semantic conception of vagueness. For a given vague object, there may be many possible ways to make it precise depending on different interpretations. An admissible interpretation is termed a *precisification*. Every precisification divides the underlying space into two regions, one which definitely belongs to the vague region and the other which does not. To determine truth or falsity of a given proposition it is valuated over all precisifications. A proposition is supertrue if it is true for all precisifications. Likewise, it is superfalse, if it is false for all precisifications.

Supervaluation theory itself does not specify what constitutes a valid interpretation. The underlying formal model employs classical logic making it usable with existing spatial information systems and databases. Computational applications are however hampered due to the practical difficulty in explicit specification of all possible precisifications (Bennett 2011).

4.7 Comparing the Methods

Table 1 compares the different methods according to our selected criteria. Although the models differ in many aspects, they also share similarities and it is sometimes possible to transform one model into another. For instance, the yolk of the egg is conceptually equivalent to the lower approximation in rough sets or the supertrue region in supervaluation. Likewise, the white of the egg is similar to the upper approximation of rough sets or that region in supervaluation which is neither supertrue nor superfalse. A similar analogy may be found for fuzzy sets. Consequently, shortcomings in a particular method may be overcome by transforming it into another which supports it. However, further discussion on this topic is beyond the scope of this chapter.

Table 1 Comparing methods to represent spatial vagueness

Method	Conceptualization of space	Formal model	Data model	Reasoning
Base representations	Crisp boundaries	None	Both raster and vector	Reasoning techniques for crisp objects are applicable
Probabilistic methods	Continuous field with objectively measurable property. Space is completely indeterminate.	Stochastic	Vector (sampled points), raster (sampled grid)	Statistical reasoning and analysis
Fuzzy sets	Completely indeterminate space with graduated or indeterminate boundaries	Fuzzy logic	Both raster and vector	Fuzzy set theoretic operations, topological reasoning using fuzzy RCC
Egg-yolk model	Partially indeterminate—core surrounded by indeterminate region	Three-valued logic	Vector (region)	Topological reasoning with RCC-5
Rough set	Partially indeterminate—determinate lower approximation and indeterminate upper approximation in finitely partitioned space	Three-valued logic	Resolution (finitely partitioned space)—either raster or a tessellation	Topological reasoning, reasoning with imprecision in multi-resolution datasets
Supervaluation	Multiple valid crisp interpretations	Classical logic	Vector (region)	Topological reasoning (esp. in case of gradual transition of boundaries between two regions)

5 Decision Making Using the Analytic Hierarchy Process

5.1 Overview of the Analytic Hierarchy Process

The analytic hierarchy process (AHP) (Saaty 2008), used when one has an objective and several alternatives to achieve it, is utilized here as a form of multi-criteria decision analysis to select a method of representation. The final decision is made based by calculating priorities of the alternatives with regards to chosen criteria. The advantage of AHP is that it allows decision makers to exercise their judgments in addition to the underlying information. The process can also be revisited in order to revise the final outcome. The overall procedure in AHP may be summarized as follows (Saaty 2008):

Step 1 Define the problem and determine what kind of knowledge is sought.
Step 2 Structure a *decision hierarchy* with the goal at the top, alternative candidates at the lowest level and the criteria and sub-criteria at the intermediate levels.
Step 3 Perform *pairwise comparison* of elements at each level with respect a node in the immediate upper level. Pairwise comparison matrices are used for this purpose.
Step 4 Compute priorities for elements from comparison and use this to weigh the priorities in the immediate lower level. Repeat this for every element. For each element in the lower level, add the weighed values to obtain its overall priority. Continue the process until priorities for the alternatives at the lowest level are computed.

The decision hierarchy defined is shown in Fig. 3. The goal is to choose a method to represent the vague place. The four criteria discussed in Sect. 3.1 are deemed important in order to reach the goal. The criteria are further partitioned into sub-criteria. The conceptualization of space, for instance, is sub-divided into *completely indeterminate, partially indeterminate and multiple interpretations*. The alternative approaches to model spatial vagueness discussed in the previous section form the lowest level of the hierarchy.

5.2 Pairwise Comparison of Representation Methods

Once the hierarchy is defined, three sets of pairwise comparisons are carried out.

1. between the *criteria* with respect to the *goal*
2. between the *sub-criteria* with respect to the *criteria*, and
3. between the *alternatives* with respect to the *sub-criteria*

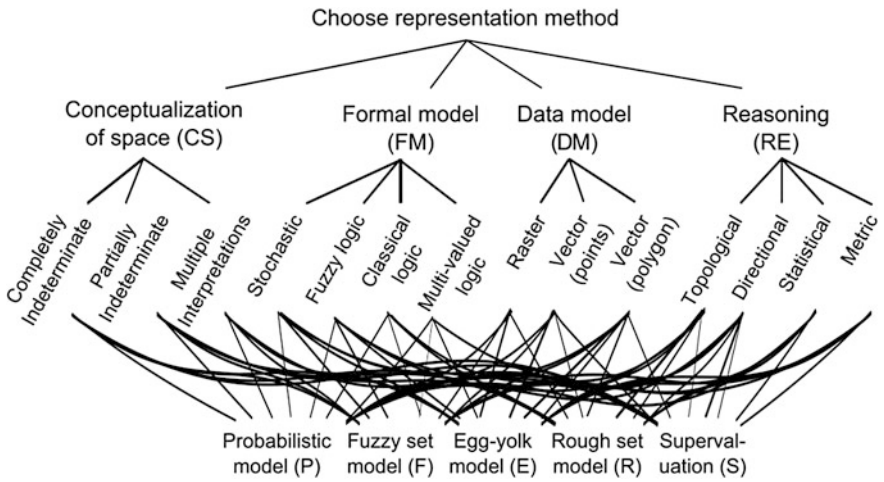


Fig. 3 AHP hierarchy for spatial vagueness representation

Comparisons (1) and (2) are dependent on the purpose of representation. Here the user can exercise judgment over what is more important for a given scenario. For instance, if reasoning is more important for an application, then it is assigned a greater weight than other criteria. Within reasoning, topological reasoning gets a higher value if it is deemed more important than other forms of reasoning for the same application. Comparisons in (3) are independent of the purpose of application and are assigned values based on conclusions drawn from our analysis of the different modelling techniques (Sect. 4).

A vector of weights is obtained for each comparison. Once pairwise comparisons have been completed at all levels, weights computed for nodes at the higher levels are propagated down the hierarchy and combined with weights at the immediate lower level to yield a global weight which signifies the priority for that node. This continues until the final priorities for the alternative methods are computed. Alternatives are ranked by descending order of their global weight i.e. greater the weight, higher the rank. Pairwise comparisons make use the fundamental scale defined for AHP (Saaty 2008). When comparisons are made between two elements in the hierarchy with respect to another, a value is assigned to one of them and the other gets the reciprocal value. In our decision making process, we distinguish five levels of priorities (and their reciprocals):

1- Both the elements contribute equally to the objective.
3- One element contributes slightly more than the other.
5- One element strongly contributes more than the other.
7- One element very strongly contributes more than the other and this is demonstrated in practice.

9- One element contributes more than the other to the highest possible order of affirmation. This also indicates the other method does not support a criterion at all (e.g. fuzzy set model does not use classical logic).

A sample pairwise comparison between the alternative representation methods with respect to the sub-criterion *completely indeterminate* (CI) is shown below as matrix M_{CI} (names of alternatives are given by their first character. Probabilistic model is designated P and so on as seen in Fig. 3). Its eigenvector W_{CI} shows the corresponding relative weights for each of the alternatives with respect to the chosen criterion.

$$M_{CI} = \begin{array}{c} \\ P \\ F \\ E \\ R \\ S \end{array} \begin{array}{c} \begin{array}{ccccc} P & F & E & R & S \end{array} \\ \begin{bmatrix} 1 & 1 & 9 & 9 & 9 \\ 1 & 1 & 9 & 9 & 9 \\ 1/9 & 1/9 & 1 & 1 & 1 \\ 1/9 & 1/9 & 1 & 1 & 1 \\ 1/9 & 1/9 & 1 & 1 & 1 \end{bmatrix} \end{array} \quad W_{CI} = \begin{bmatrix} 0.43 \\ 0.43 \\ 0.04 \\ 0.04 \\ 0.04 \end{bmatrix}$$

The comparison above shows how well a method compares against another when the space is completely indeterminate. The following rationale lies behind assignment of values for the above comparison:

- The principal diagonal consists of comparisons of an alternative with itself and the value is always 1.
- Probabilistic and fuzzy set models are assigned a value 9 against egg-yolk model, rough set model and supervaluation because the latter methods assume that space is partially indeterminate (or completely crisp as in supervaluation) (refer Table 1).
- Both probabilistic and fuzzy methods are equally good alternatives when the space is completely indeterminate, and obtain the value 1.
- The value 1/9 is the reciprocal and conversely implies the method is not preferred over the other alternatives.

The relative ranking of the alternatives is the eigenvector of the pairwise matrix and each row signifies the weight associated with the corresponding row of the matrix M_{CI}. The weights reflect the above judgments and both probabilistic and fuzzy set methods have a considerable priority over other alternatives when the space is completely indeterminate. These are combined with weights from the upper levels in order to derive the global weighting which is then used to rank the alternative.

While the criteria and sub-criteria defined here are not extensive, they are nonetheless critical factors for modelling of spatial vagueness. The alternatives for spatial vagueness representation are also limited to the commonly used ones as mentioned earlier. However the decision making process we use enables criteria and alternatives to be added or removed by simply altering the hierarchy.

6 Application of AHP and Discussion

The application scenario (Sect. 2) is revisited here. In scenario 1, the goal is to (1) obtain a representation of a lake, and (2) perform topological reasoning upon it, using (3) a satellite image. Expressing these requirements in terms of the criteria, let us say that representing the way space is conceptualized (CS) is the most important objective. Performing topological reasoning which is a sub-criterion of reasoning (RE) is the next priority. The data model (DM) is equally important since the representation is to be obtained from a raster image. These three criteria override the requirement for a specific formal model (FM), which is consequently deemed the least important. These pairwise comparisons of these criteria with respect to the objective are shown in the matrix M_{goal1} along with their computed weights W_{goal1}.

$$M_{goal1} = \begin{array}{c} \\ CS \\ FM \\ DM \\ RE \end{array} \begin{array}{c} \begin{array}{cccc} CS & FM & DM & RE \end{array} \\ \begin{bmatrix} 1 & 5 & 3 & 3 \\ 1/5 & 1 & 1/3 & 1/5 \\ 1/3 & 3 & 1 & 1 \\ 1/3 & 5 & 1 & 1 \end{bmatrix} \end{array} \quad W_{goal1} = \begin{bmatrix} 0.51 \\ 0.07 \\ 0.19 \\ 0.22 \end{bmatrix}$$

Pairwise comparison is similarly performed for each of the sub-criteria. In this scenario, space is considered partly indeterminate since there is a region which is definitely a lake along with some uncertain region. The data model used is a raster and topological reasoning has priority. Pairwise comparisons at the final level of hierarchy between the alternatives and the sub-criteria are determined by analysis of the models and remain the same for both scenarios. Space limitations allow us to show only pairwise comparisons between the goal and criteria.

The next scenario is focused more on the probabilistic nature of the vague place. The goal is to (1) represent the lake, in order to (2) compute the probability that a given location is flooded, using (3) water level observations from sensors. Expressing these in terms of criteria, the conceptualization of space is as important as the reasoning (statistical) ability and forms the main objective. Data model is the next priority since the observations sampled at different locations are used here but other sources could have been used as well. As long as the objectives are met the user is not concerned with the underlying formal model and for this reason it is least prioritized. The pairwise comparisons with respect to the goal and corresponding weights for this scenario are shown below in the matrices M_{goal2} and W_{goal2} respectively.

$$M_{goal2} = \begin{array}{c} \\ CS \\ FM \\ DM \\ RE \end{array} \begin{array}{c} \begin{array}{cccc} CS & FM & DM & RE \end{array} \\ \begin{bmatrix} 1 & 5 & 3 & 1 \\ 1/5 & 1 & 1/3 & 1/5 \\ 1/3 & 3 & 1 & 1/3 \\ 1 & 5 & 3 & 1 \end{bmatrix} \end{array} \quad W_{goal2} = \begin{bmatrix} 0.39 \\ 0.07 \\ 0.15 \\ 0.39 \end{bmatrix}$$

Table 2 Ranking of priorities for application scenarios

Alternative	Scenario 1	Scenario 2
Probabilistic	0.17 (4)	0.33 (1)
Fuzzy set model	0.25 (1)	0.25 (2)
Egg-yolk model	0.24 (2)	0.12 (3)
Rough set model	0.21 (3)	0.13 (4)
Supervaluation	0.13 (5)	0.11 (5)

Synthesizing our decision model involves multiplying the priority of nodes with those of their parent node and adding the resulting weight for each alternative to obtain the final priority. This yields the final priorities of the alternatives with respect to the goal. The computed priorities for the alternatives and their corresponding rank (in brackets) for each of the scenarios are shown in Table 2.

For scenario 1, the fuzzy set model seems to offer the best fit. This conforms to the characteristics of the fuzzy method itself since it is flexible with regards to the conceptualization of space, supports topological operations and is able to use raster the data model. The egg-yolk model is ranked as the next best choice although the model requires certain and uncertain regions to be defined and is mostly applicable for vector data models. The ranking may be explained due to the fact that the space in this scenario is conceptualized as partially indeterminate and the fact that egg-yolk models are well suited for topological reasoning which skews the results in its favour. Similarly, in scenario 2, the importance assigned to statistical reasoning means probabilistic method is the most preferred candidate compared to others. Probability that a random location is flooded can be estimated by interpolation. Fuzzy methods can also be applied in this case. The observed values can be translated into an equivalent fuzzy membership and further reasoned upon. This is a subjective process and the results may or may not correspond with those from the probabilistic method.

One of the drawbacks of using such a decision making process is that the methods are assigned a rank even though they might not be applicable. Good scores on some criteria also tend to compensate the bad scores on others. For instance, in the latter scenario the egg-yolk model is not applicable directly though it ranks third among the alternatives. The fact that a method is assigned a priority based on its calculated weight does not necessarily mean it is applicable. It only means that a subset of the requirements is addressed by the method. The best fitting method gets the highest rank. If none of the models are applicable, then the user would have to reconsider the requirements, e.g. search for a different data source etc. The number of pairwise comparisons also grows as alternatives or criteria are added to the hierarchy.

The decision making process we introduce here makes some assumptions. We assume that the models are mutually independent and that a vague place is modelled solely using one of the alternatives. This might however differ in practice. It is possible to use fuzzy modelling to partition a region into determinate and indeterminate. One can then use egg-yolk method upon this to perform topological

reasoning. Such solutions are not accounted for at present. When two different methods are used in conjunction this way, it may be treated as a two-step decision process, the first focusing on the space, the second on the reasoning ability.

7 Conclusion

Representation of vague places in information systems can be carried out using multiple methods. To aid selection of a suitable method, the contributions of this chapter can be summarized as follows. First, relevant criteria which provide a basis for comparison between different methods are identified by examining modelling requirements at differing levels of abstraction. Next, we consider some commonly cited methods in literature and analyze the methods with respect to the identified criteria. From an analysis of the methods with respect to the criteria, we are able to compare them. Finally, a formal decision making process for selecting a representation based on criteria is presented. AHP is used for this purpose and the applicability and usefulness of AHP of our approach are demonstrated with two examples.

Future work will consider models which present a different way of conceptualizing space such anchoring which constrain the vague region by anchoring it to a known location (Galton and Hood 2005). Use of expert opinion in pairwise comparison of alternative methods will increase the reliability of comparison and is being considered. Presently this is subject to results from our analysis. Further improvements are possible in the form restricting the rankings only to applicable models rather than considering every alternative as a potential choice.

Acknowledgments This research is funded by the German Research Foundation (DFG) as part of the International Research Training Group on Semantic Integration of Geospatial Information (IRTG-SIGI, GRK 1498).

References

Bennett B (2011) Spatial vagueness. In: Jeansoulin R, Papini O, Prade H, Schockaert S (eds) Methods for handling imperfect spatial information, Springer, pp 15–47

Clementini E, Di Felice P (1996) An algebraic model for spatial objects with indeterminate boundaries. Geogr Objects Indeterminate Boundaries 2:155–169

Cohn A, Gotts N (1996) The 'egg-yolk' representation of regions with indeterminate boundaries. Geogr Objects Indeterminate Boundaries 2:171–187

Couclelis H (1996) Towards an operational typology of geographic entities with ill-defined boundaries. In: Burrough PA, Frank AU (eds) Geographic objects with indeterminate boundaries. Taylor & Francis Inc, Bristol, pp 45–56

Davies C, Holt I, Green J, Harding J, Diamond L (2009) User needs and implications for modelling vague named places Spatial Cognition and Computation. Interdisciplinary J 9(3):174–194

Duckham M Sharp J (2005) Uncertainty and geographic information: Computational and critical convergence. In: Re-presenting GIS, Wiley, pp 113–124
Erwig M, Schneider M (1997) Vague regions. In: Scholl M, Voisard A (eds) Advances in spatial databases, lecture notes in computer science, vol 1262, Springer Berlin/Heidelberg, pp 298–320
Fisher P, Wood J, Cheng T (2004) Where is Helvellyn? Fuzziness of multi-scale landscape morphometry. Trans Institute British Geogr 29(1):106–128
Galton A, Hood J (2005) Anchoring: a new approach to handling indeterminate location in GIS. In: Cohn AG, Mark DM (eds) Spatial information theory, lecture notes in computer science, vol 3693, Springer Berling, pp 1–13
Humayun MI, Schwering A (2012) Representing vague places: Determining a suitable method. In: Vasardani M, Winter S, Richter KF, Janowicz K, Mackaness W (eds) Proceedings of the international workshop on place-related knowledge acquisition research (P-KAR 2012), Monastery Seeon, Germany, vol 881, pp 19–25
Kulik L (2001) A geometric theory of vague boundaries based on supervaluation. In: Montello D (ed) Spatial information theory, lecture notes in computer science, vol 2205, Springer Berling, pp 44–59
Leyk S, Boesch R, Weibel R (2005) A conceptual framework for uncertainty investigation in map-based land cover change modelling. Trans GIS 9(3):291–322
Longley PA, Goodchild MF, Maguire DJ, Rhind DW (2005) Geographic information systems and science, 2nd edn Wiley
Mallenby D (2008) Handling vagueness in ontologies of geographical information. PhD thesis, School of Computing, University of Leeds
Montello DR, Goodchild MF, Gottsegen J, Fohl P (2003) Where's downtown?: Behavioral methods for determining referents of vague spatial queries. Spatial Cognition Comput 3(2):185–204
Pawlak Z (1982) Rough sets. Int J Comput Inf Sci 11(5):341–356
Saaty TL (2008) Decision making with the analytic hierarchy process. Int J Serv Sci 1(1):83–98
Usery EL (1996) A conceptual framework and fuzzy set implementation for geographic features. In: Burrough PA, Frank AU (eds) Geographic objects with indeterminate boundaries. Taylor and Francis, London
Varzi AC (2001) Vagueness in geography. Philos Geogr 4(1):49–65
Vögele T, Schlieder C, Visser U (2003) Intuitive modelling of place name regions for spatial information retrieval. In: Spatial information theory, lecture notes in computer science, vol 2825, Springer Berling, pp 239–252
Wang F, Hall GB (1996) Fuzzy representation of geographical boundaries in gis. Int J Geogr Inf Syst 10(5):573–590
Worboys M (1998) Imprecision in finite resolution spatial data. GeoInformatica 2(3):257–279
Worboys MF, Clementini E (2001) Integration of imperfect spatial information. J Visual Lang Comput 12(1):61–80
Worboys M, Duckham M (2004) GIS: a computing perspective, 2nd edn CRC Press

Provenance Information in Geodata Infrastructures

Christin Henzen, Stephan Mäs and Lars Bernard

Abstract When it comes to usability evaluation of geodata information about its provenance or lineage are vital. Nevertheless, in practice the corresponding metadata elements are often neglected. Even if available, the tabular or listed metadata representations in current metadata catalogue user interfaces do not sufficiently support the users browsing and comparing metadata. This chapter proposes an interactive application for data provenance visualization called MetaViz. As a foundation for the MetaViz design the chapter provides a detailed analysis on modeling aspects, available standards and specifications for data provenance and presents possible design and implementation choices. A scientific geodata infrastructure that supports researchers sharing results of numerical simulations of different environmental phenomena serves as the underlying use case.

1 Introduction

In geodata infrastructures (GDI) metadata is meant to support (1) discovery, (2) evaluation and (3) integration of heterogeneous geodata sources. However, most of today's geocatalogue and geoportal developments primarily focus only on discovery aspects. Once also data evaluation gets into the focus data provenance becomes of major interest: Learning about a dataset's history, its origin, its

C. Henzen (✉) · S. Mäs · L. Bernard
Professorship of Geoinformation Systems, Technische Universität Dresden Department of Geosciences, Helmholtzstraße 10, 01069 Dresden, Germany
e-mail: Christin.Henzen@tu-dresden.de

S. Mäs
e-mail: Stephan.Maes@tu-dresden.de

L. Bernard
e-mail: Lars.Bernard@tu-dresden.de

D. Vandenbroucke et al. (eds.), *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography,
DOI: 10.1007/978-3-319-00615-4_8,

previous treatments and potentially experiences in using it are crucial aspects in assessing whether and how a considered dataset might fit for an application (Di and Yue 2011; Simmhan et al. 2005; Moreau 2010). Besides supporting usability assessments information about data provenance facilitates transparency, maintenance documentation and might even ensure reproducibility (Di and Yue 2011; Glavic and Dittrich 2007; Osterweil et al. 2010).

In current geoinformation metadata standards (ISO 19115, INSPIRE) provenance descriptions are defined using elements for a textual lineage description and if more detailed by providing references and free text documentations of data sources and data creation processes. There is not only a lack in harmonised vocabularies on describing data provenance; it also shows in practice, that creation of these metadata elements is often neglected.

Yet another issue in supporting metadata based evaluation of geodata is the way metadata is presented in geocatalogues. Problems such as the absence of customizable detail levels as well as the lack of effective communication methods for metadata contents are obvious (cp. Bowers 2012; Malaverri et al. 2012; Kindermann et al. 2007). Further, geocatalogues and geoportals mostly do not offer suitable and compact representations of the evaluable metadata. Metadata is typically presented in user interfaces consisting of long lists or tables that do not support user-friendly navigation, browsing or guidance through the metadata or even interactive analysis and comparison of metadata sets (cp. Fisher et al. 2009; Bowers 2012; Zargar 2009). Convincing (visual) inspection tools, showing how to make use of provenance information for geodata and thus motivating the provision of such metadata can hardly be found.

Focussing on data provenance this chapter proposes an interactive application for metadata visualization called MetaViz. The presented solutions and scenarios stem from the development of a Scientific GDI to support researchers in sharing input and results of numerical simulations of different environmental phenomena. The chapter provides a detailed analysis on the current state in describing data provenance as a basis for the design of MetaViz. Then implementation details and functionality of MetaViz and its integration in the GDI environment are presented. A discussion of the achieved results will help to identify future research and development needs.

2 Aspects of Provenance

Generally provenance metadata informs the user about the history of a data product, source data and processes (Moreau 2010; Simmhan 2005; Di and Yue 2011). It offers the possibility to document the data origin by references but without a need to publish all interim results or any potentially access restricted input data itself. The term provenance is often used synonymously with the terms lineage or pedigree. Additional data usage documentations describe concrete applications of the data (e.g. visualizations or analysis) or link to further processed data products.

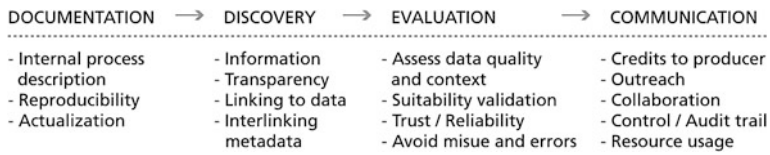


Fig. 1 Purposes of provenance information

The latter can be seen as a different view on lineage information focusing on the derived data products instead of the data history. Therefore, usage information can be deduced from provenance information and is typically described in the same schema. In this chapter usage is conceived as derivation of data products, leaving out the concrete applications of data.

The main purposes of provenance information can be categorized into documentation, discovery, evaluation and communication (Fig. 1). These categories also mirror the process steps from metadata acquisition to usage and communication of results. From the data producers perspective provenance information documents internal processes and might even facilitate the reproducibility. Discovery and evaluation correspond to the general purposes of metadata, that is enabling the user to find data and to assess whether the dataset suits the requirements of an application or not.

One well-known issue in assessing data is trust. Naturally trust strongly relates to data producers and their trustworthiness or reputation (Malaverri et al. 2012; Di and Yue 2011). Furthermore, trust can be raised by applying more formal methods such as data validation or by enabling a reproduction of data product on providing all required provenance information. Additional provenance data also helps to avoid misuse or misinterpretation of data and thus ideally helps in preventing from incorrect data usages (Devillers et al. 2005; Malaverri et al. 2012).

Another important purpose of provenance is the communication of credits to contributors of source data, and derivation algorithms or processes, outreach documentation and indicator in audit trails. The derivation of the usage is not only interesting for further data users and controller but also for the data producer.

The following subchapters provide an in-depth analysis of the different characteristics in modelling, presenting and standardizing provenance data. These characteristics will then be used to (1) to classify related work and (2) to describe the design and implementation of MetaViz.

2.1 Modelling Provenance

Describing provenance of information resources is a well-researched topic that can be examined from different perspectives such as several modelling or classification concepts, architectures and standards as well as user requirements and interfaces.

Table 1 Modeling aspects of provenance

Subjects/entities	Data		Processes	
Granularity/level of detail	Coarse (e.g. dataset level)	Fine (class or attribute level)	Coarse (e.g. only one process between datasets, without sub processes)	Fine (e.g. detailed workflows with sub processes)
Representation	Directed to provenance	Directed to usage	Bidirectional	
	Successive process steps		Complete process sequence	

Due to the application domain and specific user requirements provenance information can be modelled data- or process-oriented. Data provenance describes the history of a data product on a fine-grained level using classes and relationships (Table 1). Spéry et al. (2001) developed such a fine-grained data provenance model that is used to capture manipulations on the feature-level of spatio-temporal data. Vert et al. (2002) and Pastorello et al. (2005) analyse web-based file and document management of GIS data and define coarse-grained models for data provenance that use the document as highest granularity stored together with adapted FDGC metadata in a database or managed via services.

Process provenance, sometimes called service provenance, focuses on detailed information about the workflow and corresponding sub processes facilitating the reproduction (cp. Osterweil et al. 2010). In some cases data provenance can be deduced from process provenance, e.g. by omitting the process information and only showing data derivations (cp. Simmhan et al. 2005).

It is also possible to derive a coarse-grained provenance model from a fine-grained one, which induces different views of the data model Visualizing provenance information on different levels of granularity allows the user to get a brief summary of provenance or a very detailed view, for example on attribute level. Provenance can generally be represented either as separate successive processes or as the complete sequence of all processes:

- If represented in separated process steps each entity only contains information about the processing of the prior dataset (direct predecessor). Thus lineage information is stored step-wise in several linked metadata sets.
- Provenance can be modelled as complete provenance with fine or coarse granularity. In contrast to the successive provenance representation, the complete provenance contains information about the whole lineage process. Within a chain of processed data, provenance descriptions are stored redundant in the metadata of the data and in the metadata of its successor.

Another aspect of provenance modelling is the representation of direction or navigation links: Some provenance models only provide backward links to origin processes and source data. Others focus on usage and link only to derived data and the respective processes. Usually the missing direction can be deduced. In a

Table 2 Options for system design of provenance GUI

System implementation and architecture			
Application domain	Web	Desktop	
Storage	Tightly coupled with data	As part of the metadata	Separated storage systems
Data interchange	Standard interface	Proprietary interface	
Infrastructure	Distributed environment (service-based)	Standalone application	

bidirectional representation no further processing is needed, but the metadata storage might be redundant.

2.2 Exploring Provenance Data

Evaluating the fitness for use with the help of provenance information does not only depend on the underlying metadata model but also on the information representation techniques. The basic options for the design of a GUI representing provenance data can be discriminated in being either visual representations or being part of the query structures. Provenance visualizations such as trees or directed acyclic graphs are often used to illustrate processing workflows (Anand et al. 2010; Cheung and Hunter 2006) or linked data whereas textual descriptions are typically used in metadata catalogue systems such as GeoNetwork.[1] In such systems queries are usually formulated in textual form, like search terms or keywords. Other more formal querying methods in provenance user interfaces are textual as well as graphical query languages, such as Query Language for Provenance (QLP) (Anand et al. 2010) or Little-JIL (Cass et al. 2000).

When analysing approaches based on the implementation characteristic distinctions are the provenance storage, interchange format and infrastructure. Storage of provenance data can be either coupled with data or with metadata holdings (Di and Yue 2011) (Table 2). The latter can for example be done in geocatalogues following the Catalogue Service for the Web (CSW) interface (OGC 2007) to provide standardized access to geometadata storage systems.

Managing provenance in service-oriented architectures is a pressing challenge (Wang et al. 2008; Yue et al. 2011; Kindermann et al. 2007; Di and Yue 2011) and approached in different ways. Wang et al. developed a three-tier architecture with a web service layer that handles storing, searching and browsing requests, a logic layer and a repository layer that contains the spatial data store and the separated semantic repository. Yue et al. (2011) extended a geospatial metadata catalogue to manage data and service provenance. Automatic metadata generation during data

[1] http://geonetwork-opensource.org

production or process execution, actualization and exchange has been discussed as an additional aspect of provenance (Di and Yue 2011).

Further approaches on geoinformation provenance management such as GeoOpera, GOOSE, ESSW and Geolineus are reviewed by Bose and Frew (2005). Research in provenance of geoinformation mostly addresses modelling and implementation aspects of workflow management or origin and processing of sensor observations. Only a few approaches also address the graphical representation of provenance. Lanter (1991), for instance, introduced a graphical language and user interface for layer-based geographic data. The interface displays an interactive flow diagram and focusses data provenance but does not address the processing steps. Current investigations on provenance of geodata do either focus on interactive visualizations or on using provenance standards.

Outside the geoinformation domain several visualisation methods for provenance data can be found. There are provenance clients such as Provenance Browser (Anand et al. 2010) or Provenance Explorer (Cheung and Hunter 2006) as well as graphical languages, e.g. Little-JIL (Fig. 2a) (Cass et al. 2000). These approaches suggest context-sensitive provenance views, as a data-dependency view or an invocation graph and also present different granularity levels for provenance graphs.

2.3 Standards and Specifications

A provenance information model should follow a standard, but should also be tailored to the specific context (Malaverri et al. 2012; Di and Yue 2011). Geosciences mostly lack appropriate provenance metadata as well as suitable standards (Tilmes 2008; Di and Yue 2011; Yue et al. 2011).

A generic and thematically independent provenance model is the Open Provenance Model (OPM) (Moreau et al. 2011), which specifies a provenance model in a technology-agnostic manner. The three basic elements of this specification are (1) artefacts that describe entities e.g. datasets, (2) agents as a kind of controlling units and (3) processes. Moreover OPM realizes a role as well as a view concept dealing with hierarchical and overlapping accounts.

Another commonly used provenance model is the qualified Dublin Core.[2] Being very compact and producer focussing it is typically used to model web resources. Standards used to describe lineage information in GDI are ISO 19115 part 2 (ISO 2005) and the FGDC Content Standard for Digital Geospatial Metadata (CSDGM).[3] Both define basically the same entities to describe data provenance and only differ in naming conventions. While some researchers argue about too static views and technical names (Fisher et al. 2009; Zargar 2009) these standards

[2] http://dublincore.org/documents/dcmi-terms/

[3] http://www.fgdc.gov/metadata/geospatial-metadata-standards#csdgm

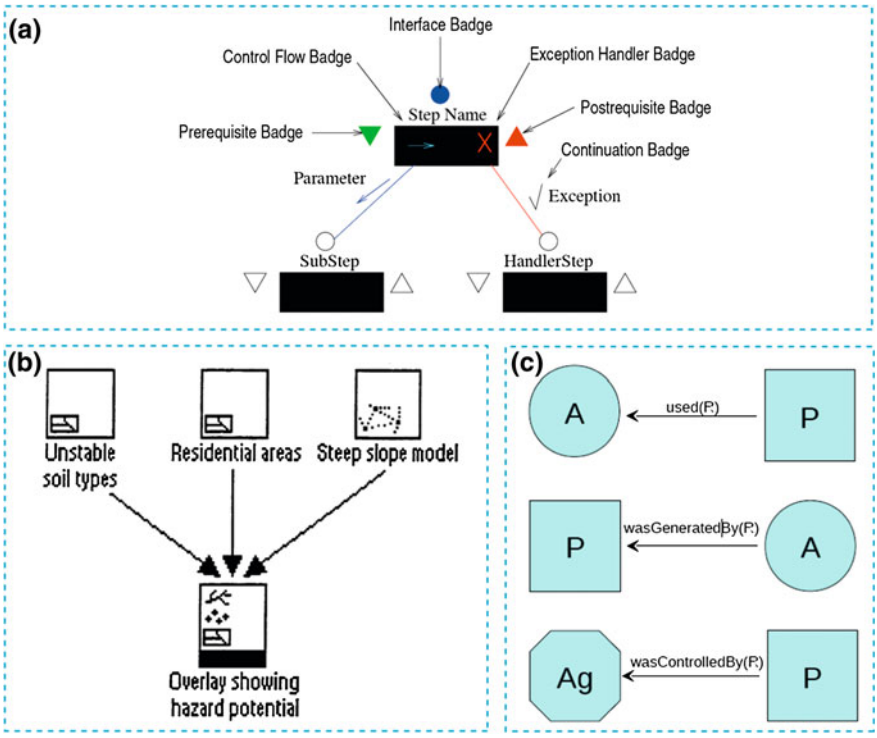


Fig. 2 Provenance visualizations. **a** Legend of the language Little-JIL (Cass et al. 2000). **b** Flow diagram with intermediate and product layer (Lanter 1991). **c** Provenance graph in OPM-Notation (Moreau 2010)

are often applied and need only a few adaptations to support the concepts of OPM, because they are defined more precisely focussing on data provenance of geodata.

All specifications, apart from Dublin Core, represent provenance information at fine or coarse granularity (Table 3)[4] and allow different views on the provenance.

To harmonize these specifications their main elements have been identified (Table 4). Thus, a process step is described by inputs and outputs, a process or model description and the data producer. It is sometimes controlled or facilitated by agents (Moreau et al. 2011) and further explained in (separate) documentations. The specifications do not always allow a one-to-one mapping (Nogueras-Iso et al. 2005), but a transformation between them is possible.

Sometimes specifications get combined: Malaverri et al. (2012) introduce a coarse-grained data provenance model based on a combination of OPM and FGDC illustrated by a use case on map generation. They analyse several quality indicators, such as timeliness of data or reputation of data provider with regard to trustworthiness.

[4] http://dublincore.org/documents/dcmi-terms/#terms-provenance

Table 3 Data provenance concepts in metadata and provenance standards

Provenance standard/ specification	ISO 19115-2 Lineage subclasses	FGDC CSDGM	OPM	Qualified Dublin Core
Concept definition	"Specify lineage of imagery and gridded data datasets" (ISO2005)	"Description of the source material […] and the methods of derivation […] (for the) digital files" (FGDC 2000)	"Represent provenance for "any" thing" (Moreau et al. 2011)	"A statement of any changes in ownership and custody of the resource […]"
Subjects	Data provenance	Data provenance	Data or process provenance	Data provenance
Granularity/ level of detail	Mainly coarse (dataset level)	Mainly coarse (dataset level)	Fine or coarse	Coarse
Representation	Directed to provenance	Directed to provenance	Directed to provenance or usage or bidirectional	Directed to provenance

Table 4 Mapping of provenance elements of metadata and provenance standards

Provenance element	ISO 19115-2 Lineage subclasses	FGDC CSDGM	OPM	Qualified Dublin Core
Input	Source	Source information	Artifact	Source
Output	Source	Source information	Artifact	Source
Process/Model	Process step	Process step	Process	Provenance
Agents	Source, Process step	Source information, Process step	Agent	Provenance
Documentation of processes	Documentation	Source used citation abbreviation	Annotation	Provenance
Data producer	Processor	Process Contact	Agent	Creator, Contributor, Publisher

3 MetaViz: An Interactive Provenance Visualization Client

The need for an application visualizing data provenance arose during the GLUES Project, which is the coordination project of the international interdisciplinary research program 'Sustainable Land Management'[5] of the German Ministry of Education and Research. Within this funding measure several so called regional collaborative projects are researching the impacts of climate and socio-economic changes and a corresponding optimization of the use of land and natural resources

[5] http://modul-a.nachhaltiges-landmanagement.de

in different regions. Since this interdisciplinary research is policy-oriented the projects cooperate with regional scientists and stakeholders. The major aims of GLUES are to support communication, coordination, facilitation of data exchange and integration of results by developing a common data platform and consistent scenarios on land use, climate and social-economic change (Eppink et al. 2012; Mäs et al. 2011).

Technically, the access to the results of GLUES and the regional projects will be provided by means of a scientific GDI.[6] The provided data can be used by other scientists as input into their simulation models. To avoid misinterpretations and misapplication of data a major focus of the GDI implementation is on acquisition and representation of meaningful metadata and, in particular, provenance data. Due to the complexity and the high amount of metadata, a visual illustration of the interrelationships between different datasets and models is essential. Therewith, scientists can, for example, get a comprehensive view which models provide data for a certain scenario or whether an input data also served into other models. Beside the scientific work, such provenance visualization can also be of interest for research assessment and outreach analysis, since it represents the data exchange and collaboration between different research institutions.

There are several restrictions on provenance modelling and visualization based on the properties of the data and the user requirements within the project. For instance, detailed model descriptions (i.e. detailed descriptions of subprocesses) are not available in the metadata and would possibly not be feasible due to the models' complexity.

However, the importance of linking a model and its scientific publications has been pointed out on a GLUES workshop on models and consistent datasets. The following list characterizes models and data in the GLUES context:

- A model is represented as a single process step
- A model is described by a short summary, several scientific publications and a reference to the modeler or scientific institution
- A model can have several inputs and outputs, but is not directly connected to another model
- A dataset can be input of one model and simultaneously output of another model
- Pre-processing steps, such as cleaning the data, will only be collected as textual description of a process, but not as further process steps
- The provenance of a model is not considered.

The users in this research project are as wide-ranging as its thematic fields. However, the scientific community is just one of the four identified user groups, namely policy makers, stakeholder, society and scientist. The main objective of all user groups is the discovery of data to get a general overview of existing data or to find relevant data for a specific problem (Table 5). Data provenance information can support the assessment of the data quality and evaluation of the fitness for use.

[6] http://geoportal.glues.geo.tu-dresden.de

Table 5 User groups and their objectives within the research project

User groups	Objectives and activities in the project
Scientific Community	(Internal) documentation
	Discover relevant data
	Communicate (scientific) results
	Evaluate data
	Use results
Policy maker	Get overview
	Communicate
	Transfer (scientific) results
Stakeholder	Get information
	Implement results
	Communicate
	Transfer results
Society	Get general information

Descriptions of numerical models and their output data are complex and a quality evaluation requires detailed knowledge about the model and its initialisation, basic assumptions and research goals. For scientists having this background knowledge the provenance visualisation shows dependencies among data sets and it can indicate how model assumptions, restrictions and even errors propagate. Further, data provenance illustrates the data exchange and collaboration between the scientists.

The technical basis for this interdisciplinary work is a geoportal as an entry point to the GDI that provides a common metadata pool for the documentation of global long and midterm scenarios, its resulting datasets and synthesis results. The integrated metadata catalogue supports for the manual or scripted acquisition of ISO 19115 metadata including lineage information about source data and processes.

Figure 3 shows an extract of the catalogue user interface. It displays a part of the lineage information for a dataset that is generated by a model, named CAPRI. The user interface shows information about the model, such as a documentation link and brief summary. The model has at least two input sources, listed below the model information. Although this extract does not contain all lineage information, it shows that the table-like and non-interactive information representation is complex and difficult to apprehend for users.

3.1 Requirements for an Interactive Provenance User Interface

The interpretation of provenance information gets a significant support through the design of an easy-to-use and comprehensive user interface. Design dimensions are the provenance information model, information and interaction design as well as

Process information:	**Identifier:**	CAPRI	
	Software reference:		
	Procedure description:	The CAPRI model is a comparative static global partial equilibrium model for the agricultural sector. It endogenously determins market balances, area use and yields and many other variables for agricultural raw products and a number of processed products. CAPRI has been developed within EU Framework projects and is been used widely for policy impact analysis.	
	Documentation:	**Title:**	CAPRI model documentation
		Date:	2011-01-01
		Datatype:	publication
		Identifier:	
		Other citation details:	http://www.capri-model.org /docs/capri_documentation_2011.pdf
Source:	**Description:**	FAOSTAT data	
	Source citation:	**Title:**	FAOSTAT data
		Date:	2012-08-07
		Datatype:	publication
		Identifier:	faostatdatabasedomains (Codespace: urn:glues-ext:fao:metadata:dataset)
		Other citation details:	
	Description:	AGLINK-COSIMO (OECD,FAO)	
	Source citation:	**Title:**	AGLINK-COSIMO (OECD,FAO)
		Date:	2012-08-07
		Datatype:	publication
		Identifier:	oecdfaoagriculturaloutlook209-2018 (Codespace: urn:glues-

Fig. 3 Extract of the provenance information of a metadata record in the catalogue

the technical design. Although the successive data provenance (Table 1) information model based on ISO 19115 is quite static it is used here to support the integration in the existing GDI environment and connected to the Open Geospatial Consortium Web Catalogue Service (CSW) services and the existing metadata. The information design shall enable to answer the following questions, which summarize user requirements as defined by Kunde et al. (2008) and adapted to the introduced use case:

- Which data was used for the generation of a dataset?
- Which data was generated using a given dataset?
- Which actors (organizations, tools…) have been involved?
- Which resources from other models have been used in the generation of a dataset?
- In which stage of a processing chain is a given dataset?
- Did the model the dataset is part of reach a satisfactory conclusion by some given regulations or criteria?

Therefore the user interface should support an objective data representation of who, what, why, when and how the data is generated. The way of representing this information should be efficient using visualizations instead of long textual descriptions. The visualization should also display relationships and qualify the

user to position the dataset or model in space and time (Edwards et al. 2010; Bowers 2012; Zargar 2009).

The views should be adaptive and allow the typical scientific iterative data exploration (cp. Wang et al. 2008). In addition, all user groups require efficient and user-friendly navigations through lineage and usage information and dataset hierarchies as well as linked context-sensitive data representations such as the visualization of data in a map client. The information should be presented in a way understandable for users who are not familiar with the application and are going to use it rather seldom such as citizens or politicians. At the same time the presentation has to fit the purposes of scientists searching for detailed model or data descriptions and the corresponding publications. Conclusively, the interaction design should allow the navigation among related metadata, datasets or its visualization (Di and Yue 2011) and supports guidance through the data instead of complex querying.

Technically, the logic of the user interface has to offer possibilities to request and process ISO compliant metadata. This indicates that the application has to deal with the successive provenance steps of ISO 19115-2 deducing dataset inheritance and hierarchies based on scripted ID-matching. As shown in Table 4, a mapping from ISO to other specifications, especially FGDC, can be made easily to use the user interface with different data schemes.

Finally, a brokering mechanism that generates parameterized application links such as a link to a metadata's detail page of the catalogue has to be included.

3.2 Architecture and Implementation

The application MetaViz[7] is an interactive web-client consisting of a Java-based backend that contains the application logic and a user interface realized with HTML and JavaScript. Since MetaViz uses the standardized CSW 2.0.2 interface it does not need further storage systems, but directly requests the configured metadata catalogue for the ISO 19115 compliant metadata with provenance information (Table 6). The catalogue's response is processed and transformed into an intern and condensed JSON model with respect to the specific requirements of provenance visualization, such as sorting and pre-selection of required lineage and usage information.

Pre-processing of metadata is a computing expensive process. This is particularly because the ISO metadata schema does not directly fit the necessary requests that answer the user's requirements:

[7] A lineage example for the dataset PROMET shown in MetaViz application: http://geoportal.glues.geo.tu-dresden.de:8080/MetaViz/detail.jsp?id=glues:lmu:metadata:dataset:promet

Information about MetaViz is summarized in a factsheet: http://geoportal.glues.geo.tu-dresden.de/geoportal/documents/Fact_Sheet_MetaViz.pdf

Table 6 Characteristics of MetaViz

Property	Realization in MetaViz
Subject/entities	Data
Granularity/Level of detail	Coarse
Representation	Bidirectional
	Successive provenance steps
Visual representation	Graph
Querying	Visual query interface
Application domain	Web
Storage	Tightly coupled with metadata
Interfaces and Data Formats	Standard interface using ISO 19115-2, CSW 2.0.2
Infrastructure	Distributed environment (service-based)

- Lineage and usage information
- Parent–child-relations between datasets and data series
- Connected view services (or other services supporting further exploration)

ISO metadata stores lineage as sequential provenance steps instead of a complete provenance graph in one metadata entry. Due to this, several catalogue requests have to be made to compose the whole lineage of a dataset. Furthermore, usage information has to be deduced from the lineage descriptions, as the 'usage of a dataset' is only stored as a lineage of another dataset. Usage is considered here only in terms of processes that lead to new data products, leaving out direct applications like data visualization).

Parent–child-relations are also not stored bidirectional: metadata sets contain links to parents, but not to children (cp. Nogueras-Iso et al. 2005, p. 37). Links of datasets and their connected view services are likewise stored within the metadata of the service instead of the metadata for the dataset. Thus all data offered by the CSW has to be analyzed to get the children, the usage or the linked view services of a dataset. This pre-processing is quite computing expensive. To increase the query performance MetaViz can be switched between a direct database mode or a CSW mode. In the database mode the metadata is requested directly from the underlying database of a metadata catalogue, resulting in much better response times. Using the standardized CSW interface in the CSW mode lacks in performance but allows more flexibility in being less tightly coupled to the database scheme of the used catalogue.

As illustrated in Fig. 4, MetaViz can be linked with other clients being used in a GDI like the geocatalogue GUI or geovisualization clients. This allows for a continuous user interaction. MetaViz is not only requesting data from a metadata catalogue but also linking back to a catalogue's detail page, which shows the entire list of metadata elements in the traditional manner. Furthermore the application is coupled with a map client to visualize the data if the metadata contains a reference to a Web Map Service (WMS). By calling MetaViz parameterized with a dataset id, it can be embedded into other websites or applications.

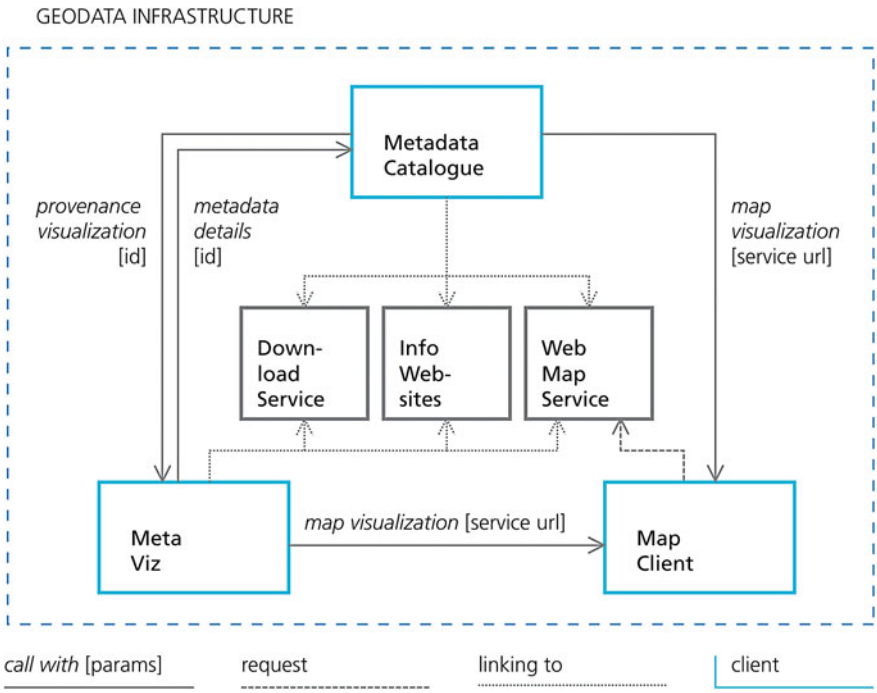


Fig. 4 Linkage and information exchange between MetaViz and other GDI clients

3.3 User Interface

MetaViz focuses on a user-friendly and compact visual description of lineage and usage of datasets within a GDI. The main element of the application is a tree-like interactive lineage and usage graph (Fig. 5)[8] showing the provenance of a dataset with its name displayed above the graph on the left. Next to the graph some general information such as name, temporal and spatial extent, tagged keywords and (interactive) relations to parent or child datasets are listed.

Below the graph extended provenance information is displayed. Process descriptions and publications are separated visually to arrange information in a well-structured and easy-readable way. The process description contains free-text about rationale of the process step and process parameters such as software reference, processor and time of process execution. Pre studies with scientists in the GLUES project lead to a design which does not display all elements of the ISO 19115-2, such as detailed runtime parameters, to keep a rather simple and quickly to grasp user interface.

[8] The example shown in the screenshot is available in the web: http://geoportal.glues.geo.tu-dresden.de:8080/MetaViz/detail.jsp?id=f872b5b8-bb23-4df5-a906-0b396c99cc22

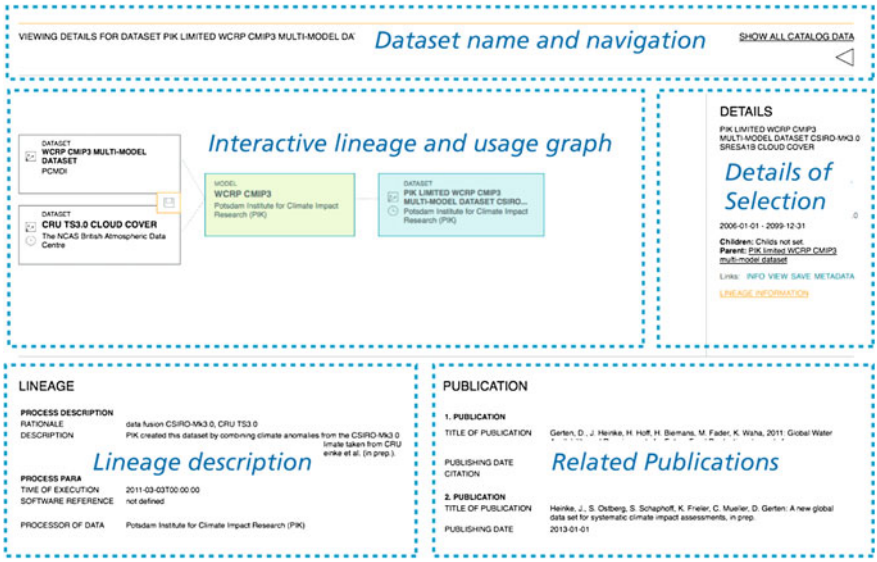


Fig. 5 Overview of areas in MetaViz application GUI showing the lineage of a dataset

The publication information displayed on the right lists a BibTex-like reference, the date of publication as well as a link to the publication, if available.

The lineage graph shows the dataset derivation. Explored from left to right, it shows lineage information on the left and usage information, if available, on the right. It connects the focused dataset (blue box) to the process (green box) where it originates from as well as the source datasets (white boxes) of these processes.

Each graph contains a maximum of one lineage and one usage step to keep the presentations comprehensible. It is possible to focus a presentation either on lineage (Fig. 6)[9] or on usage (Fig. 7).

The application does not only visualize the lineage as a graph but also displays relevant information to assess data quality. This information like data provider, data type or time-variant are displayed in the visualization as texts or symbolized via icons (Fig. 8)[10]. All icons are explained with short tooltips texts to facilitate the application handling.

The general navigation concept of the application is as simple as the visualization. The users do not have to formulate complex queries. Navigating through the lineage graph or to the context-sensitively linked applications is done by clicking on an icon, link or button.

[9] The example shown in the screenshot is available in the web: http://geoportal.glues.geo.tu-dresden.de:8080/MetaViz/detail.jsp?id=glues:lmu:metadata:dataset:promet

[10] The example shown in the screenshot is available in the web: http://geoportal.glues.geo.tu-dresden.de:8080/MetaViz/detail.jsp?id=glues:pik:metadata:dataset:csiro-mk3.0sresa1bcloudcover

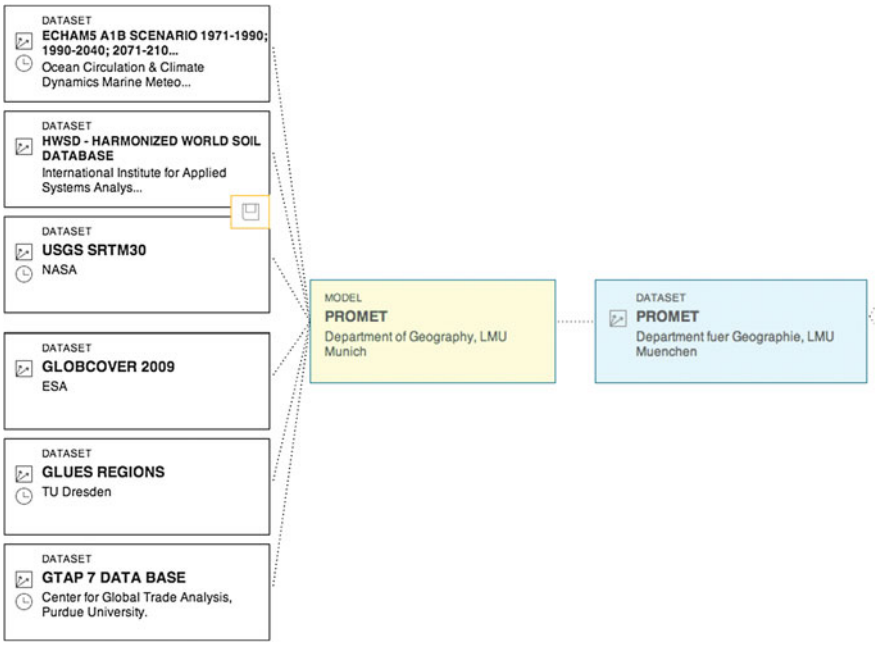


Fig. 6 Lineage graph of the dataset PROMET visualized in MetaViz

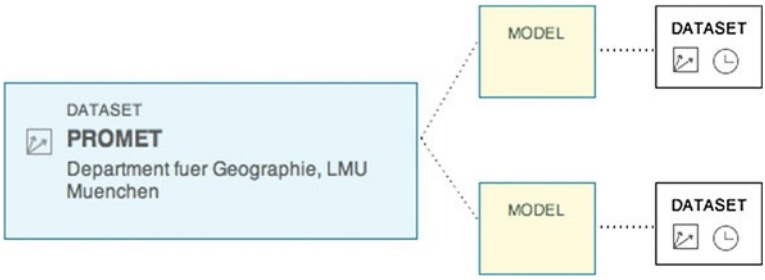


Fig. 7 Simplified usage graph of the dataset PROMET

MetaViz can be used as standalone application or integrated in other website with a parameterized call as well as called from the catalogue GUI as one view of the standardized metadata. This offers the different user groups, such as the data modellers, the possibility to integrate the application in their own research website, link to it from scientific publications or use a screenshot of the graphical presentation in their publication.

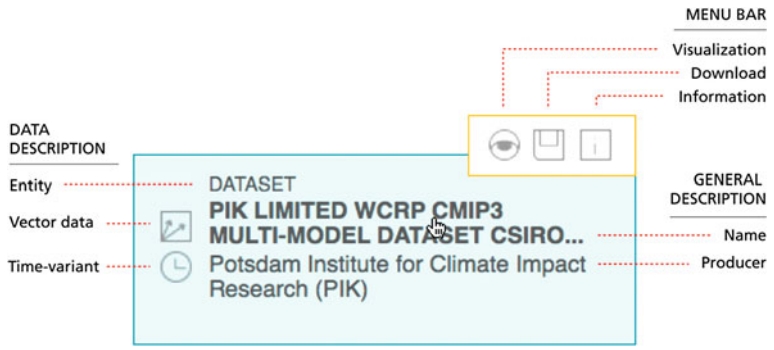


Fig. 8 Visual representation of dataset in MetaViz and context-sensitive menu

4 Conclusions and Future Work

Management and representation of provenance information in GDI has not gained much attention so far. The introduced provenance visualization client illustrates how the presentation of metadata in GDI can be enriched by interactive, intuitive and user-friendly interfaces. Such visualization supports the communication of data quality, enhances interpretation and prevents misinterpretation or misapplication of geodata. Moreover, it is felt that intuitive and convincing metadata applications—as intended with MetaViz—can further stimulate the willingness to generate and maintain metadata. The representation of the processing steps is understandable, even for non-expert users. In scientific GDI usage information can play a major role for the evaluation of scientific outputs, comparable to the way citations are used to rank scientific publications.

The current metadata standards, including ISO 19115, do not entirely fit the requirements of our use case. The description of numerical models and their output data would require data elements explicitly representing information about model initialisation, scenarios, drivers and basic assumptions of the model. In particular scenarios, that define a projection of a potential future based on a coherent set of assumptions (Nakićenović et al. 2000), would be useful to classify and compare datasets. Statistical analyses, like in spatial econometrics, require provenance information for analysis workflows and the applied (spatial) weights (Anselin and Rey 2012). At least for scientists working with these data such information is indispensable for evaluation.

Another issue are parent–child-relations among datasets, which are typical for the machine generated data in our use case. Although very useful for structuring data, these relations are hardly represented and navigable in current GDI catalogue user interfaces. In particular for datasets with a high number of child datasets it is not sufficient to store the relation without an explicit description of the concrete commonalities and differences of the sub datasets (Nogueras-Iso et al. 2005). The navigation and illustration of these relations is a possible future extension of MetaViz.

Scientists, but also data producers in general, are clearly not passionate in collecting detailed metadata descriptions. Therefore, the application MetaViz requires only a minimal set of lineage attributes, being evaluated and approved by domain modelling experts within the GLUES project, and integrates existing descriptions, such as publications. Nevertheless, automatic metadata derivation and acquisition remains a big issue for future research.

So far, MetaViz considers usage only in terms of processes that lead to new data products. To also include direct applications of the data (e.g. visualizations or analysis) future extensions for direct user feedback and a rating system are planned. Additionally, a usability evaluation of the current GUI shall help in improving the design. Further, the suitability of the application for other use cases, such as metadata created by web processing services will be analyzed.

References

Anand MK, Bowers S, Ludäscher B (2010) Provenance browser: displaying and querying scientific workflow provenance graphs, ICDE, 2010

Anselin L, Rey SJ (2012) Spatial econometrics in an age of CyberGIScience. Int J Geogr Inf Sci 26(12):2211–2226

Bose R, Frew J (2005) Lineage retrieval for scientific data processing: a survey. ACM Comput Surv 37(1):1–28

Bowers S (2012) Scientific workflow, provenance, and data modeling challenges and approaches. J Data Semant 1(1):19–30

Cass AG, Lerner E, McCall K, Osterweil LJ, Sutton SM, Wise E (2000) Little-JIL/Juliette: a process definition language and interpreter. International Conference on Software Engineering, 2000

Cheung K, Hunter J (2006) Provenance explorer—customized provenance views using semantic inferencing. ISWC 2006, (LNCS), vol 4273. pp 215–227

Devillers R, Bédard Y, Jeansoulin R (2005) Multidimensional management of geospatial data quality information for its dynamic use within GIS. Photogram Eng Remote Sens 71(2):205–215

Di L, Yue P (2011) Provenance in earth science cyberinfrastructure. A White Paper for NSF EarthCube, 2011

Edwards P, Pignotti E, Reid R (2010) Weaving a provenance fabric to support next generation science. IEEE internet computing special issue on provenance in web applications, 2010

Eppink F, Werntze A, Mäs S, Popp A, Seppelt R (2012) Land management and ecosystem services: how collaborative research programmes can support better policies. GAIA: Ecol Perspect Sci Soc 21(1):55–63

FGDC (2000) Content standard for digital geospatial metadata workbook (For use with FGDC-STD-001-1998), Version 2.0. Federal geographic data committee, May 1 2000

Fisher P, Comber AJ, Wadsworth R (2009) What's in a name? Semantics, standards and data quality. In: Devillers R, Goodchild H (eds) Spatial data quality: form process to decisions. CRC Press, Boca Raton, pp 3–16

Glavic B, Dittrich K (2007) Data provenance: a categorization of existing approaches, BTW'07, pp 227–241

ISO 19115-2 (2005) International standard on geographic information—Part 2: metadata for imagery and gridded data

Kindermann S, Stockhause M, Ronneberger K (2007) Intelligent data networking for the earth system science community, German e-Science

Kunde M, Bergmeyer H, Schrieber A (2008) Requirements for a provenance visualization component, Provenance and annotation of data and processes (Lecture notes in computer science), vol 5272. pp 241–252

Lanter DP (1991) User-centered graphical user interface design for GIS. National center for geographic information and analysis, report. pp 91–96

Malaverri JEG, Medeiros CB, Camargo R (2012) A provenance approach to assess quality of geospatial data. 27th symposium on applied computing (SAC)

Mäs S, Müller M, Henzen C, Bernard L (2011) Linking the outcomes of scientific research: requirements from the perspective of geosciences. Proceedings of the first international workshop on linked science 2011 (LISC2011), CEUR Workshop Proceedings, vol 783

Moreau L (2010) The foundations for provenance on the web. Found Trends Web Sci 2(2–3):99–241

Moreau L, Clifford B, Freire J, Gil Y, Groth P, Futrelle J, Kwasnikowska N, Miles S, Missier P, Myers J, Simmhan Y, Stephan E, Van den Bussche J (2011) The open provenance model—core specification (v1. 1). Future generation computer systems

Nakićenović N, Alcamo J, Davis G (2000) IPCC special report on emissions scenarios (SRES) Cambridge. Cambridge University Press, NY

Nogueras-Iso J, Zaragaza-Soria FJ, Muro-Medrano PR (2005) Geographic information metadata for spatial data infrastructures: resources, interoperability and information retrieval. Springer, Berlin 2005

OGC (2007) OpenGIS catalogue services specification. Version 2.0.2, OGC 07-006r1

Osterweil L, Clarke L, Ellison A, Boose E, Podorozhny R, Wise A (2010) Clear and precise specification of ecological data management processes and dataset provenance. IEEE Trans Autom Sci Eng 7(1):189–195

Pastorello G, Medeiros C, Resende S, Rocha H (2005) Interoperability for GIS document management in environmental planning. J Data Semanti III (LNCS), vol 3534. pp 100–124

Simmhan Y, Plale B, Gannon D (2005) A survey of data provenance in e-science. SIGMOD record 34:31–36

Spéry L, Claramunt C, Libourel T (2001) A spatio-temporal model for the manipulation of lineage metadata. Geoinformatica 5(1):51–70

Tilmes C, Fleig A (2008) Provenance tracking in an earth science data processing system. Lect Notes Comput Sci 5272:221–228

Vert G, Stock M, Jankowski P, Gessler P (2002) An architecture for the management of GIS data files. Trans GIS 6(3):259–275

Wang S, Padmanabhan A, Myers J, Tang W, Liu Y (2008) Towards provenance-aware geographic information systems, GIS. ACM, NY

Yue P, Wei Y, Di L, He L, Gong J, Zhang L (2011) Sharing geospatial provenance in a service-oriented environment. Comput Environ Urban Syst 35:333–343

Zargar A (2009) An operation-based approach to the communication of spatial data quality in GIS

Part IV
Formal Semantics

Matching Formal and Informal Geospatial Ontologies

Heshan Du, Natasha Alechina, Mike Jackson and Glen Hart

Abstract The rapid development of crowd-sourcing or volunteered geographic information both challenges and provides opportunities to authoritative geospatial information. Matching geospatial ontologies is an essential element to realizing the synergistic use of disparate geospatial information. We propose a new semi-automatic method to match formal and informal real life geospatial ontologies, at both terminology level and instance level, ensuring that overall information is logically coherent and consistent. Disparate geospatial ontologies are matched by finding a consistent and coherent set of mapping axioms with respect to them. Disjointness axioms are generated in order to facilitate detection of errors. In contrast to other existing methods, disjointness axioms are seen as assumptions, which can be retracted during the overall process. We produce candidates for retraction automatically, but the ultimate decision is taken by domain experts. Geometry matching, lexical matching and cardinality checking are combined when matching geospatial individuals (spatial features).

1 Introduction

In recent years, the emergence and development of crowd-sourcing or volunteered geographic information has challenged and also provided opportunities to the traditional model of geospatial data collection, storage and updates. Allowing amateurs to collect geospatial data helps lower the cost, capture richer user-based information and reflect real world changes more quickly. At the same time it may

H. Du (✉) · N. Alechina · M. Jackson
The University of Nottingham, Nottingham, UK
e-mail: psxhd1@nottingham.ac.uk

G. Hart
Ordnance Survey of Great Britain, Southampton, UK

D. Vandenbroucke et al. (eds.), *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography,
DOI: 10.1007/978-3-319-00615-4_9,

also dilute information quality, such as completeness, consistency and accuracy (Jackson et al. 2010). It is desirable to use volunteered and authoritative geospatial information as complements to each other, taking the best of both.

Ontology refers to an explicit specification of a shared conceptualization (Gruber 1993) and plays an important role in establishing shared formal vocabularies. A spatial individual has a certain and verifiable location and a meaningful label, which together distinguish itself from others. Geospatial ontologies describe conceptual hierarchies and interrelations of terminologies in the domain of geospatial science, which are used to describe facts (classifications, relations, attributions and locations) about spatial individuals. Compared to other ontologies, geospatial ontologies have some special properties. Firstly, many geospatial terminologies are commonly used in daily life and their meanings vary in different contexts. For example, "College" may refer to an institution within a university in one ontology, whilst meaning a secondary school in another. In addition, geospatial ontologies often do not have a huge number of classes as ontologies in several other subject areas (for example, biomedicine) do, but may represent many real world spatial individuals, whose locations, at least in theory, can be verified. For example, *Space*, a large-scale geospatial ontology constructed using WordNet, GeoNames and Thesaurus of Geographical Names, contains 845 classes and 6, 907, 417 individuals (Giunchiglia et al. 2012). Since geospatial ontologies for authoritative and volunteered data sets are developed independently, matching geospatial ontologies is an essential step to use them synergistically.

We propose a new semi-automatic method for matching geospatial ontologies, at both terminology level and instance level. Geographic information quality includes several aspects, viz., completeness, logical consistency, positional accuracy, thematic accuracy, temporal quality and usability (International Organization for Standardization 2011). We focus on logical consistency, ensuring that, after adding a mapping, overall information is logically coherent and consistent, without any contradictions. We assume that the TBoxes[1] of geospatial ontologies are not very large, but contain concepts which are more ambiguous, compared to, for example, biomedical ontologies. The matching process is reduced to the problem of finding a coherent and consistent set of assumptions (including disjointness axioms, equivalence and inclusion relations between concepts from different ontologies, and "sameAs" and "partOf" relations between spatial instances) with respect to input ontologies. Unlike a premise, an assumption is believed by default, but can be retracted later if found to be not reasonable later. Disjointness axioms are generated in order to facilitate detection of errors or contradictions. Geospatial individuals are matched using location, lexical and classification information.

The rest of the chapter is organized as follows. Section 2 reviews related work on ontology matching and geospatial data integration. We describe our new method in Sect. 3, and evaluate it in Sect. 4. Finally, Sect. 5 provides conclusions.

[1] Definitions of concepts and roles.

2 Related Research

Ontology matching is the task of finding a mapping, i.e. a set of correspondences, between entities from different ontologies (Euzenat and Shvaiko 2007). It includes two main levels, the terminology level and instance level. Many ontology matching methods and systems have been developed in recent years (Euzenat and Shvaiko 2007; Shvaiko and Euzenat 2012). Most of them are based on lexical and structural analysis and similarity measurements. However, mappings generated by these methods often contain logical contradictions. Logical reasoning is employed for either mapping generation or verification in some systems, including the early logic-based attempts, such as CtxMatch (Bouquet et al. 2003) and its extension S-Match (Giunchiglia et al. 2004), and more recently, ASMOV (Jean-Mary et al. 2010) which verifies mappings against five specified inconsistent patterns, KOSIMap (Reul and Pan 2010) based on description logic coherence checking assuming the disjointness of siblings, ContentMap (Jimenez-Ruiz et al. 2009) which computes new entailments from initial mappings generated by other systems, LogMap (Jimenez-Ruiz and Grau 2011) and CODI (Niepert et al. 2010).

LogMap is a logic-based ontology matching tool, designed for large-scale biomedical ontologies. It employs lexical and structural methods to compute an initial mapping. LogMap iterates two main steps. In Step 1, unsatisfiable classes will be detected using propositional Horn representation and satisfiability checking, and be repaired using a greedy diagnosis algorithm. However, the propositional Horn satisfiability checking is sound but incomplete, and the underlying semantics is restricted to propositional logic, and thus cannot guarantee the coherence of the mapping between more expressive ontologies. In Step 2, new mapping relations will be generated based on the similarity of classes related to established correspondences. Only newly discovered correspondences can be eliminated in the repair step, whilst correspondences found in earlier iterations are seen as established or valid. In other words, each mapping relation will be checked once, against the available information at that time, which, however, cannot guarantee its correctness when new information is discovered later.

CODI is a probabilistic logical alignment system based on Markov logic (Domingos et al. 2008). It transforms the matching problem to a maximum-a posterior optimization problem subject to cardinality constraints, coherence constraints and stability constraints. The GUROBI optimizer (Gurobi Optimization Inc 2012) is employed to solve the optimization problems. CODI reduces incoherence during the alignment process for the first time, compared to all other existing methods repairing alignments afterwards. CODI is based on the rationale of finding the most likely mapping by maximizing the sum of similarity-weighted probabilities for potential correspondences. However, during the optimization process, some valid correspondences can be thrown away.

It is a central problem within the context of Linked Data to identify correspondences between instances from different sources. Wolger et al. (2011) provide a summary of the existing data interlinking methods. Most of them are based on

lexical methods, such as string matching and word relation matching, machine learning and natural language processing techniques. Only within some systems, such as L2R (Sais et al. 2007), KnoFuss (Nikolov et al. 2007) and RDF-AI (Scharffe et al. 2009), consistency is checked.

In addition, there is some recent work on debugging and repairing ontology mappings (Meilicke and Stuckenschmidt 2009; Qi et al. 2009; Wang and Xu 2008), which is still at an early stage. However, to the best of our knowledge, all of them, as well as the ontology matching or data interlinking systems described above, treat disjointness axioms as premises, rather than retractable assumptions, and none of them have addressed the special properties of geospatial information.

In geospatial information science, several data conflation methods have been developed for matching or integrating geospatial vector data, mainly based on the similarities of geometries or topological relations, as well as attributes, if available. Most of them focus on conflating road vector data. However, few of these techniques check and ensure the logical coherence and consistency of integrated information (Du et al. 2012). In addition, several ontology-driven methods have been developed for integrating geospatial terminologies. Most of them are based on similarity measures or a predefined top-level ontology. Logical reasoning is only employed when formal ontologies commit to a same top-level ontology (Buccella et al. 2009). However, when ontologies are developed independently, the common top-level ontology is not usually available. Additionally, there exist some other methods (Volz and Walter 2004; Jain et al. 2010), following the bottom-up approach to linking geospatial schemas or ontologies, inferring terminology correspondences from instances correspondences. Though this works well when instance data is representative and overlapping, it uses a very strong form of induction from particular to the universal, which leads to lack of correctness and completeness (Bouquet 2007). Therefore, more research is required to fill in the gap, exploring logic-based approaches to matching geospatial ontologies.

3 Method

We propose a new semi-automated method for matching geospatial ontologies. Initial mappings between concepts and between individuals are generated using lexical matching and geometry matching. Logical coherence and consistency is ensured by automatically generating sets of assumptions responsible for incoherence or inconsistency using description logic reasoner Pellet (Sirin et al. 2007), and asking domain experts to decide which assumptions from these sets should be removed to restore coherence and consistency. Due to limited space, we recommend Baader et al. (2007) for the basic notions of description logic.

Definition 1 (*Ontology*) An ontology O has a TBox which contains knowledge at the conceptual level, and an ABox which describes facts about individuals using terminologies described in the TBox.

Definition 2 (*Coherence*) An ontology O is coherent if there is no class which only admits an empty interpretation. Otherwise, it is incoherent.

Definition 3 (*Consistency*) An ontology O is consistent if

- there exists no individual name a can be shown to belong to a concept C and to its negation, C;
- there exists no individual names a, b can be shown to belong to a role R and its negation, R;

Otherwise, O is inconsistent.

This method matches ontologies from the terminology level to the instance level. It includes four main steps. Since the original ontologies are often lightweight, disjointness axioms are generated in *Step 1*, to facilitate detection of incoherencies and inconsistencies. Ontology TBoxes are matched in *Step 2*. In *Step 3*, we match ABoxes of geospatial ontologies using location and lexical information. The whole ontologies are matched in *Step 4*. Mapping relations are represented as axioms in standard description logics, making use of existing and highly optimized reasoning techniques, for example Pellet (Sirin et al. 2007). Differing from other existing methods, this method treats generated disjointness axioms and the mappings between different ontologies, as *assumptions*, rather than premises. Users are allowed to retract or enable existing assumptions, and add new assumptions, during the matching process.

Definition 4 (*Premise and Assumption*) A premise is believed all the time, whilst an assumption is believed by default, but may be retracted later.

To represent and reason with two ontologies O^i and O^j, where i, j are their names, as well as the mapping M between them, as if they all belong to one super ontology ($O^i \cup O^j \cup M$), we label all atomic concepts, roles and individual names in each ontology by the name of the ontology. An atomic concept C and an individual name a from ontology i are represented as i: C and i: a respectively.

Definition 5 (*Union of Ontologies*) The union of ontologies O^i and O^j, represented as ($O^i \cup O^j$), is an ontology containing all axioms in O^i and O^j.

3.1 Matching Terminologies

A terminology mapping is a set of correspondences between concepts from different ontologies. A terminology correspondence is represented in one of the two basic forms:

$$B^i \sqsubseteq C^j \tag{1}$$

$$B^i \sqsupseteq C^j \tag{2}$$

where B, C denote concepts.[2] The relation (1) states that the concept B from the ontology i is more specific than or equivalent to the concept C from the ontology j. The relation (2) states that the concept B from the ontology i is more general than or equivalent to the concept C from the ontology j. The equivalence relation (3) holds if and only if (1) and (2) both hold.

$$B^i \equiv C^j \tag{3}$$

It states that the concept B from the ontology i and the concept C from the ontology j are equivalent.

A disjointness axiom states that two or more concepts are pairwise disjoint, having no common element. For example, *Person* and *Place* are disjoint, which can be represented as $Person \sqsubseteq \neg\ Place$, where $\neg$ denotes negation. Disjointness axioms in ontologies play an important role in debugging ontology mappings. However, within original geospatial ontologies, disjointness axioms are not always available or sufficient. Adding disjointness axioms manually, especially for large ontologies, is time-consuming and error-prone. Many existing systems employ more automatic approaches, either assuming the disjointness of siblings (e.g. Reul and Pan 2010), or employing machine learning techniques to detect disjointness (e.g. Meilicke et al. 2008a). After disjointness axioms are generated by whatever means, all existing ontology matching or debugging methods, to the best of our knowledge, use them as premises, though the input disjointness axioms can be insufficient or too restrictive. Differing from these methods, we use generated disjointness axioms as assumptions, and ensure the assumption set is coherent.

Definition 6 (*Coherence of an Assumption Set*) An assumption set A_s is incoherent with respect to an ontology O, if $O \cup A_s$ is incoherent, but O is coherent. Otherwise, it is coherent with respect to an ontology O.

When some incoherence is introduced by assumptions, minimal incoherent assumption sets (MIA) will be computed. The notion of MIA is defined by extending the minimal conflict set defined for mappings (Meilicke et al. 2008b) to this context.

Definition 7 (*Minimal Incoherent Assumption Set*) Given a set of assumptions A_s, a set $C \subseteq A_s$ is a minimal incoherent assumption set (MIA) iff C is incoherent and each $C' \subset C$ is coherent.

A minimal incoherent assumption set can be fixed by removing any axiom from it. When a MIA contains more than one element, one needs to decide which axiom to remove. Most of the existing methods remove the one either with the lowest confidence value or which is the least relevant. However, there is no consensus with respect to the measure of the degree of confidence or relevance. In several cases, confidence values or relevance degrees might be unavailable or difficult to

[2] When B, C denote atomic concepts, $B^i = i{:}\ B$, $C^j = j{:}\ C$.

compute or compare. Rather than relying on them, we allow domain experts to make ultimate decisions.

Algorithm 1 is designed to generate a coherent assumption set (CAS) with respect to an ontology.[3] The set of minimal incoherent assumption sets will be visualized clearly (Line 5). Domain experts are employed to take repair actions (Line 6). Currently, a repair action can be retracting an assumption axiom. Users are allowed to take several repair actions at one time.

ALGORITHM 1: CAS
Input: O: a coherent ontology
A_s: an assumption set for O
Output: A_{cs} : a coherent assumption set with respect to O. $A_{cs} \subseteq A_s$.
1. $A_{cs} \coloneqq A_s$;
2. $O_{tmp} := O \cup A_{cs}$;
3. **while** O_{tmp} is incoherent **do**
4. $S_{mia} := MIA(O_{tmp})$;
5. *visualization*(S_{mia});
6. * *repair*(O_{tmp}, S_{mia});
7. *update*(A_{cs});
8. **end while**
9. **return** A_{cs}

Step 1: Generating coherent disjointness assumption sets (CDAS). For each coherent ontology TBox,[4] as shown in *Algorithm 2*, we generate disjointness axioms as assumptions for sibling classes and refine them by applying *Algorithm 1*.

ALGORITHM 2: CDAS
Input: T: a *coherent* ontology TBox
Output: D_{cs}: a coherent disjointness assumption set with respect to T.
1. $D_s \coloneqq disjointnessOfSiblings(T)$;
2. $D_{cs} \coloneqq CAS(T, D_s)$;
3. **return** D_{cs}

Step 2: Matching terminologies (*Algorithm 3*). Currently, an initial terminology mapping is generated by using a very simple lexical matching method, i.e. stating equivalence of atomic concepts with identical names (Line 1). *Definition 6* can be extended from one ontology O to two ontologies T_1 and T_2, given that the union of two ontologies, $T_1 \cup T_2$, is an ontology. A coherent disjointness assumption set for TBoxes (union of CDAS for each TBox) and an initial terminology mapping form an initial assumption set, from which, a coherent assumption set with respect to T_1

[3] In an algorithm, lines marked with * may require manual intervention.

[4] An ontology only with a TBox.

and T_2 is calculated by applying *Algorithm 1*. An assumption in a minimal incoherent assumption set can be a disjointness axiom or a terminology correspondence axiom. Domain experts are consulted to decide which assumption(s) to retract when incoherence arises.

ALGORITHM 3: Matching Terminologies
Input T_1, T_2: *coherent* ontology TBoxes
D_{cs}: a coherent disjointness assumption set with respect to T_1, T_2
Output T_{cs}: a coherent terminology assumption set with respect to T_1, T_2

1. $M_{st} := lexicalMatching(T_1, T_2)$;
2. $T_{cs} := CAS(T_1 \cup T_2, D_{cs} \cup M_{st})$;
3. **return** T_{cs}

3.2 Matching Geospatial Individuals

An instance level mapping is a set of individual correspondences. An individual correspondence is represented in one of the following forms:

$$(i : a, j : b) \in sameAs \tag{4}$$

$$(i : a, j : b) \in partOf \tag{5}$$

where a, b denote individual names. The relation (4) states that the individual name a from the ontology i and the individual name b from the ontology j refer to the same object. The relation (5) states that the individual name a from the ontology i refers to an object which is a part of the object the individual name b from the ontology j refers to.

ALGORITHM 4: Matching Geospatial Individuals
Input A_1, A_2: ontology ABoxes
Output M_{sa} : an initial geospatial instance mapping between A_1, A_2

1. M_{sa}:={};
2. **for each** spatial individual a_1 in A_1 **do**
3. **for each** spatial individual a_2 in A_2 **do**
4. **if** *geo_poss_match*(a_1.*geometry*, a_2.*geometry*)
5. **and** *lex_poss_match*(a_1.*lexicons*, a_2.*lexicons*) **then**
6. add (a_1, a_2) ∈ *sameAs* to M_{sa};
7. **end if**
8. **end for**
9. **end for**
10. *cardinalityChecking* (*Msa*)
11. **return** *Msa*

Step 3: Matching Geospatial Individuals. *Algorithm 4* is designed to match geospatial individuals whose geometries are represented using the same coordinate reference system (CRS). The geometry of a spatial object can be represented in different accuracy levels, granularities or world views in different ontologies. In other words, for the same spatial object, the recorded geometry in ontology *i* may not be exactly the same as the recorded geometry in ontology *j*.

The *geo_poss_match* (Line 4) between two geometries returns true if the geometries are similar enough given a margin of error in representation. Currently, it requires input geometries as polygons. Two polygons are possibly matched if one of them is the smallest polygon containing the characteristic point from the other.[5]

The *lex_poss_match* (Line 5) between two lexical descriptions returns true if the lexicons (meaningful labels indicating identity) are similar enough, tolerating partial differences, for example, a full name and its abbreviation, and recognizing different names for the same location. Currently, it employs a series of basic string matching strategies, such as equivalence, inclusion and abbreviation.

For each pair of spatial individuals a_1 and a_2 from different ontologies, if their geometries are possibly matched (Line 4) and their lexicons are possibly matched (Line 5), then they can be assumed to be the same. We generate a "sameAs" relation linking them and add it to an instance mapping M_{sa} (Line 6).

It is currently assumed that, within a local ontology, a spatial individual has at most one representation. In other words, there are no "sameAs" relations within a local ontology. The cardinality checking (Line 10) revises M_{sa}, a set of "sameAs" relations, ensuring that "sameAs" is one-to-one. If not, we remove them from M_{sa}, and add corresponding "partOf" relations. For example, if M_{sa} contains (i: a, j: b) ∈ *sameAs* and (i: c, j: b) ∈ *sameAs*, we replace them with (i: a, j: b) ∈ *partOf* and (i: c, j: b) ∈ *partOf*. [6]

The geometry matching, lexical matching and cardinality checking complement each other to cope with the following possibilities. Different geospatial individuals may share the same label or the same location in an ontology. In addition, a same geospatial individual may be represented as a whole in one ontology, whilst as several parts of it in another.

---

[5] Individuals from the Ordnance Survey of Great Britain (OSGB) Buildings and Places ontology and the OpenStreetMap ontology (See Sect. 4.2) are spatially linked by finding the smallest OSM polygon containing a point from OSGB address Layer 2. See Fig. 1 for examples. Polygons containing the same red point are linked.

A more sophisticated geometry matching method for generating spatial "sameAs" and "partOf" relations is under development and evaluation.

[6] We are aware that, an individual a in one ontology O^i can be part of an individual b in another ontology O^j, even if there are no other individuals in O^i who can be part of b. This has be considered when designing our new geometry matching method.

Definition 8 (*Consistency of an Assumption Set*) An assumption set A_s is inconsistent with respect to an ontology O, if $O \cup A_s$ is inconsistent, but O is consistent. Otherwise, it is consistent with respect to an ontology O.

Definition 9 (*Minimal Inconsistent Assumption Set*) Given a set of assumption A_s, a set $C \subseteq A_s$ is a minimal inconsistent assumption set (MIA),[7] iff C is inconsistent and each $C' \subset C$ is consistent.

Similarly, a minimal inconsistent assumption set can be fixed by removing one element from it. The algorithm for calculating consistent assumption set (CAS) can be generated from *Algorithm 1*, changing coherence checking to consistency checking. *Definition 8* can be easily extended to deal with two ontologies.

ALGORITHM 5: Matching Geospatial Ontologies

Input $O_1 = (T_1, A_1)$, $O_2 = (T_2, A_2)$: coherent and consistent geospatial ontologies
T_{cs}: a coherent assumption set with respect to T_1, T_2
M_{sa}: an initial geospatial instance mapping between A_1 and A_2.

Output O_{cs}: a consistent assumption set with respect to O_1 and O_2

1. $O_{cs} := CAS(O_1 \cup O_2, T_{cs} \cup M_{sa})$[8];
2. **return** O_{cs}

Step 4: Matching Geospatial Ontologies (*Algorithm 5*). A coherent assumption set with respect to TBoxes is obtained in *Step 2*. An initial geospatial instance mapping between ABoxes is generated in *Step 3*. The union of them is an assumption set with respect to input ontologies. Applying *CAS*, if overall information is inconsistent, a set of minimal inconsistent assumption sets will be calculated, and visualized appropriately to help domain experts to repair them. These steps iterate until a consistent assumption set is found.

4 Evaluation

The method described above is implemented as a system called GeoMap. Pellet (Sirin et al. 2007) is employed for coherence and consistency checking. Minimal incoherent assumption sets are calculated from explanations for unsatisfiable classes, and minimal inconsistent assumption sets from explanations for inconsistencies.

We evaluate GeoMap using the Ordnance Survey of Great Britain (OSGB) Buildings and Places ontology (Hart et al. 2008) and the OpenStreetMap (OSM) controlled vocabularies (OpenStreetMap 2012). OSGB and OSM are

[7] MIA refers to *minimal incoherent assumption set* when matching terminologies, and refers to *minimal inconsistent assumption set* when matching instances.

[8] *CAS* here refers to the calculation of consistent assumption set.

representatives of authoritative and crowd-sourced geospatial information sources respectively. OSGB is the national topographic mapping agency of Great Britain. OSM is a collaborative project to create a free editable map of the world, relying on volunteers for data collection. Currently, OSM has not established a standard ontology, but maintains a collection of commonly used tags for main map features. An OSM feature ontology is generated automatically from the existing classification of main features. For example, given "Restaurant" is a value under the key "Amenity" in the OSM classification, we formulate this as *OSM*: *Restaurant* $\sqsubseteq$ *OSM*: *Amenity*. Both ontologies are written in the OWL 2 Web Ontology Language (W3C 2009). The OSGB Buildings and Places ontology has 692 classes and 1,230 logical axioms. There are 663 classes and 677 logical axioms in the OSM ontology. Both ontologies, containing no disjointness axioms, are coherent.

4.1 Evaluating Terminology Mapping

Evaluating Step 1. Applying *Algorithm 2*, a coherent disjointness assumption set containing 32,299 pairwise disjointness axioms is generated with respect to the OSGB Building and Places ontology. With respect to the OSM ontology, the coherent disjointness assumption set contains 9,348 pairwise disjointness axioms. A sample of 323 and a sample of 93 are taken randomly from these coherent disjointness assumption sets respectively. Based on manual evaluation, the rates of correctness are 0.951 and 0.892 respectively.

Evaluating Step 2. GeoMap, CODI (Niepert et al. 2010) and LogMap (Jimenez-Ruiz and Grau 2011) are employed to match the OSGB Buildings and Places ontology and the OSM ontology (TBoxes), given the generated coherent disjointness assumption sets. The experiments are performed on an Intel Dual Core 2.00 GHz, 3.00 GB RAM personal computer from command line. The experimental results are summarized in Table 1.

GeoMap time in Table 1 is for generating equivalence relations for same-named classes from different ontologies and checking coherence using Pellet. Total time including human interaction (choosing which assumption(s) to retract, time in average is 105.6 s) is 124.4 s. Based on manual evaluation, the precision rates of GeoMap, CODI and LogMap mappings are 89, 76 and 70 % respectively.

Table 1 GeoMap time

	GeoMap	CODI	LogMap
Time[a]	18.8 s (automatic part)	167.72 s	8.65 s
Output	84	105	91
Precision	0.89	0.76	0.70
Recall[b]	0.71	0.76	0.41

[a] Times are in seconds, averaged over 5 runs

[b] The recalls are calculated based on the ground truth shown in Table 2

Table 2 Equivalence relations provided by domain experts

Ground truth[a]	GeoMap	CODI	LogMap
OSGB: Bank ≡ OSM: Bank	1	1	1
OSGB: Chapel ≡ OSM: Chapel	1	1	0
OSGB: Church ≡ OSM: Church	1	1	0
OSGB: Fire Station ≡ OSM: Fire_Station	1	1	1
OSGB: Hotel ≡ OSM: Hotel	1	1	0
OSGB: House ≡ OSM: House	1	1	1
OSGB: Nursery School ≡ OSM: Kindergarten	0	1	0
OSGB: Library ≡ OSM: Library	1	1	1
OSGB: Market ≡ OSM: Marketplace	0	0	0
OSGB: Museum ≡ OSM: Museum	1	1	1
OSGB: Car Park ≡ OSM: Parking	0	−1	0
OSGB: Police Station ≡ OSM: Police	0	−1	−1
OSGB: Public House ≡ OSM: Pub	0	1	0
OSGB: Restaurant ≡ OSM: Restaurant	1	1	1
OSGB: Shop ≡ OSM: Shop	1	0	0.5
OSGB: Town Hall ≡ OSM: Townhall	1	1	1
OSGB: Warehouse ≡ OSM: Warehouse	1	1	0
Score	12	11	6.5

[a] The ground truth is a small set of equivalence relations provided by domain experts. As future work, we will extend the current ground truth to a larger set, to get a more realistic evaluation

The recalls are calculated based on a small set of "ground truth", i.e. equivalence relations provided by domain experts shown in Table 2. In Table 2, "1" means the mapping contains that relation, "0" means not. "−1" means the mapping contains a "wrong" relation. For example, CODI mapping contains an incorrect relation *OSGB*: *Parking*[9] ≡ *OSM*: *Parking* rather than *OSGB*: *Car Park* ≡ *OSM*: *Parking*. "0.5" means the mapping contains partially the relation. For example, LogMap mapping contains *OSGB*: *Shop* ⊑ *OSM*: *Shop* instead of *OSGB*: *Shop* ≡ *OSM*: *Shop*.

Table 1 shows that, LogMap calculates a mapping very quickly, the precision rate of GeoMap mapping is the highest, whilst the recall of CODI mapping is the highest. LogMap is designed for matching large-scale ontologies, especially in biomedical domain, in a reasonable time. As mentioned before, we assume that the TBoxes of geospatial ontologies are not very large, but contain concepts which are more ambiguous, compared to biomedical ontologies. We will focus on precision and recall at the current stage of research.

The precision of GeoMap mapping is high, since domain experts are involved to make ultimate decisions. Consider the following example. In the OSM ontology, several classes, such as *Bicycle*, *Clothes*, *Hardware* and *Kitchen*, are defined as subclasses of *Shop*, indicating what a shop sells. In this context, *OSM*: *Clothes* does not refer to clothes, but a clothes shop. Domain experts retract *OSGB*:

[9] The meanings of OSGB concepts are usually normal. *OSGB*: *Parking* ⊑ *OSGB*: *Purpose*.

Clothes[10] ≡ *OSM*: *Clothes*, whilst, CODI removes *OSGB*: *Shop* ≡ OSM: *Shop* and keeps the existing correspondences of subclasses of *Shop*, optimizing the sum of similarity-weighted probability. LogMap weakens the equivalence relation to *OSGB*: *Shop* ⊑ *OSM*: *Shop*. Though mappings calculated by GeoMap and CODI are always coherent[11] with respect to input ontologies, the results based on optimization may not be reasonable in several cases, especially when informal information exists. Domain experts do not necessarily make the same choices, and an individual domain expert may make different decisions on different occasions.

CODI produces more correct correspondences, such as *OSGB*: *Nursery School* ≡ *OSM*: *Kindergarten* and *OSGB*: *Public House* ≡ *OSM*: *Pub*, owing to its usage of more intelligent lexical matching techniques. However, CODI trades off its precision rate, since it also produces incorrect relations, like *OSGB*: *Race Horse*[12] ≡ *OSM*: *Horse_Racing*. The recall of CODI is better than that of GeoMap or LogMap, however, it is still far from covering all ground truth relations. All three fail to calculate the relation *OSGB*: *Market* ≡ *OSM*: *Marketplace* and miss relations of other types, such as inclusions[13] and overlaps. To improve the recall, more sophisticated lexical matching methods are required, and domain experts are needed, at least at the current stage of development.

The experimental results show that domain experts are indispensable when matching terminologies in order to obtain 100 % precision and recall. Mappings produced by fully automatic methods, such as CODI and LogMap, require final validation by experts, which is difficult and time-consuming. Our method reduces human effort by directing experts to make ultimate decisions during the matching process. As future work, more methods need to be developed to support the manual intervention stage, minimizing human efforts.

4.2 Evaluating Geospatial Instance Mapping

We currently require that, geospatial individuals from different ontologies have polygonal geometries, represented as two dimension vector data. The instance data for OSGB Buildings and Places ontology is extracted from the OSGB Address Layer 2 and the OSGB Topology Layer (Ordnance Survey 2012). The Address Layer 2 is a point layer, containing lexical and classification information for spatial individuals. The Topology Layer is a polygon layer, containing geometries of spatial individuals. These two layers are linked together by finding the smallest Topology Layer polygon containing a point from the Address Layer 2. The OSM instances are from the building layer (containing polygonal geometries, names and

[10] *OSGB: Clothes* refers to "garments worn over the body". It is a secondary concept.

[11] A LogMap mapping may not.

[12] *OSGB*: *Race Horse* ⊑ *OSGB*: *Animal*. It is used to define *OSGB*: *Racing Stables*. .

[13] LogMap weakens some equivalence relations to inclusions, but also does not produce enough.

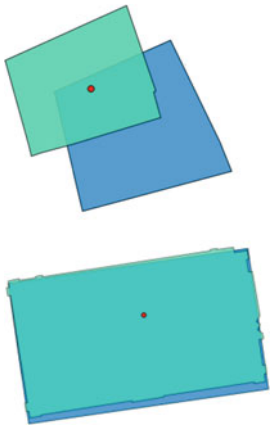


Fig. 1 Prezzo Ristorante (*Up*) and capital one

types) of OSM data, downloaded through Geofabrik (Geofabrik GmbH Karlsruhe 2012) in April, 2012. From the studied area of Nottingham city centre, 713 geospatial individuals are added to OSGB Buildings and Places ontology, 253 geospatial individuals are added to OSM ontology automatically, resulting in two consistent ontologies. Each geospatial individual is classified to a class based on its type information, and has geometry information and lexicon information as its data properties.

Evaluating Step 3 and Step 4. When matching geospatial individuals, geometry matching (*geo_poss_match*) is necessary since two different spatial objects may have the same name. For example, *OSGB*: *1000002308426350* refers to a restaurant called 'PREZZO RISTORANTE'. So does *OSGB*: *1000002309000257*. However, they are actually different restaurants which are distant from each other. Without any geometry checking, the existing data interlinking tools, for example KnoFuss (Nikolov et al. 2007), will map them to the same spatial object *OSM*: *116824670*. In *Step 3*, only *OSGB*: *1000002309000257* and *OSM*: *116824670* are linked (Fig. 1), since their polygons contain the same point from the OSGB Address Layer 2. Lexical matching (*lex_poss_match*) is necessary since two different objects may share the same location. For example, *OSGB*: *1000002308427059* refers to an *OSGB*: *Clinic*, labelled as 'N E M S PLATFORM ONE PRACTICE', while *OSGB*: *1000002308427060* refers to a general commercial company, labelled as 'CAPITAL ONE (EUROPE) PLC', in the same building. Without lexical matching, both will be mapped to *OSM*: *17505332*, labelled as 'CAPITAL ONE', based on geometry similarity (Fig. 1). Cardinality checking is necessary since the same object may be represented as a whole in one ontology, whist as several parts in the other. For example, *OSGB*: *1000002308430942* refers to a NatWest bank in the Victoria Centre. *OSGB*: *1000002308429872* refers to Millies Cookies, a bakery in the Victoria Centre. Without cardinality checking, both will be 'sameAs', rather than 'partOf', *OSM*: *16469518*, the Victoria Centre (Fig. 2).

In *Step 4*, domain experts are consulted to make decisions to repair inconsistencies. For example, *OSGB*: *1000002308476718* refers to an *OSGB*:

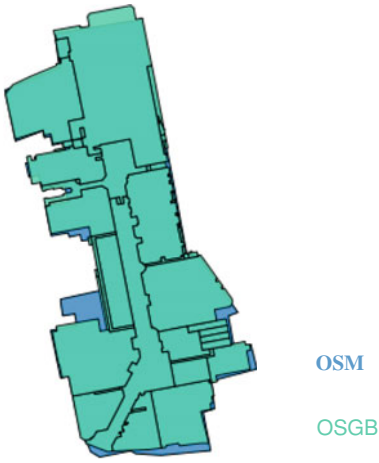


Fig. 2 Victoria centre

HealthCentre labelled as 'SNEINTON HEALTH CENTRE'. *OSM*: *62134030* refers to an *OSM*: *Clinic* labelled also as 'SNEINTON HEALTH CENTRE'. Their geometries are very similar. However, the existence of the following assumptions leads to inconsistency.

$$(OSGB : 1000002308476718, \\ OSM : 62134030) \in sameAs \tag{6}$$

$$OSGB : Clinic \equiv OSM : Clinic \tag{7}$$

$$OSGB : Clinic \sqsubseteq OSGB : HealthCentre \tag{8}$$

Domain experts are consulted to decide which assumption(s) to retract. To keep the individual correspondence (6), it is reasonable to retract (8) or weaken (7) to *OSGB*: *Clinic* $\sqsubseteq$ *OSM*: *Clinic*. This differs from all other methods, which use (8) as a premise, which is not retractable.

Based on manual evaluation, more than 95 percent of the output 139 individual correspondences (37 "sameAs" and 102 "partOf") are reasonable.

Though the initial experimental results seem promising, we are aware that there is still a long way to go before being able to apply this method into practice. Firstly, the "semantic gap" that exists between databases and their corresponding ontologies makes it difficult to populate all individuals from databases to ontologies automatically. For example, "Bar", "POBox" and "Cafe" are individual types in the OSGB database, but are not defined as concepts in the OSGB Buildings and Places ontology. Additionally, some lexical and classification information in OSM data set might be missing, in which case, only geometry matching can be applied. Furthermore, though several geometry matching and lexical matching techniques have been developed, almost none of them ensure overall correctness and consistency of results. As future work, we will explore new ways to make full use of geometry, lexical and classification information for

matching geospatial individuals, aimed for overall correctness and consistency, and minimized human effort.

5 Conclusion

In conclusion, we propose a new semi-automatic method to match disparate geospatial ontologies, guaranteeing the coherence and consistency of overall information. Differing from other existing methods, disjointness axioms and mappings are seen as assumptions, which can be retracted later if found to be too restrictive or inappropriate. A series of algorithms are designed to match disparate ontologies from terminology level to instance level by calculating a coherent and consistent assumption set with respect to them. Geometry matching, lexical matching and logical consistency checking are combined for matching geospatial individuals. The initial experiments show promising results, which indicate that, when matching geospatial ontologies, using geometry or location information helps and domain experts are indispensable. As future work, we plan to develop more sophisticated matching methods, aimed at obtaining 100 % precision and recall, and minimizing human effort.

References

Baader F, Calvanese D, McGuinness DL, Nardi D, Patel-Schneider PF (eds) (2007) The description logic handbook. Cambridge University Press, Cambridge

Bouquet P (2007) Contexts and ontologies in schema matching. Context and ontology representation and reasoning. Roskilde University, Denmark

Bouquet P, Serafini L, Zanobini S (2003) Semantic coordination: a new approach and an application. International semantic web conference, pp 130–145

Buccella A, Cechich A, Fillottrani P (2009) Ontology-driven geographic information integration: a survey of current approaches. Comput Geosci 35:710–723

Domingos P, Lowd D, Kok S, Poon H, Richardson M, Singla P (2008) Just add weights: markov logic for the semantic web. Uncertainty reasoning for the semantic web I, ISWC International Workshops, URSW 2005–2007, Revised Selected and Invited Papers, 2008, pp 1–25

Du H, Anand S, Alechina N, Morley J, Hart G, Leibovici D, Jackson M, Ware M (2012) Geospatial information integration for authoritative and crowd sourced road vector data. Transactions in GIS, Blackwell Publishing Ltd, 2012, 16, 455–476

Euzenat J, Shvaiko P (2007) Ontology matching. Springer, Berlin

Geofabrik GmbH Karlsruhe: Geofabrik (2012) http://www.geofabrik.de

Giunchiglia F, Dutta B, Maltese V, Farazi F (2012) A facet-based methodology for the construction of a large-scale geospatial ontology. J Data Semant 1:57–73 Springer

Giunchiglia F, Shvaiko P, Yatskevich M (2004) S-Match: an algorithm and an implementation of semantic matching. European semantic web conference (ESWC), pp 61–75

Gruber TR (1993) A translation approach to portable ontology specifications. Knowl Acquisition 5:199–220

Gurobi Optimization Inc (2012) Gurobi optimizer reference manual. http://www.gurobi.com

Hart G, Dolbear C, Kovacs K, Guy A (2008) Ordnance survey ontologies. http://www.ordnancesurvey.co.uk/oswebsite/ontology

International Organization for Standardization (2011) ISO/DIS 19157: Geographic information—Data quality

Jackson MJ, Rahemtulla H, Morley J (2010) The synergistic use of authenticated and crowd-sourced data for emergency response. The 2nd international workshop on validation of geo-information products for crisis management (VALgEO). Ispra, Italy, pp 91–99, 11–13 Oct 2010. Available online: http://globesec.jrc.ec.europa.eu/workshops/valgeo-2010/proceedings

Jain P, Hitzler P, Sheth AP, Verma K, Yeh PZ (2010) Ontology alignment for linked open data. Int Semant Web Conf 1:402–417

Jean-Mary YR, Shironoshita EP, Kabuka MR (2010) ASMOV: results for OAEI 2010. The 5th international workshop on ontology matching (OM-2010)

Jiménez-Ruiz E, Grau BC (2011) LogMap: logic-based and scalable ontology matching. Int Semant Web Conf 1:273–288

Jiménez-Ruiz E, Grau BC, Horrocks I, Llavori RB (2009) Ontology integration using mappings: towards getting the right logical consequences. The 6th european semantic web conference (ESWC), pp 173–187

Meilicke C, Stuckenschmidt H (2009) An efficient method for computing alignment diagnoses. In: Third international conference on web reasoning and rule systems, pp 182–196

Meilicke C, Stuckenschmidt H, Tamilin A (2008a) Reasoning support for mapping revision. J Logic Comput 19:807–829

Meilicke C, Völker J, Stuckenschmidt H (2008b) Learning disjointness for debugging mappings between lightweight ontologies. Proceedings of the 16th international conference on knowledge engineering: practice and patterns, Springer, Berlin, pp 93–108

Niepert M, Meilicke C, Stuckenschmidt H (2010) A probabilistic-logical framework for ontology matching. American association for artificial intelligence for ontology matching. In: Proceedings of the 24th AAAI Conference on Artificial Intelligence, Atlanta, Georgia, AAAI Press

Nikolov A, Uren V, Motta E (2007) KnoFuss: a comprehensive architecture for knowledge fusion. The 4th international conference on knowledge capture, ACM, NY, pp 185–186

OpenStreetMap (2012) The Free Wiki World Map. http://www.openstreetmap.org

Ordnance Survey (2012) Ordnance Survey. http://www.ordnancesurvey.co.uk/oswebsite

Qi G, Ji Q, Haase P (2009) A conflict-based operator for mapping revision: theory and implementation. In: Proceedings of the 8th *international semantic web conference*. ISWC '09, Springer, Berlin, Heidelberg, pp 521–536

Reul Q, Pan JZ (2010) KOSIMap: use of description logic reasoning to align heterogeneous ontologies. The 23rd international workshop on description logics (DL 2010)

Sais F, Pernelle N, Rousset MC (2007) L2R: A logical method for reference reconciliation. In: AAAI conference on artificial intelligence. pp 329–334

Scharffe F, Liu Y, Zhou C (2009) RDF-AI: an architecture for RDF datasets matching, fusion and interlink. IJCAI 2009 workshop on Identity, Reference and Knowledge Representation (IR-KR)

Shvaiko P, Euzenat J (2012) Ontology matching: state of the art and future challenges. IEEE Transactions on Knowledge and Data Engineering

Sirin E, Parsia B, Grau BC, Kalyanpur A, Katz Y (2007) Pellet: a practical OWL-DL reasoner. Web semantics: science, services and agents on the World Wide Web, Elsevier Science Publishers B. V., vol 5, pp 51–53

Volz S, Walter V (2004) Linking different geospatial databases by explicit relations. International society for photogrammetry and remote sensing (ISPRS) congress, communication vol IV, pp 152–157

W3C (2009) OWL 2 Web Ontology Language. http://www.w3.org/TR/owl2-overview

Wang P, Xu B (2008) Debugging ontology mappings: a static approach. Comput Artif Intell 27(1):21–36

Wolger S, Siorpaes K, Bürger T, Simperl E, Thaler S, Hofer C (2011) A survey on data interlinking methods. Semantic Technology Institute (STI) Innsbruck, University of Innsbruck. Available online: http://www.insemtives.eu/publications/A_Survey_on_Data_Interlinking_Methods.pdf

On the Formulation of Conceptual Spaces for Land Cover Classification Systems

Alkyoni Baglatzi and Werner Kuhn

Abstract Cognitive approaches to knowledge representation improve man–machine communication, as they are close to human reasoning. Conceptual spaces have been proposed as one such knowledge formalization method. Our research investigates the theory of conceptual spaces as a methodology for implementing semantic reference systems. Conceptual spaces are spanned by quality dimensions. Concepts are represented as regions in n-dimensional spaces and instances as n-dimensional vectors. The land cover domain is chosen for applying this theory with the view to formulating a conceptual space from textual descriptions. Based on a land cover classification system and its descriptions, the methodology for extracting the quality dimensions is demonstrated and their measurement scales are discussed. The usefulness of formalizing a classification system as conceptual space is demonstrated in the process of semantically transforming instances from one classification system to another.

1 Introduction

The challenge of communicating geospatial concepts results from the fact that conceptualizations of geographic entities vary widely. This causes semantic interoperability problems (Bishr 1998), making information exchange and integration procedures error prone and cumbersome. The same terms may refer to

A. Baglatzi (✉)
School of Rural and Surveying Engineering, National Technical University of Athens,
H. Polytechniou Street 9, Zografos Campus, 15780 Athens, Greece
e-mail: baglatzi@mail.ntua.gr

W. Kuhn
Institute for Geoinformatics, University of Muenster,
Weseler Street 253, D-48151 Muenster, Germany
e-mail: kuhn@uni-muenster.de

D. Vandenbroucke et al. (eds.), *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography,
DOI: 10.1007/978-3-319-00615-4_10,

different entities or different terms may be assigned to similar entities. The different conceptualizations, therefore, need to be captured, formalized, and related to each other.

Semantic reference systems (Kuhn 2003) have been proposed as a way to formalize different conceptualizations in the same way as coordinate reference systems define locations on the earth's surface. As a first step, what is needed for a semantic reference system to be realized, is the extraction of its semantic primitives and their formalization. In this work, the case of extracting semantic primitives from land cover category definitions is analyzed. The theory of conceptual spaces is utilized then for the formalization process. Conceptual spaces are built from quality dimensions. Concepts are represented as regions and instances as points in them. The semantic primitives extracted from the definitions constitute the quality dimensions of a conceptual space.

The remainder of this chapter is organized as follows. Section 2 provides some background on conceptual spaces and related work. Section 3 analyzes the procedure of formalizing a land cover classification system as a conceptual space. Section 4 describes the usefulness of the formalization based on a semantic transformation example and Sect. 5 concludes the chapter and discusses possible future work directions.

2 Background

2.1 On the Theory of Conceptual Spaces

Gärderfors (2004) introduced the theory of conceptual spaces, underlining the importance of cognition in knowledge representation. Conceptual spaces are spatial representations of concepts with a presumed cognitive basis. Cognitive semantics views meaning as constructed in minds, in contrast to realist semantics where the meaning is supposed to exist in the world (Gärdenfors 2004). In such a way, cognitive semantics acknowledges and captures different conceptualizations of the world.

Conceptual spaces are spanned by quality dimensions representing the properties of an entity in a geometrical structure. Quality dimensions are either directly perceived by our sensory system or innate and inferred from our experiences. Rickard (2006, p. 313) categorizes the quality dimensions into psychophysical and scientific. A set of integral (non-separable) dimensions represents a domain. For instance, the colour domain consists of the three integral dimensions hue, saturation, and brightness. Domains and dimensions are freely chosen by experts and there are no restrictions on their nature. Instances of properties are represented as points with values on all quality dimensions. Concepts are represented as regions.

As argued in Gärdenfors (2004), the main advantage of conceptual spaces is their geometrical structure, providing representational capacities that are absent in traditional symbolic representations. The notion of distance, which is inherent in

conceptual spaces, eases the comparison of concepts and the measurement of their similarity.

In contrast to feature or network based approaches, it takes into consideration the structure of the properties, leading to more representative similarity metrics. Raubal (2004) proposed the formalization of conceptual spaces as n-dimensional vector spaces. He introduced z-transformations as a statistical standardization method for harmonizing the different quality dimensions types. By assigning weights to quality dimensions, contextual information about their importance is added. Aisbett and Gibbon (2001) formalized conceptual spaces considering that the theory is ordered between the symbolic and sub-conceptual level. In a continuation of that work, ambiguity and its influence on reasoning and calculations is analyzed and captured in the formalization (Rickard et al. 2007). For bringing the theory of conceptual spaces into the semantic web, the Conceptual Space Markup Language (CSML) has been proposed in Adams and Raubal (2009a) based on the metric algebra for conceptual spaces (Adams and Raubal 2009b).

In the geographic domain, Schwering and Raubal (2005a) provided a methodology for calculating semantic similarity between geographic entities represented in conceptual spaces. This work was extended by taking into account spatial relations in similarity measurement (Schwering and Raubal 2005b). Keßler used conceptual spaces for the description of data and services applied on a landmark dataset (Keßler 2006) and Ahlqvist enriched the theory of conceptual spaces with fuzzy logic (Ahlqvist 2004), demonstrating its use for conceptual mappings between land cover classes (Ahlqvist 2005).

2.2 Semantic Reference Systems and the Role of Conceptual Spaces

Semantic reference systems (Kuhn 2003) have been introduced as an analogy to spatial reference systems for enabling semantic interoperability via a formalized understanding and sharing of geospatial concepts. A semantic reference system consists of a semantic reference frame and a semantic datum (Janowicz and Scheider 2010). A semantic reference frame captures the conceptualization of a certain universe of discourse and with the aid of ontologies restricts the interpretation of terms. An ontology is understood in the usual way, as an "explicit specification of a conceptualization" (Gruber 1995, p. 908). Building blocks of a semantic reference frame are the semantic primitives. These are atomic concepts, which cannot be further defined within the universe of discourse. For instance, height, leaf phenology and vegetation types are the semantic primitives of the "vegetation" universe of discourse in the land cover domain. These primitives define the semantic reference frame and serve to define more complex concepts in this universe of discourse (Kuhn and Raubal 2003) like the land cover types Evergreen Needleleave Forest, Shrublands, Woody Savannas etc.

A semantic datum is needed in order to ground the interpretation of the primitives (Kuhn 2009). Grounding is accomplished based on observations (Janowicz and Scheider 2010). According to the most recent definition, a semantic datum "fixes free parameters of measurement scales by grounding them in measurement procedures" (Kuhn 2012). In the vegetation example, the semantic primitive height can be grounded in the ratio measurement scale with meter as the unit of measurement and its known measurement procedure. Only when people share a semantic datum, communication can be established without misunderstandings or misinterpretations. Concept definitions, as traditionally used for communicating shared conceptualizations, can be seen as the source of information for extracting the semantic reference frame. In our research, conceptual spaces are investigated as a framework for implementing semantic reference systems. The quality dimensions of the conceptual space stand for the primitives of a universe of discourse. These quality dimensions span the semantic reference frame. As a semantic reference frame is an ontology of primitives, all the concepts of a universe of discourse need to be formulated as combinations of these semantic primitives. In the same way, concepts in a conceptual space are defined as combinations of the values on the quality dimension. What is then needed is to fix the interpretation of the quality dimensions (Janowicz and Scheider 2010) in order to ground the conceptual space. This is achieved by assigning measurement scales and reproducible measurement procedures to the quality dimensions (Scheider 2011).

Knowing the conceptual space for different universes of discourse, the ultimate goal is to establish transformations among them. In such a way, different conceptualizations can be related. In contrast to spatial transformations, which are well studied in geodesy, there is no established methodology for semantic transformations in conceptual spaces.

3 Constructing a Conceptual Space for Land Cover Classification Systems

The following sections describe the process of designing a conceptual space. We describe the quality dimensions and their values and discuss the advantages and weaknesses of conceptual spaces with respect to the measurement scales and the geometry. For demonstration purposes we use the IGBP-DIS land cover classification system (Townshend et al. 1994).

3.1 The Quality Dimensions

In order to construct a conceptual space, the quality dimensions representing properties, have to be identified first. Information sources for dimensions are definitions, describing the properties of land cover classes either directly or via

some values. For instance, in the following definition, the presence of trees, canopy cover exceeding 60 %, tree height exceeding 2 m, the presence of vegetation throughout the year, the needleleaf tree type, are some of the properties that characterize an instance classified as Evergreen Needleleaf Forest:

> Evergreen needleleaf forest: lands dominated by trees with a percent canopy cover >60 % and height exceeding 2 m. Almost all trees remain green all year. Canopy is never without green foliage.

Specifically, the property tree canopy cover has the value "greater than 60 %", the property leaf phenology has the value "remains green all year" (evergreen). One can see that such descriptions contain information about the values of the properties used in the classification. Sometimes quality dimensions and their values are both present in the descriptions i.e. the quality dimension tree canopy cover and its value ">60 %" or the height quality dimension and its value ">2 m" are explicitly stated in the given description.

Yet, the process of identifying quality dimensions can be more complicated when the properties have to be inferred from their values. For instance, the value "ever-green" refers to a certain type of trees that keep their leaves throughout the year and is found in the literature as leaf phenology. Leaf phenology is defined as "the arrangement of leaves in time" (Kikuzawa 1995, p. 1) and takes values "evergreen", "deciduous" and "mixed". In this case, the quality dimension leaf phenology is inferred from the value "evergreen". The same applies to the value "needleleaf", which implies the quality dimension leaf type. A basic quality dimension is the cover dimension, whose values "vegetation", "water", "bare land" lead to an elementary classification of the earth surface. The quality dimension life form, which is inferred from different vegetation values like "trees", "shrubs", "bushes" etc., leads to the categorization of the main vegetation classes (forest, shrubland, grassland etc.). The quality dimensions height, cover, leaf type and phenology are also seen in Mayaux et al. (2006) as classifiers of the GLC2000 land cover classification system, that is the basic parameters that define this scheme.

The relation between qualities and their values is described in several ways i.e. determinable and determinate (Johansson 2000), quantity dimension and quantity (ISO 2004), quality type and quale (Masolo et al. 2003). Values on the quality dimensions can be also seen as magnitudes, i.e. a "conceptualization of the corresponding property" (Bunge 1973, p. 108). Our understanding of quality dimensions and values (or magnitudes) follows the explanation of Probst (2007), according to which "each quality has a certain magnitude that can be located on a quality dimension". These quality dimensions can also be seen as the semantic primitives of the land cover classification system. In the following, each description of the IGBP-DIS land cover classification system is analyzed in order to identify the quality dimensions.

The quality dimensions cover, leaf phenology, tree canopy cover and height from the previous definition are also common in the Evergreen Broadleaf Forest definition. Additionally, the new value "broadleaf" implies the quality dimension

leaf type which was identified in the previous definition via the value "needleleaf".

> Evergreen broadleaf forest: lands dominated by trees with a percent canopy cover >60 % and height exceeding 2 m. Almost all trees remain green all year. Canopy is never without green foliage.

In the following three descriptions the same quality dimensions were found.

> Deciduous needleleaf forest: lands dominated by trees with a percent canopy cover >60 % and height exceeding 2 m. Consists of seasonal needle-leaf tree communities with an annual cycle of leaf-on and leaf-off periods.

> Deciduous broadleaf forest: lands dominated by trees with a percent canopy cover >60 % and height exceeding 2 m. Consists of seasonal broadleaf tree communities with an annual cycle of leaf-on and leaf-off periods.

> Mixed forest: lands dominated by trees with a percent canopy cover >60 % and height exceeding 2 m. Consists of tree communities with interspersed mixtures or mosaics of the other four forest cover types. None of the forest types exceeds 60 % of landscape.

For Closed and Open Shrublands the quality dimension cover (value = vegetation), life form (value = woody vegetation), height (value = 2 m), shrub canopy cover value = 60 %) and shrub phenology (value = evergreen or deciduous) were identified. For Woody Savannas and Savannas the quality dimensions cover (value = vegetation), life form (value = herbaceous and other understorey systems), shrub canopy cover (value = 10–30 %, 30–60 %) and height (value = >2 m) were found.

> Closed shrublands: lands with woody vegetation less than 2 m tall and with shrub canopy cover is >60 %. The shrub foliage can be either evergreen or deciduous.

> Open shrublands: lands with woody vegetation less than 2 m tall and with shrub canopy cover between 10-60 %. The shrub foliage can be either evergreen or deciduous.

> Woody savannas: lands with herbaceous and other understorey systems and with forest canopy cover between 30-60 %. The forest cover height exceeds 2 m.

> Savannas: lands with herbaceous and other understorey systems and with forest canopy cover between 10-30%. The forest cover height exceeds 2 m.

For Grasslands the corresponding quality dimensions are cover (value = vegetation), life form (value = herbaceous), tree canopy cover (value = <10 %) and shrub canopy cover (value = <10 %).

> Grasslands: lands with herbaceous types of cover. Tree and shrub cover is less than 10 %.

For Permanent Wetlands the quality dimensions cover (value = water and vegetation), life form (value = herbaceous or woody vegetation) and water quality (value = salt, brackish, fresh) were found.

> Permanent wetlands: lands with a permanent mixture of water and herbaceous or woody vegetation that cover extensive areas. The vegetation can be present in either salt, brackish, or fresh water.

For Croplands the quality dimensions cover (value = vegetation), life form (temporary crops) was found and for Croplands/Natural Vegetation Mosaics cover (value = vegetation) and life form (value = mosaic of croplands, forest, shrublands and grasslands).

> Croplands: lands covered with temporary crops followed by harvest and a bare soil period (e.g., single and multiple cropping systems. Note that perennial woody crops will be classified as the appropriate forest or shrub land cover type.
>
> Croplands/Natural vegetation mosaics: lands with a mosaic of croplands, forest, shrublands, and grasslands in which no one component comprises more than 60 % of the landscape.

For Urban and Built-in only the quality dimension cover (value = buildings and other man-made structures) was found and for Barren (value = exposed soil) and bare soil type (value = sand, rocks, snow).

> Urban and built-up: land covered by buildings and other man-made structures. Note that this class will not be mapped from the AVHRR imagery but will be developed from the populated places layer that is part of the Digital Chart of the World.
>
> Barren: lands of exposed soil, sand, rocks, or snow and never has more than 10 % vegetated cover during any time of the year.

For Water Bodies the quality dimensions cover (value = water) and water quality (value = fresh or salt) were found while for Snow and Ice only the quality dimensions cover (value = permanent snow and/or ice) were identified.

> Water bodies: oceans, seas, lakes, reservoirs, and rivers. Can be either fresh or salt water bodies.
>
> Snow and ice: lands under snow and/or ice cover throughout the year.

The quality dimensions are summarized in Table 1. They span the conceptual space of the IGBP-DIS land cover classification system. Not all of them are applicable to all classes. For instance, leaf phenology or height are only applicable to vegetation classes, and water quality to water-related classes. Examples of the values (or magnitudes) on these quality dimensions are shown in Table 2.

Table 1 Quality dimensions of the land cover classification systems

	Quality dimensions
q_1	Cover
q_2	Life form
q_3	Tree canopy cover
q_4	Shrub canopy cover
q_5	Shrub phenology
q_6	Height
q_7	Leaf phenology
q_8	Leaf type
q_9	Bare soil type
q_{10}	Water quality

Table 2 Quality dimensions of the land cover conceptual space and their values

Quality dimensions	Values
Cover	Vegetation, soil, water, bare land…
Life form	Trees, bushes, shrubs…
Tree canopy cover	60, 40 %…
Shrub canopy cover	60, 40 %…
Shrub phenology	Evergreen, deciduous
Height	5, 3 m…
Leaf phenology	Evergreen, deciduous…
Leaf type	Needleleaved, broadleaved…
Bare soil type	Ice, sand, rock…
Water quality	Salt, fresh, brakish…

The notion of magnitude implies measurements and scales. Scale is the "mode of representation of the corresponding property" (Bunge 1973, p. 119) and influences the representation of the quality dimensions of the conceptual space. According to Stevens (1946), there are four kinds of scales: nominal, ordinal, interval, and ratio scale.

Qualitative (a.k.a. categorical) values are measured on nominal and ordinal scales. On a nominal scale, labels are assigned to the values on the quality dimensions i.e. trees, shrubs, etc. These values have neither natural distance nor natural ordering. Valid operations on this scale are only equivalence and set membership.

On the ordinal scale, values are ranked according to a common criterion like the brightness of colour. Quantitative (a.k.a. numerical) values are represented on interval and ratio scales. On the interval scale, values are ordered in a quantifiable manner independently of a true zero point. The ratio scale is the one most often used in physical sciences and presupposes the existence of an absolute zero point. Table 3 shows the basic operations for each measurement scale. In order to further analyze the quality dimensions, we follow the classification of Stevens (1946) for the scales. Extensions of the Stevens scales, such as cyclic or absolute, are potentially relevant as well (Chrisman 1998).

In line with Stevens (1946), the correspondence of quality dimension of the conceptual space and measurement scales is shown in Table 4.

Many quality dimensions have categorical values from a nominal scale. This is one of the major challenges of the conceptual space theory when applied to the land cover domain. Although the theory itself does not restrict the type of scales on quality dimensions, there is an immanent difficulty in the geometrical representation of nominals. The assignment of numerical values or even orderings to

Table 3 Scales and their operations from Stevens (1946, p. 678)

Scale	Basic empirical operations
Nominal	Determination of equality
Ordinal	Determination of greater or lesser
Interval	Determination of equality of intervals or differences
Ration	Determination of equality of rations

Table 4 Measurement scales of the quality dimensions of the IGBP-DIS classification system

Quality dimensions	Measurement scale
Cover	Nominal
Life form	Nominal
Tree canopy cover	Ratio
Shrub canopy cover	Ratio
Shrub phenology	Nominal
Height	Ratio
Leaf phenology	Nominal
Leaf type	Nominal
Bare soil type	Nominal
Water quality	Nominal

categorical values, to enable distance computation in the conceptual space, is not supported by nominals. This challenge has not been addressed comprehensively in the conceptual space theory yet.

This is a problem already addressed in different domains i.e. in geology (Brodaric et al. 2004), agriculture (Abdullah et al. 2005) or visualization techniques (Rosario et al. 2004). Ahlqvist and Ban (2007) provide an interesting view of nominal (or, as they call them, non-ordered categorical) data. Although they investigate methods that are used for transforming categorical data, i.e. classes to numerals, they do not further analyze the categorical nature of the values of their properties. This is of little benefit for the present work because what is needed is a method to assign numbers to nominal values of properties and not to the classes themselves.

The methodologies for mapping nominals to numerals can be summarized by two broader categories. One option is to transform nominals into binary values where one and zero represent the existence and absence of the value. The advantage of this is that nominals become a way "quantifiable"; the disadvantage is that the values have no semantics. The second option is to transform nominals into numerals based on the number of occurrences in a dataset. This solution allows for more complex mathematical transformations but does not provide any substantial information about the semantics of the nominals either.

Nominal values are concepts and, as such, should be representable by conceptual spaces themselves. The colour example shows how a concept gets analyzed along quality dimensions (red, green, blue or hue, saturation, intensity). For the land cover domain, this would mean to analyze nominals like evergreen, deciduous, needleleaf, broadleaf etc. into quality dimensions spanning conceptual spaces for these values. The land cover classification system would define one conceptual space for the land cover types and subspaces for the nominal values of these land cover types. The source of information for the conceptual spaces of nominal values would be their descriptions. However, there was not enough information about the description of the nominals in the IGBP-DIS land cover classification system. Adopting descriptions from external sources would be risky, as their use and understanding in the IGBP-DIS classification system is not guaranteed to be the same as in these external sources.

In terms of vector spaces, the conceptual space for land cover can be represented as:

$$\mathrm{CS}_n = (q_1, q_2, q_3, \ldots, q_n), \text{ where } q_i = \text{quality dimension}$$

The land cover classes are then represented as points in this multidimensional vector space. For instance, the class Evergreen Needleleaf Forest is a point in the conceptual space and its coordinates are:

$$c_{ENF} = (\text{vegetation, trees, } > 60\,\%, \text{ NA, NA, } > 2 \text{ m, evergreen, needleleaf, NA}),$$

with NA, the non applicable dimensions are defined.

We have shown how to construct a conceptual space for a land cover classification system. The next step is to demonstrate its use in semantic transformations between classification systems. This procedure is hampered by the lack of a theory or practice accounting for quality dimensions on a nominal scale and by the resulting inability of defining subspaces for nominal values. We will assign existence or absence values to nominals to establish semantic transformations.

4 On the Use of Conceptual Spaces for Land Cover Classification Systems

4.1 Defining Semantic Transformations

Semantic transformation is the process of relating two semantic reference systems and transforming instances from one to the other. The main objective of semantic transformation is to establish a mechanism that enables the interpretation of a term in two different semantic reference systems and allows for mappings between them. A typical need for semantic transformations is seen in the land cover domain where a piece of land is classified under type A.x in the land cover classification system A and under type B.y in the land cover classification system B. By transforming from A to B (and vice versa) the two classification systems become interoperable: land that is classified in system A can be re-classified in system B and vice versa. The domain experts preserve their classification systems, but can communicate about them.

In the spatial case of semantic reference systems, transformations are used for changing the coordinates of points from one system to another. The idiosyncrasy of the earth surface calls for optimized spatial reference systems with ellipsoids, datums and coordinates systems that differ from place to place. As a result of the transformation process, the uniquely assigned coordinates of one point in reference system A are interpretable in reference system B via the transformation functions. Spatial transformations require the parameters of the relative position of the two reference systems (shifts, rotations etc.). If the two spatial datums of the reference

systems are known, transformation parameters can be directly calculated based on the relative position. If the spatial datums are not known, control points, that is, specific points on the earth's surface with known coordinates in the two reference systems, are used to define the transformation parameters. The geometric dimensions of the transformations are fixed, predefined and have the same measurement scale in all cases of transformations, spatial or otherwise.

4.2 Semantic Transformation: An Example

When domains share their semantic primitives, a common conceptual space can be established. In the common conceptual space, land cover types from all classification systems are represented as regions and instances as points (Fig. 1). The perspectives on the instances in the domain are the same as well as the criteria for classification.

Yet, there are aspects of semantic heterogeneity that mostly concern the values on the quality dimensions and the way they are combined to form the land cover types. In order to deal with these problems, semantic transformations are needed. For this type of transformations we adopt the term "conversions" which can be found in geodesy as the transformations that "exclude any change of datum" (Illife and Lott 2008, p. 91). Semantic conversions depend on the relative position of the instances in the common conceptual space.

An example of a common conceptual space is that for the IGBP-DIS and the University of Maryland (UoM) land cover classification system (Hansen et al. 2000). Although the labels, the number of land cover types, and the descriptions are not identical, a common classification ground exists. The purpose of both is the creation of a global land cover dataset. Also, the data (remote sensing image with similar temporal, spatial and spectral characteristics) and the methods (remote sensing image classifications techniques) are compatible. Mappings between these

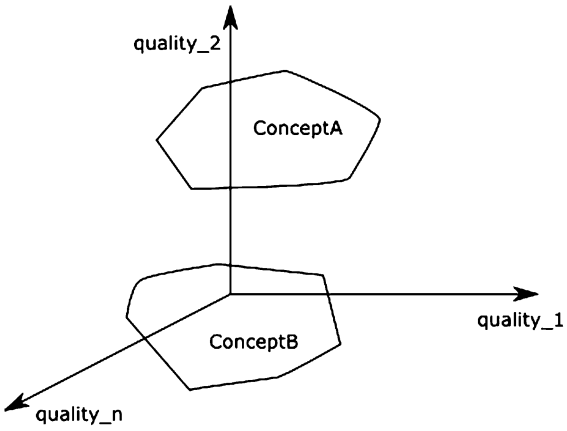


Fig. 1 Concepts in the same conceptual space

Table 5 Quality dimensions and values of the shared conceptual space

Quality dimensions	Values
Cover	Vegetation
Life form	Herbaceous understories, woody understories
Tree canopy cover	30, 40, 60 %
Height	2, 5 m

two classification systems developed by domain experts are used as examples for semantic transformations, for example, between the land cover types 'WoodlandsUoM' of UoM and 'WoodysavannasIGBP' of the IGBP-DIS land cover classification system. The descriptions of the two types are shown below:

> Woodlands: lands with herbaceous or woody understories and tree canopy cover of >40 % and <60 %. Trees exceed 5 m in height and can be either evergreen or deciduous. (UoM)

> Woody savannas: lands with herbaceous and other understorey systems and with forest canopy between 30 % and 60 %. The forest cover height exceeds 2 m. (IGBP-DIS)

The quality dimensions of the common conceptual space and their values are shown in Table 5.

The two concepts, 'WoodlandsUoM' and 'WoodySavannasIGBP', are depicted in Fig. 2. A simplified model has been chosen in a two dimensional space and the two concepts are represented with rectangles. Topologically,[1] 'WoodlandsUoM' is proper part of 'WoodySavannasIGBP'. Conversion in this sense is a scaling between the two concepts.

The conversion from 'WoodySavannasIGBP ' to 'WoodlandsUoM' follows the equation below:

$$Woodlands_{UoM} = k * WoodySavannas_{IGBP}, \text{ where } k > 1,$$

It is evident that there can be a loss of information when converting instances from one system to the other. These losses are due to the limited accuracy of the transformation.

For example, instances that are 'within' the definition of 'WoodySavannasIGBP' but excluded from the definition of 'WoodlandsUoM' will be uncategorized when transforming from the IGBP-DIS classification system to the UoM classification system. In order to quantify the loss, the distance of one instance from the rectangular 'WoodlandsUoM' is measured. It can be calculated as the distance between instances i1, i2,… in that are not described by both definitions, and the prototype P of the narrower concept. In Fig. 3 the distance of the instances i1, i2, i3, i4 of 'WoodySavannasIGBP' from the prototype P of 'WoodlandsUoM', is depicted. The number of the instances that are covered by both definitions divided by the total number of instances indicates the quality of the transformation of a particular pair of datasets.

[1] Following the RCC8 topological relations.

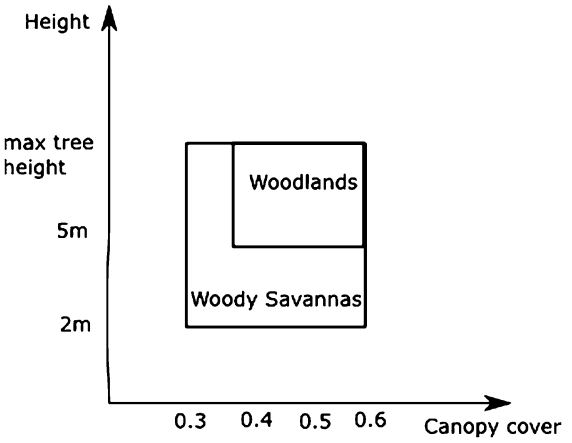


Fig. 2 Woody savannas and woodlands in the same conceptual space

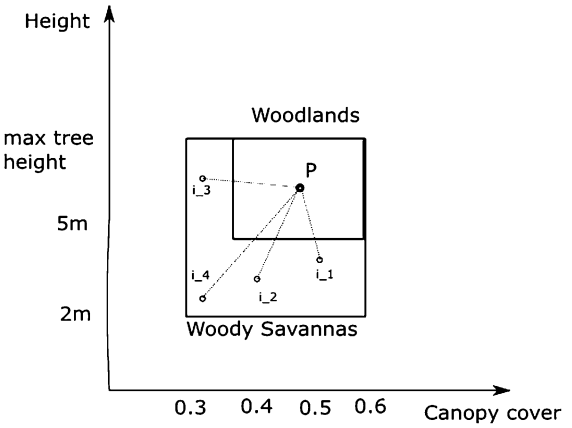


Fig. 3 Calculation of the accuracy of the semantic transformation

$$q = \frac{N_s}{N_{total}},\ where \begin{cases} N_s\ the\ number\ of\ shared\ instances \\ N_{total}\ the\ total\ number\ of\ instances\ of\ the\ two\ datasets \end{cases}$$

The reverse conversion from 'WoodlandsUoM' to 'WoodySavannasIGBP' would follow the equation:

$$Woodlands_{UoM} = k * WoodySavannas_{IGBP},\ \text{where } k < 1.$$

In this case, no losses are observed but the non-overlapping part will be empty, because no values can be assigned to instances of 'WoodlandsUoM'.

In general, semantic conversions in a common conceptual space where one concept is a proper part of another can be expressed as a scaling between the concepts, described as follows:

$$Concept_A = k * Concept_B, where \begin{cases} k > 1 & from\ the\ narrower\ to\ the\ broader\ concept \\ k < 1 & from\ the\ broader\ to\ the\ narrower\ concept \\ k = 0 & infeasible\ conversion\ (i.e.\ disjoint\ concept) \end{cases}$$

5 Conclusions and Future Work

Semantic Reference Systems have been proposed as a means of enabling semantic interoperability of geographic concepts. In this chapter we have used the theory of conceptual spaces to formalize semantic reference systems. We analyze the process of extracting quality dimensions from geographic definitions and discuss their measurement scales on an example from the land cover domain, emphasizing the difficulty in coping with categorical values in conceptual space. Semantic transformations are suggested as a way of translating instances from one conceptual space to another.

One of the main future directions is the development of semantic transformations between multiple classification systems with conceptual spaces of different dimension numbers. The increasing complexity calls for additional transformation types with more complex mathematical operations. Investigating whether multilingual aspects influence the formulation of the conceptual spaces or their relations can be a future perspective.

Acknowledgments This research has been partially funded by the EC funded project ENVISION (contract number 217951) and the European Union Seventh Framework Programme—Marie Curie Actions, Initial Training Network GEOCROWD under grant agreement n FP7-PEOPLE-2010- ITN-264994.

References

Abdullah A, Bulbul R, Mehmood T (2005) Mapping nominal values to numbers by data mining spectral properties of leaves. In: Proceedings of 3rd international symposium on intelligent information technology in agriculture, Beijing

Adams B, Raubal M (2009a) Conceptual space markup language (csml): towards the cognitive semantic web. In: Proceedings of the 2009 IEEE international conference on semantic computing, ICSC '09, IEEE computer society, Washington, DC, USA, pp 253–260

Adams B, Raubal M (2009b) A metric conceptual space algebra. In: Proceedings of the 9th international conference on spatial information theory, COSIT'09, Springer, Berlin, pp 51–68

Ahlqvist O (2004) A parameterized representation of uncertain conceptual spaces. Trans GIS 8(4):493–514

Ahlqvist O (2005) Using uncertain conceptual spaces to translate between land cover categories. Int J Geogr Inf Sci 19(7):831–857

Ahlqvist O, Ban H (2007) Categorical measurement semantics: a new second space for geography. Geogr Compass 1(3):536–555

Aisbett J, Gibbon G (2001) A general formulation of conceptual spaces as a meso level representation. Artif Intell 133(1–2):189–232
Bishr Y (1998) Overcoming the semantic and other barriers to gis interoperability. Int J Geogr Inf Sci 12(4):299–314
Brodaric B, Gahegan M, Harrap R (2004) The art and science of mapping: computing geological categories from field data. Comput Geosci 30(7):719–740
Bunge M (1973) On confusing measure with measurement in the methodology of behavioral science. In: The methodological unity of science, vol 3. p 105
Chrisman NR (1998) Rethinking levels of measurement for cartography. Cartography Geogr Inf Sci 25(4):231–242
Gärdenfors P (2004) Conceptual spaces: the geometry of thought. The MIT Press, Cambridge
Gruber TR (1995) Toward principles for the design of ontologies used for knowledge sharing. Int J Hum Comput Stud 43(5):907–928
Hansen MC, DeFries RS, Townshend JRG, Sohlberg R (2000) Global land cover classification at 1 km spatial resolution using a classification tree approach. Int J Remote Sens 21(6–7):1331–1364
Illife J, Lott R (2008) Datums and map projections: for remote sensing, GIS and surveying, 2nd edn. Whittles Publishing
ISO VIM (2004) International vocabulary of basic and general terms in metrology (vim). Int Organ 2004:09–14
Janowicz K, Scheider S (2010) Semantic reference systems. In: Warf B (ed) Encyclopedia of geography. SAGE Publications, California
Johansson I (2000) Determinables as universals. http://hem.passagen.se/ijohansson/ontology6.htm. Last accessed: 09 Oct 2012
Keßler C (2006) Conceptual spaces for data descriptions. In: The cognitive approach to modeling environments (CAME), workshop at GIScience, pp 29–35
Kikuzawa K (1995) Leaf phenology as an optimal strategy for carbon gain in plants. Can J Bot 73(2):158–163
Kuhn W (2003) Semantic reference systems. Int J Geogr Inf Sci 17:405–409
Kuhn W (2009) Semantic engineering. In: Research trends in geographic information science, chapter 5. Springer, Berlin, pp 63–76
Kuhn W (2012) Semantic reference systems. Lecture notes, reference systems course, ifgi
Kuhn W, Raubal M (2003) Implementing semantic reference systems. In: Gould M, Laurini R, Coulondre S (eds) AGILE 2003—6th AGILE conference on geographic information science, Collection des sciences appliquees de l' INSA de Lyon, Lyon, France, Presses Polytechniques et Universitaires Romandes, pp 63–72
Masolo C, Borgo S, Gangemi A, Guarino N, Oltramari A (2003) WonderWeb deliverable D18 ontology library (final). Technical report, IST project 2001-33052 WonderWeb: ontology infrastructure for the semantic web
Mayaux P, Eva H, Gallego J, Strahler AH, Herold M, Agrawal S, Naumov S, De Miranda EE, Di Bella CM, Ordoyne C et al (2006) Validation of the global land cover 2000 map. IEEE Trans Geosci Remote Sens 44(7):1728–1739
Probst F (2007) Semantic reference systems for observations and measurements. PhD thesis, University of Münster, Germany
Raubal M (2004) Formalizing conceptual spaces. In: Formal ontology in information systems, proceedings of the third international conference (FOIS 2004), vol 114. pp 153–164
Rickard JT (2006) A concept geometry for conceptual spaces. Fuzzy Optim Decis Making 5(4):311–329
Rickard JT, Aisbett J, Gibbon G (2007) Reformulation of the theory of conceptual spaces. Inf Sci 177(21):4539–4565
Rosario GE, Rundensteiner EA, Brown DC, Ward MO, Huang S (2004) Mapping nominal values to numbers for effective visualization. Inf Vis 3(2):80–95

Scheider S (2011) Grounding geographic information in perceptual operations. Unpublished Ph. D Dissertation, University of Münster, Germany. Available at http://geographicknowledge.de/pdf/MyThesis.pdf

Schwering A, Raubal M (2005a) Measuring semantic similarity between geospatial conceptual regions. In: GeoSpatial semantics—first international conference, GeoS 2005, pp 90–106

Schwering A, Raubal M (2005b) Spatial relations for semantic similarity measurement. Perspectives in conceptual modeling. Springer, Berlin, pp 259–269

Stevens SS (1946) On the theory of scales of measurement. Science 103(2684):677–680

Townshend J, Justice C, Skole D, Malingreau JP, Cihlar J, Teillet P, Sadowski F, Ruttenberg S (1994) The 1 km resolution global data set: needs of the international geosphere biosphere programme. Int J Remote Sens 15(17):3417–3441

Part V
Data Mining, Agregation and Disagregation

The Impact of Classification Approaches on the Detection of Hierarchies in Place Descriptions

Daniela Richter, Kai-Florian Richter and Stephan Winter

Abstract The chapter investigates the identification of hierarchical structures in place descriptions. Different approaches to classify spatial granularity will be compared and applied to a corpus of human place descriptions. Results show how hierarchical structures as well as deviations depend on the respective classifications. They further indicate certain difficulties in developing a suitable classification of spatial references. Findings contribute to the understanding of human spatial language, and thus the development of flexible mechanisms for their interpretation and integration in location-based systems.

1 Introduction

Place descriptions are a common way to describe *where* things are. Their hierarchic organization is broadly recognized and evident from various studies (e.g., Plumert et al. 2001; Shanon 1979; Richter et al. 2013). People employ different concepts and perspectives when dealing with space dependent on the tasks that they perform. They conceptualize the world at different granularities (or grain-sizes) by abstracting from it those things that serve their present interests (Hobbs 1985). Accordingly, hierarchical structures emerge in place descriptions that reflect a

D. Richter (✉)
Institute of Photogrammetry and Remote Sensing,
Karlsruhe Institute of Technology, Karlsruhe, Germany
e-mail: daniela.richter@kit.edu

K.-F. Richter · S. Winter
Department of Infrastructure Engineering, The University of Melbourne,
Melbourne, Australia
e-mail: krichter@unimelb.edu.au

S. Winter
e-mail: winter@unimelb.edu.au

D. Vandenbroucke et al. (eds.), *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography,
DOI: 10.1007/978-3-319-00615-4_11,

hierarchical organization of spatial knowledge in the mind (Hirtle and Jonides 1985; Stevens and Coupe 1978), and serve the purpose of anchoring the location of a thing or event to known places and of disambiguating places of specific granularity levels.

Over the years, a number of classification schemes for spatial granularity have been proposed for various purposes. It often depends on which classification scheme is applied whether a particular place description has a recognizable hierarchical structure, and whether this structure is sequential or contains gaps. For example, using a typical address scheme of *street name*, *city*, *country* the place description 'Grattan St, Australia' is hierarchical, but contains a gap. Using instead a scheme inspired by embodied experience, say *personal space* (everything at arm's length) and *environmental space* the same place description becomes flat, i.e., non-hierarchical.

Since human place descriptions, for reasons of efficiency, usually follow Grice's maxims of conversation (Grice 1975) they usually refer to relevant places. Then levels of granularity matter: to characterize the resolution of a place description, and to identify the places of coarser resolution that disambiguate the ones of finer resolution (e.g., disambiguating Melbourne, Victoria, from Melbourne, Florida). While hierarchical structures in place descriptions are important in human conceptualization of and communication about space, we claim that their identification in automated processes will depend on the applied classification of spatial granularity. This may also cause the detection of gaps or flat structures. Alternatively, there may be other reasons for these deviations, for example, cognitive principles of salience and prominence, that may explain these deviations. We will investigate two research questions:

1. How are hierarchies of place descriptions related to the applied classification scheme; and,
2. Can deviations (in form of flat structures or gaps) be avoided by improving the classification?

To address these questions we will study a corpus of human place descriptions collected in a mobile game, in which participants had to answer the question 'Tell us where you are' (Winter et al. 2011). While place descriptions are context-specific and can refer to all kinds of things in space, our focus here is on people's locations, which suggests a finest resolution limit related to the size of the human body. The corpus was previously analyzed to learn about general hierarchical structures in place descriptions using a particular classification scheme (Richter et al. 2013). In this chapter the previous approach will be compared with other classification schemes to see whether we find support for the research questions above, and whether issues emerge from the comparison that were not known before.

The research findings will contribute to our knowledge on hierarchical organization principles in place descriptions. Granularity plays a crucial role for developing systems for automated interpretation of and reasoning on spatial information (e.g., answering *where* questions); understanding hierarchical structures is essential in this regard.

The chapter is structured as follows. Section 2 reviews related work; Sect. 3 presents our approach, Sect. 4 introduces the corpus and analysis methods, Sects. 5 and 6 then present and discuss the results.

2 Related Work

This section reviews related research regarding place descriptions and their hierarchical organization. It further summarizes approaches to classifying spatial granularity.

2.1 Hierarchical Place Descriptions

Spatial mental representations are acquired through direct and indirect interaction with the environment (Ishikawa and Montello 2006; Siegel and White 1975). The mental organization of spatial knowledge is based on an individual's acquisition of this knowledge, and distorted by preferential reasoning through anchor points (Sadalla et al. 1980), i.e., asymmetric relationships caused by different salience, and hierarchical structures defined by partonomies (Hirtle and Jonides 1985; Stevens and Coupe 1978). If salience causes asymmetric relationships it imposes an order independent from partonomies. Measures for salience have been suggested (e.g., Raubal and Winter 2002; Sorrows and Hirtle 1999), but they are local measures, providing an order only in a given context. They do not lend themselves to building a global hierarchy.

Verbal place descriptions reflect these cognitive organization principles of spatial knowledge. They inherit the hierarchical organization of spatial knowledge (Plumert et al. 2001; Richter et al. 2013; Shanon 1979). Hierarchical structures are employed to decrease the cognitive effort of storing and retrieving information, and to decrease ambiguity in spatial knowledge sharing. From a linguistic perspective, place descriptions are referring expressions (Dale 1992) to locations of objects. Gestalt theory suggests that their focus on the object identifies the figure, and other location references are taken from the ground (the environment) (Talmy 1983). Normally people select the most relevant referents from a possible set of referents. The principles of relevance are two-fold: a cognitive principle that human cognition is geared to the maximisation of relevance, and a communicative principle that utterances create expectations of optimal relevance (Sperber and Wilson 1986). These conversation principles are reflected, for example, in *generating* hierarchically organized place descriptions (Kelleher and Kruijff 2006; Tomko and Winter 2009).

2.2 Classification Approaches for Spatial Granularity

Granularity describes varying levels of abstraction of a phenomenon, which form a hierarchy. Either the finer levels of a hierarchy contain representations that are more detailed than the coarser levels, as in cartographic generalizations, or the finer levels will contain smaller objects that are aggregated at coarser levels, as in partonomies (Timpf 1998). Different understandings and (formal) definitions of granularity exist (e.g., Bittner and Smith 2003; Hobbs 1985; Keet 2006). A hierarchy will necessarily contain at least two different levels of granularity.

Rosch et al. (1976) introduced the concept of *basic objects*, which relate to the preference of basic level categories in cognitive categorization from sub- or supercategories (e.g., the preference for using the word 'table' instead of 'kitchen table' or 'furniture', when asked 'where is the cup?'). Similarly, basic-level geographic categories exist (Smith and Mark 2001) (e.g., 'country', or 'city', with their superordinate category 'place', or subordinate categories such as 'home country').

To select references for destination descriptions Tomko and Winter (2009) developed a model based on three types of hierarchically structured data: a containment hierarchy of districts, the likelihood of using specific streets, and the visual and semantic salience of landmark buildings. Also *SpatialML* (Mani et al. 2010), a markup language for the annotation of natural language references to places, uses spatial granularity in form of tags for different feature types, such as *country*, *state* or *populated place*. Granularity was also applied for the study of place descriptions (Plumert et al. 2001; Richter et al. 2013; Tenbrink and Winter 2009).

3 Place Descriptions and Their Hierarchical Classification

This chapter studies spatial granularity as in differences in perceived or actual size of geographic entities. If these entities are related (e.g., by containment) a partonomy hierarchy emerges. In an address scheme of *street name*, *city*, *state*, 'Grattan St, Parkville, Victoria' is a hierarchical place description by virtue of the entities on finer granularity levels being contained in those on coarser level. Skipping 'Parkville' in that description, a *gap* emerges because there is an element of the schema missing in the sequence. Removing 'Grattan St' or 'Victoria' on the other hand would not create a gap as the remaining elements adhere to the sequence of granularity levels.

In their seminal work on basic categories, Rosch et al. (1976) pointed out that cognitive economy requires a balancing between fine-grained distinctions and fewer categories. With too few categories, relevant distinctions cannot be made, and with too many categories, cognitive representation and reasoning become slow and hard to maintain. Applied to place descriptions and their contained hierarchical structures, there needs to be a balance between enough granularity levels to

pick up these structures, and few enough levels not to introduce gaps that are not really there.

This section will define place descriptions and the types of possible hierarchical structures (Sect. 3.1), and then introduce a selection of existing classification approaches (Sect. 3.2) that will be analyzed in the remainder of the chapter.

3.1 Hierarchies in Place Description

A place description is a verbal description answering a *where* question. A typical form to describe the location of something is:

$$PD : [[subject\ verb]\ preposition]\ NP$$

with brackets indicating optional elements of the place description *PD*.

The noun phrase *NP* is a locative noun phrase, as in '[[I'm] in] Brunswick'. It can consist of just a noun ('Brunswick'), a compound ('Brunswick Baths'), or a complex phrase aggregated from simpler noun phrases and relationships ('Brunswick, near the train station'). The noun phrases refer to geographic entities of a particular level of granularity. For example, 'intersection' is of finer granularity than 'downtown'. A hierarchical structure in a place description is defined as a structure consisting of 1 to *n* granularity levels. L_1 is the lowest level (the most fine-grained) in the hierarchy *H*; L_n the highest (most coarse-grained) level: H: $(L_1)(L_2)(L_3)\ldots(L_n)$. A place description can expose one or multiple levels of granularity that form one of the following hierarchy patterns (cf. Richter et al. 2013 for further details):

- *Strictly hierarchical*: a place description showing a strictly monotonically increasing or decreasing behavior towards the spatial hierarchy. The sequence of granularity levels is either zooming in or zooming out; no duplicates of the same levels occur.
- *Partially hierarchical*: a place description showing a monotonically increasing or decreasing behavior. Duplicates of the same levels occur.
- *Flat*: a place description that shows constant behavior towards the spatial hierarchy. At the same time monotonically increasing and decreasing (no zooming in or out), they form a special type of partially hierarchical descriptions.
- *Unordered*: a non-monotonic place description.

The order of granularity levels in hierarchical place descriptions determines zooming behavior, but also whether gaps within the pattern occur or a gap-less sequence is formed.

3.2 Classifying Spatial Granularity

Freundschuh and Egenhofer (1997) reviewed a number of classifications of spatial granularity with a focus on scales of *human conceptions of space*. The reviewed classification models cover varying numbers of distinguishing levels. Some only distinguish between large and small space (Kuipers 1978), or between elements of different dimensions to model the structure of a city (Lynch 1960). Others have four (Montello 1993; Zubin 1989), five (Couclelis and Gale 1986), or six levels of granularity (Kolars et al. 1975).

We will investigate a subset of these reviewed models here, selected by their potential to adequately capture spatial granularity in place descriptions. Coming back to the argument of distinctions versus number of categories (Rosch et al. 1976), having at least four different levels of granularity was identified as one prerequisite. The number of different levels alone is not sufficient, however. For example, while Zubin's taxonomy of spatial objects and spaces (Zubin 1989) has four levels, the classification is highly dependent on the view point. An object may be classified as type *A* in one situation, and as type *B* in another, which leads to ambiguous and unstable classifications.

Freundschuh and Egenhofer also developed their own categorization of space. In general, this classification is similar to that of Montello (1993) (see below), however, it also includes *panoramic* and *map* space, which are not useful for classifying human place descriptions. Their scheme is excluded from the investigation in favor of Montello's. The other classification scheme selected from their review is the one by Kolars et al. (1975). These two will be compared with a classification scheme by Richter et al. (2013), and the geocoding scheme of the Google Geocoding API Version 2.0, which provides also a geocoding accuracy value. In the remainder of the chapter these classifications will be referred to as *Montello*, *Kolars*, *Richter*, and *Google*, respectively.

Montello

Montello (1993) proposed four major classes of *psychological spaces*:

- *Figural space* is small in scale relative to the human body and is apprehended without any locomotion. It includes both the flat pictorial space and the space of small manipulable objects.
- *Vista space* is larger than the human body but can be visually scanned from a place without moving around.
- *Environmental space* is large in scale relative to the body. It includes the spaces of buildings, cities and neighborhoods and typically requires locomotion for its apprehension. It is learned over time.
- *Geographical space* is much larger than the human body and cannot be experienced directly, instead it is perceived only over time and typically through symbolic representations, such as maps.

Theoretically, vista space can range from small objects up to the world ('the surface of the earth as seen from an airplane, however, would constitute a vista space because of its small projective size and our consequent ability to apprehend it directly from our seat in the plane' (Montello 1993, pp. 315–316). However, place descriptions focus on the question *where* people are in their surrounding space ('in the plane'). The notion of vista space has thus been adapted to better reflect the context of place descriptions and to get a clearer separation between vista and environmental space.

Kolars

Kolars et al. (1975) defined a hierarchy of six geographic spaces based on the level of interaction among and between people and their surrounding environment:

- *Personal space* is the small space within a person's arm's length that involves only a few people and primary interaction modes of voice, touch, taste, and smell.
- *Living and working space* is the space of the personal daily life (the home, the office), the space of effective personal communication. It involves one-to-one interactions among 40–500 people and audio or visual modes of communication.
- *House and neighborhood space* constitutes, for example, a group of houses or structures along a street, such as gardens or small parks. It is the space of impersonal interactions beyond face-to-face communication that use amplified audio-visual communication modes. Distances may range from 30 to 300 m (may be limited by the line of sight). It may involve larger groups of people (100–1,000) who do not know each other but share some common purpose (e.g., places for meetings, private domiciles, parks, playgrounds).
- *City-hinterland space* consists of neighborhoods and specialized areas which have different functions than a cluster of households (e.g., towns and cities). It is the space of the daily information field within about 60 min travel distance, within the range of the local news media, and urban institutions such as local and metropolitan governments, involving 50,000–10 million people.
- *Regional-national space* constitutes of clusters of cities, is the space of legal-economic-political systems, and involves interaction via national network of news media among 200+ million people.
- *Global space* is the space of the trade and cultural exchange with interaction via international communication networks; that involves five billion and more people.

Richter

Richter et al. (2013) distinguished seven levels of spatial granularity. The *furniture level* refers to locations of furniture (in- or outdoor), including small vehicles (a bike) or natural features (a tree), whereas the *room level* describes a specified location within a building. It can also include medium vehicles, such as cars or

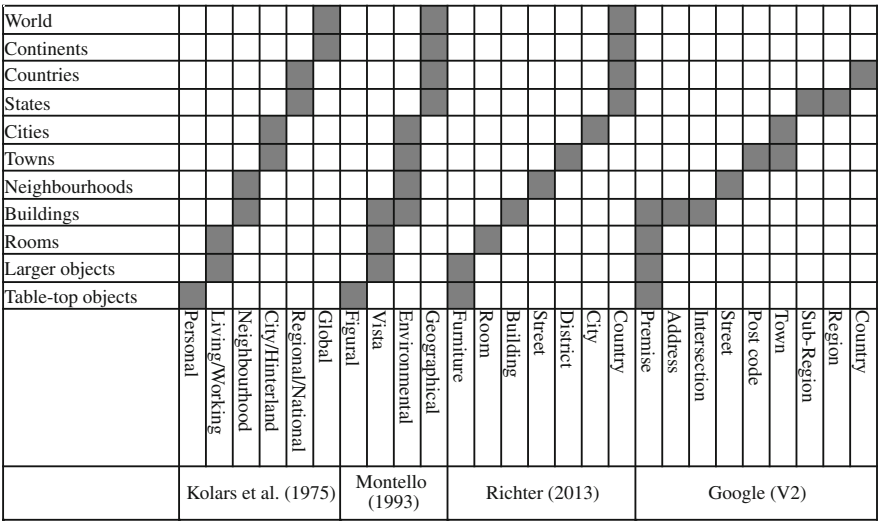


Fig. 1 Models of spatial granularity [extending on Freundschuh and Egenhofer (1997, p.369)]

boats. The *building level* refers to locations of a building, for example a house number, a street corner, a building name, or large vehicles (train, ferry). *Street level* refers to institutions, public spaces or streets. *District level* includes suburbs, rural districts, localities and post code areas. Finally, *city level* refers to a town or a city, including metropolitan areas, and *country level* to everything beyond city level, including highways, rivers, national parks, and states.

Google

The Google Geocoding API Version 2.0[1] returns an accuracy value with each returned placemark, which indicates the resolution of the given result. For example, the geocode '28/207 Queens Ave, Hawthorn 3122 Melbourne' may return 8 (address level accuracy), i.e., the geocode is on the order of resolution of a street address. A geocode for 'Australia' would return 1 (country level accuracy). Overall, Google distinguishes the following accuracy levels: *0-unknown location*, *1-country*, *2-region* (state, province, prefecture), *3-sub-region* (county, municipality), *4-town* (city, village), *5-post code* (zip code), *6-street*, *7-intersection*, *8-address*, and *9-premise* (building name, property name, shopping center) level accuracy.

Summary

Figure 1 summarizes the four classification models investigated in this research, by matching their respective space categories to the categories introduced in

[1] https://developers.google.com/maps/documentation/geocoding/v2/index

Table 1 Categories of space and the corresponding geographic entities [category names according to Freundschuh and Egenhofer (1997)]

Class	Geographic entities
Table-top objects	Small manipulable objects (pen, book)
Larger objects	Furniture (desk, bed, bench), small vehicles (bike), or small natural features (tree)
Rooms	Location within a building, or within parts belonging to it (office, floor), or medium vehicles (car, boat, bus)
Buildings[a]	Location of a house or building (residential, commercial), e.g., house number or building name (engineering dept, Spencer street station, at work), home, street intersections, large vehicles (train, ferry)
Neighbourhoods	Institution, public space or street level, this includes infrastructure (railway, tramline), public spaces (golf course, university, cemetery, hospital, mall), natural features (port, bay, lake, hill, park, reserve, paddock)
Towns	Rural locality, suburb, post code areas (Brunswick, South Melbourne), or categorical information (central business district, downtown, city center)
Cities	Town or city level, and metropolitan areas (Canberra, Melbourne)
States	References beyond city level up to state level (Victoria)
Countries	References on country level, including highways, national parks, rivers
Continents	Continent level (Europe, Australia)
World	World level (Planet Earth)

[a] The distinction between *houses* and *buildings* made by Freundschuh and Egenhofer (1997) is not made here

(Freundschuh and Egenhofer 1997). Different geographic entities are assigned to these different categories according to Table 1.

4 Comparing Different Classification Models

The presented classification models were compared using a corpus of place descriptions collected through a mobile game (Winter et al. 2011). Participants in the game were asked to first confirm their GPS self-localization, and then to submit a textual description of their location to answer the question 'tell us where you are'. Apart from these tasks and knowing they could win a gift voucher, no further context was given to the participants. 2,221 geocoded place descriptions of Australian locations were collected.

From these place descriptions, a subset of 722 place descriptions contains at least two spatial cues. These were analyzed regarding their hierarchical structure according to the different patterns identified in Sect. 3.1. Each of the four classification models was applied to each place description.

In the classification, all spatial relationships were ignored. A single exception to this has been made in classifying references at building level for Montello's vista and environmental space. References at building level are classified as *environmental*

space if the person is inside, taken either from prepositions such as *in*, *inside*, *at* (cf. Vasardani et al. 2012), or from the lack of prepositions (e.g., addresses). It is also classified *environmental space* if the person is outside, but uses a preposition synonymous to *near*. In all other cases, for example, in presence of prepositions such as *in front of* or *opposite*, a reference at building level will be classified as *vista space*.

Generally, references to multitudes of objects (e.g., apartments) are classified at their next coarser granularity level. For example, in some classification scheme 'apartment' may be classified as room level, but 'apartments' as building level. The finest and the coarsest granularity level in each classification schema are collectives of everything at and below, or everything at and beyond this level of granularity. For example, Google's classification scheme does not provide a granularity level below *premises*, so in this scheme everything smaller than a premise (e.g., an apartment) will be classified on this level.

As stated in Sect. 3.1, place descriptions may exhibit a strictly hierarchical, partially hierarchical, flat, or unordered structure. Applying the different classification schemes to the place descriptions will reveal how these schemes may result in different structures, i.e., how well they pick up variations in spatial granularity, or produce gaps in these structures. For example, 'I am at Union House, located in the University of Melbourne in Parkville, Melbourne' would result in a flat structure when applying Montello's classification, because all four references would be classified on an *environmental space* granularity level. Using Kolar's classification, the same description would be partially hierarchical (without gaps), classifying the Union House and the University of Melbourne as *houses and neighborhood space*, and Parkville and Melbourne as *city-hinterland space* granularity. The classification of Richter would identify a strictly hierarchical structure with a sequential order of the four levels *building* (Union House), *street* (the University of Melbourne), *district* (Parkville), and *city* (Melbourne). Likewise, Goolge's scheme results in three levels of granularity: *premise* (Union House), *street* (the University of Melbourne), and *town* (Parkville and Melbourne). With more than one cue on the same level of granularity, the latter structure is partially hierarchical.

5 Results

The text length of a place description in the subset varies between nine and 586 characters, and is 55 characters on average. The average number of cues (NPs) varies between two and 20 and is 2.9 on average. In total 2,071 NPs have been classified using all four classification schemes.

None of the place descriptions contains a reference to a table-top object. Likewise, none of the participants referred to an object on *world* level. Accordingly, only three of Montello's levels get used in the classification (namely *vista*, *environmental*, *geographic*); with Kolar's scheme only *living/working*, *neighborhood*, *city/hinterland*, and *regional/national* get used.

Using Montello's classification, most entities are on *environmental space* granularity (1,722, or 83 %), while *vista space* and *geographic space* levels only make up for 12 and 5 % of the NP respectively. Kolar's classification has the *neighborhood* level as predominant level, with 1,465 (71 %) of the NPs on this level; 395 (19 %) are on *city/hinterland* level, and 5 % each on *living/working* and *regional/national* level. The classification by Richter yields 16 NPs (1 %) in *furniture* level, 95 (5 %) in *room* level, 620 (30 %) in *building* level, 845 (41 %) in *street*, 275 (13 %) in *district*, 120 (6 %) in *city*, and 100 (5 %) in *country* level. Google's classification results in the following distribution: 568 (27 %) in *premise* level, 121 (6 %) in *address* level, 42 (2 %) in *intersection* level, 845 (41 %) in *street* level, 19 (1 %) *post code* level, 376 (18 %) *town* level, 29 (1 %) in *region* level, 71 (3 %) in *country* level. The granularity level *sub-region* for counties or municipalities was not used in the subset of place descriptions.

Table 2 presents the 722 place descriptions for which hierarchic (strict or partial), flat or unordered patterns have been identified for the different classification schemes. The numbers in brackets indicate the number of place descriptions that contain gaps.

Kolars' and Montello's classifications both result in a large number of flat patterns, 341 (47 %) patterns in Kolars' classification and 459 (64 %) in Montello's, respectively. On the other hand, both Montello's and Kollar's classifications only result in a few gaps, while both Richter's and Google's schemes have a significant number of gaps. However, in case of Google's scheme, the classes *premise*, *intersection*, and *address* do not truly form a (sequential) hierarchical structure; stating an address as place description, for example, essentially excludes also stating a street intersection. And a postcode actually provides the same information as the name of a town and, thus, can be considered optional (i.e., it is on the same granularity level as *town*). Taking this into account, the column Google* in Table 2 shows more realistic results for Google's classification scheme. Most notable, 80 % of the previous gaps disappear; on the other hand there are slightly fewer strictly, and more partially hierarchical and flat structures. In the following Google* will be used.

Most gaps only skip one or two levels in each of the classification schemata. Since Montello's scheme only covers three levels, only one-level gaps appear here. These gaps will also appear when applying the other schemata, because all other

Table 2 Comparison of hierarchical structures in different models

Hierarchy	Kolars	(gaps)	Montello	(gaps)	Richter	(gaps)	Google	(gaps)	Google*	(gaps)
Strict	181	(26)	91	(7)	386	(125)	381	(368)	354	(76)
Partial	130	(9)	134		128	(43)	123	(112)	135	(33)
Flat	341		459		99		111		127	
Unordered	70	(5)	38		109	(23)	107	(101)	106	(10)
Total	722	(40)	722	(7)	722	(191)	722	(581)	722	(119)

*This merges premise, intersection, address to one class, and *post code* and *town* to another, and excludes the class *sub-region* for counties and municipalities

schemata distinguish between finer levels of granularity. There are seven place descriptions that contain one gap in Montello's *environmental space*. For Kolar's scheme, 33 of the place descriptions with gaps only skip one level (*neighborhood* or *city/hinterland space*); two of them skip two levels. The other two schemata have some place descriptions that skip over more than two levels, i.e., up to three (Google*) or four (Richter).

In some cases these gaps may also not be sequentially linked. For example, using the classification by Richter, 'on the skybus to the airport, entering tullamarine fwy' contains a reference on *room* level (the Skybus), on *street* level (airport), and on *country* level (Tullamarine Freeway), skipping *building*, *suburb*, and *city* level. In Montello's classification this description would be strictly hierarchical and sequentially linked with references located in *vista space* (the Skybus), *environmental space* (airport), and *geographical space* (Tullamarine Freeway).

In Kolar's classification 31 of 37 (or 84 %) of gaps are located on *city/hinterland* level ('just off the burwood highway at mcdonalds', 'billabong in the national park', or 'at the royal park opposite the princess highway'), 6 (16 %) on *neighborhood space* ('up at eildon this weekend on the drag boat'). As mentioned before, in Montello's classification seven place descriptions skip the *environmental space* (e.g. 'traveling down the nepean highway in the car'). In Richter's classification, from 244 gaps in hierarchical structures nine (or 4 %) are on *room* level ('in bed at home'), 25 (10 %) on *building* level, 71 (29 %) on *street* level ('in wallan outside coles'), 86 (35 %) on *district* level ('melbourne ligon street'), and 53 (22 %) on *city* level ('241 royal parade parkville vic 3,052'). Finally, the Google* classification yields 69 (or 45 %) on *street* level, 31 (21 %) on *postcode/town* level, and 51 (34 %) on *sub-region/region* level.

6 Discussion

Overall, the presented results support our hypothesis: Different classification schemes yield different results in their identification of hierarchical structures. They also differ in the identification of deviations from regular, sequential hierarchical structures. Montello's or Kolars' classifications, which distinguish fewer classes, tend to produce more *flat* structures than those by Richter or Google. However, the latter two result in many more gaps that appear in the hierarchical structures.

In some more detail, the results show an interplay between the number of categories used in a classification scheme, the distinctions they can pick up, and the deviations that appear. Some of the gaps in one scheme disappear by applying another. For example, applying Richter's classification scheme to 'Under the tree at marinda park' would skip *room* and *building* levels (trees are classified on *furniture* level, while parks are on *street* level), whereas applying Montello's classification scheme would result in only two granularity levels without a gap

(classifying a tree on *vista space* and a park on *environmental space*). Some descriptions that are considered flat in other classification schemes, are classified as hierarchical in Montello's classification due to the special consideration of buildings regarding their categorization into *vista* or *environmental space* granularity. For example 'tram stop near myer', 'i'm in front of ella bache near the foodcourt, or 'In Gopal's restaurant, diagonally opposite of Melbourne City Hall would contain both references on *vista* ('tram stop', 'in front of ella bache', 'opposite or Melbourne City Hall') and also *environmental space* ('myer', 'near the footcourt', 'in Gopal's restaurant), whereas Richter's scheme would consider all these references to be on *building* level.

In general, the application of classification schemes to place descriptions (or any spatial description) requires to assign geographic entities to specific granularity levels. This may introduce some biases and may lead to results that are not always correct. For example, 'Melbourne' may be categorized to be on *city* level, however, the term 'Melbourne' is ambiguous, as it may refer to the suburb Melbourne, the 'City of Melbourne', which is the local government area incorporating the city center and a number of inner-city suburbs, or the region 'Greater Melbourne', which comprises of all suburbs that form the metropolitan area 'Melbourne'.

Other terms, such as 'home', are underspecified regarding the geographic area they refer to. It was classified on *building* level in Richter's classification scheme, but it could also refer to a city or country, depending on the context. And there are types of geographic entities which instances may be of significantly different scales, such as islands, rivers or highways. These would require a more flexible, case-based categorization. The same holds for businesses, such as cafes or restaurants, which sometimes may be part of a larger building (being on granularity level *room*), and sometimes occupying a whole building. Implementing such flexible categorization would avoid some of the gaps that emerged in the presented experiment.

Still, in the end there are deviations that cannot be explained just by the particularities of the respective classification schemes. There are 87 place descriptions that exhibit a flat structure regardless of the chosen classification scheme. These include locomotion descriptions (e.g., 'walking down greeves street to spring street'), and descriptions that just mention multiple references on the same granularity level (referring to several geographic entities of the same type), such as 'between melville rd and reynolds pde'. Furthermore seven place descriptions contain gaps regardless of the classification approach. Descriptions such as 'a loud street intersection, just before crossing the yarra', 'travelling down the napean highway in the car', 'yarra river sitting on the docks', 'whale rock, tidal river' contain all a gap in Montello's *environmental space*, and thus, as well when applying the other schemes.

7 Conclusion

We have investigated several classification schemes to characterize the levels of spatial granularity in place descriptions. These schemes were applied to human place descriptions to characterize their hierarchical structures. Place descriptions were collected through a mobile game a largely underspecified context, resulting in a wide range of different descriptions being collected. The aim of the paper was to test the hypothesis that the identification of hierarchical structures in place descriptions depends on the chosen classification schema. The results show support for the hypothesis. Most of the deviations from hierarchical structures can be related to the respective classification. However, a remaining 10 % can not be explained by the applied schemes, such as flat structures where people seem to employ hierarchies of salience, or locomotion descriptions.

We argued that too few categories in a scheme prevent from making relevant distinctions, and too many categories could exacerbate cognitive representation and reasoning. Applied to place descriptions, a balance between enough granularity levels to pick up these structures, and few enough levels to avoid artificial gaps is desirable. In this respect the classification schemata of Richter and Google behave better.

Studies of this kind will be context-dependent—place descriptions of the location of people will show different expectations to a classification scheme than place descriptions of geological faults line or of a fork on a table. However, in this chapter we have compared schemata all designed for the particular purpose. We have found strong evidence to use Richter's (or alternatively Google's) scheme for complex place descriptions at human scale.

Acknowledgments This work was funded by the Australian Research Council under its Linkage Scheme (LP100200199).

References

Bittner T, Smith B (2003) A theory of granular partitions. In: Duckham M, Goodchild MF, Worboys M (eds) Foundations of geographic information science. Taylor & Francis, London, pp 117–151

Couclelis H, Gale N (1986) Space and spaces. Geografiska Annaler Series B Human Geography, pp 1–12

Dale R (1992) Generating referring expressions: constructing descriptions in a domain of objects and processes. MIT Press, Cambridge

Freundschuh S, Egenhofer M (1997) Human conceptions of spaces: Implications for geographic information systems. Trans GIS 2(4):361–375

Grice HP (1975) Logic and conversation. In: Cole P, Morgan JL (eds) Speech acts, syntax and semantics, vol 3. Academic Press, New York, pp 41–58

Hirtle SC, Jonides J (1985) Evidence of hierarchies in cognitive maps. Mem Cogn 13(3):208–217

Hobbs JR (1985) Granularity. In: Joshi AK (ed) Proceedings of the 9th international joint conference on artificial intelligence, Morgan Kaufmann, Los Angeles, pp 432–435

Ishikawa T, Montello DR (2006) Spatial knowledge acquisition from direct experience in the environment: Individual differences in the development of metric knowledge and the integration of separately learned places. Cogn Psychol 52(2):93–129
Keet C (2006) A taxonomy of types of granularity. In: IEEE conference in granular computing (GrC2006), pp 10–12
Kelleher J, Kruijff GJM (2006) Incremental generation of spatial referring expressions in situated dialog. In: 21st international conference on computational linguistics, association for computational linguistics, Sydney, pp 1041–1048
Kolars J, Nystuen J, Bell D (1975) Physical geography: environment and man. McGraw-Hill, New York
Kuipers B (1978) Modeling spatial knowledge. Cogn Sci 2(2):129–153
Lynch K (1960) The image of the city. The MIT Press, Cambridge
Mani I, Doran C, Harris D, Hitzeman J, Quimby R, Richer J, Wellner B, Mardis S, Clancy S (2010) SpatialML: annotation scheme, resources, and evaluation. Lang Res Eval 44(3): 263–280
Montello D (1993) Scale and multiple psychologies of space. In: Frank A, Campari I (eds) Spatial information theory, lecture notes in computer science, vol 716. Springer, Berlin, pp 312–321
Plumert JM, Spalding TL, Nichols-Whitehead P (2001) Preferences for ascending and descending hierarchical organization in spatial communication. Mem Cogn 29(2):274–284
Raubal M, Winter S (2002) Enriching wayfinding instructions with local landmarks. In: Egenhofer MJ, Mark DM (eds) Geographic information science. Springer, Berlin, Lecture notes in computer science, vol 2478, pp 243–259
Richter D, Vasardani M, Stirling L, Richter KF, Winter S (2013) Zooming in—zooming out: hierarchies in place descriptions. In: Krisp JM (ed) Progress in location-based services, series, lecture notes in geoinformation and cartography, vol 25. Springer, Berlin, pp 339–355
Rosch E, Mervis C, Gray W, Johnson D, Boyes-Braem P (1976) Basic objects in natural categories. Cogn Psychol 8(3):382–439
Sadalla EK, Burroughs WJ, Staplin LJ (1980) Reference points in spatial cognition. J Exp Psychol: Hum Learn Mem 6(5):516–528
Shanon B (1979) Where questions. In: 17th annual meeting of the association for computational linguistics. ACL, University of California at San Diego, La Jolla
Siegel AW, White SH (1975) The development of spatial representations of large-scale environments. In: Reese HW (ed) Advances in child development and behavior, vol 10. Academic Press, New York, pp 9–55
Smith B, Mark DM (2001) Geographical categories: an ontological investigation. Int J Geogr Inf Sci 15(7):591–612
Sorrows ME, Hirtle SC (1999) The nature of landmarks for real and electronic spaces. In: Freksa C, Mark DM (eds) Spatial information theory. Lecture notes in computer science, vol 1661. Springer, Berlin, pp 37–50
Sperber D, Wilson D (1986) Relevance-communication and cognition. Basil Blackwell, Oxford
Stevens A, Coupe P (1978) Distortions in judged spatial relations. Cogn Psychol 10(4):422–437
Talmy L (1983) How language structures space. In: Herbert L, Pick J, Acredolo LP (eds) Spatial orientation: theory, research, and application. Plenum Press, New York, p 225
Tenbrink T, Winter S (2009) Variable granularity in route directions. Spat Cogn Comput 9(1):64–93
Timpf S (1998) Hierarchical structures in map series. Phd thesis, Technical University Vienna, Vienna
Tomko M, Winter S (2009) Pragmatic construction of destination descriptions for urban environments. Spat Cogn Comput 9(1):1–29
Vasardani M, Winter S, Richter KF, Stirling L, Richter D (2012) Spatial interpretations of preposition “at”. First ACM SIGSPATIAL international workshop on crowdsourced and volunteered geographic information (GEOCROWD)’12, November 6, 2012 Redondo Beach

Winter S, Richter K, Baldwin T, Cavedon L, Stirling L, Duckham M, Kealy A, Rajabifard A (2011) Location-based mobile games for spatial knowledge acquisition. In: Cognitive engineering for mobile GIS. Workshop at COSIT 2011, Belfast, pp 1–8

Zubin D (1989) Natural language understanding and reference frames. In: Mark D, Frank A, Egenhofer M, Freundschuh S, McGranaghan M, White RM (eds) Languages of Spatial Relations: Initiative 2 Specialist Meeting Report Technical Paper, pp 13–16

Error-Aware Spatio-Temporal Aggregation in the Model Web

Christoph Stasch, Edzer Pebesma, Benedikt Graeler and Lydia Gerharz

Abstract Spatio-temporal aggregation of observed or predicted values for environmental phenomena is needed for fusing sensor data or coupling sensors and environmental models. However, estimates from sensors or environmental models can never represent our world precisely and are subject to errors. Hence, there is uncertainty in the estimates that needs to be considered in environmental model workflows. This chapter presents an approach for an error-aware spatio-temporal aggregation in the Web, where probabilistic uncertainties are used within a Monte Carlo simulation. The approach is applied in a Web-based model chain that provides uncertain crop yield predictions on field parcel level that are aggregated to larger regions.

1 Introduction

The Model Web envisions discovery and access of environmental observations and models using the internet as mediating platform (Geller and Turner 2007; Nativi et al. 2012). Where environmental models, even those of same domains, currently exist in parallel and do not benefit from each other, the Model Web could ease the coupling of such models. To achieve this vision, the environmental

C. Stasch (✉) · E. Pebesma · B. Graeler · L. Gerharz
Institute for Geoinformatics, University of Münster, Münster, Germany
e-mail: staschc@uni-muenster.de

E. Pebesma
e-mail: edzer.pebesma@uni-muenster.de

B. Graeler
e-mail: ben.graeler@uni-muenster.de

L. Gerharz
e-mail: gerharz@uni-muenster.de

D. Vandenbroucke et al. (eds.), *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography,
DOI: 10.1007/978-3-319-00615-4_12,

observations and models should be exposed via publicly available standardized Web service interfaces such as those defined by the Open Geospatial Consortium (Maue et al. 2011). Spatio-temporal aggregation (Jeong et al. 2004; Vega Lopez et al. 2005; Stasch et al. 2012) is needed in the Model Web for two reasons: The spatio-temporal resolution of the sensor output might not match the resolution required by a model and, when chaining environmental models, the resolution of the output of one model might differ from the resolution required by another model.

However, environmental observations and models are subject to error due to the observation methods or due to simplified representation of real world phenomena by models. As a result, there is uncertainty in observations and model results that needs to be considered (Heuvelink 1998). The UncertWeb project aims to provide tools for managing and communicating uncertainties in the Model Web (Bastin et al. 2013). Such an uncertainty-enabled Model Web requires an error-aware spatio-temporal aggregation that explicitly considers uncertainties in input data and allows to propagate uncertainties to the aggregated estimates. As uncertainty can be reduced by aggregation in model workflows, an error-aware spatio-temporal aggregation also provides means to adjust the uncertainty, for example by averaging out some of the variability in the data.

The core contribution of this chapter is an approach for an error-aware spatio-temporal aggregation in the Model Web relying on open standards. A probabilistic approach is chosen for representing the uncertainties and a Monte Carlo simulation is used to propagate uncertainties in aggregation processes. To provide error-aware aggregation processes in the Model Web, a common Web service interface is defined and implemented in a Web-based model workflow for predicting land-use and crop yield response to climatic and economic change in England (Jones et al. 2012).

The remainder of the chapter is structured as follows: Sect. 2 provides an overview of error-aware spatio-temporal aggregation. Afterwards, the approach for Web-based error-aware aggregation is presented in Sect. 3. The application of the approach in a case study for aggregating yield predictions is described in Sect. 4, followed by the presentation of results in Sect. 5 and a discussion of the approach in Sect. 6. In the last section, conclusions are drawn and next steps are presented.

2 Error Aware Spatio-Temporal Aggregation

An aggregation process computes a single value, an aggregate, for a group of attribute values using an aggregation function. The values are grouped by partitioning predicates. Spatio-temporal aggregation groups spatio-temporal features by spatial and/or temporal predicates and applies aggregation functions to those features in order to change the spatio-temporal resolution of datasets (Jeong et al. 2004; Vega Lopez et al. 2005; Stasch et al. 2012). An example of a spatio-temporal aggregation process is shown in Fig. 1a. Temperature observations gathered hourly at monitoring stations are aggregated temporally to daily maxima

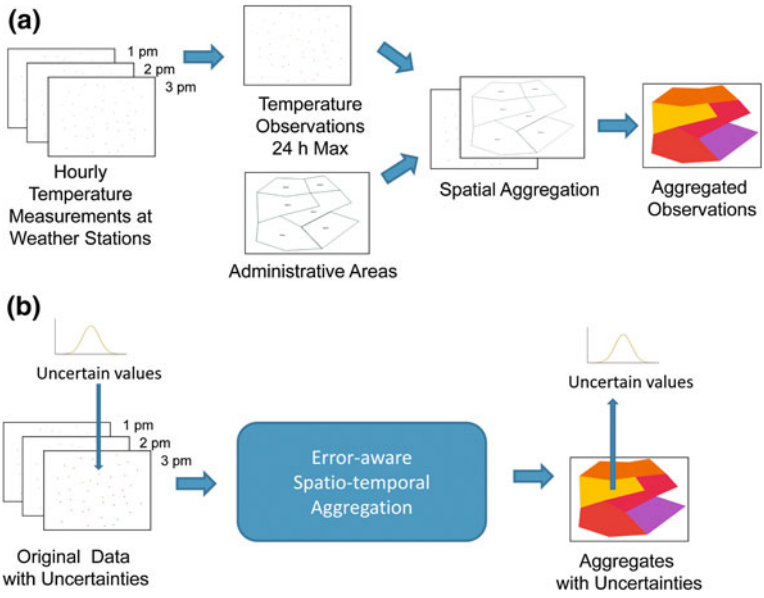


Fig. 1 Illustration of (error-aware) spatio-temporal aggregation (modified from Stasch et al. 2011). **a** Spatio-temporal aggregation. **b** Error-aware spatio-temporal aggregation

and spatially to means of spatial regions. The temporal grouping predicates consist of days, the temporal aggregation function is MAX, the spatial grouping predicates are spatial polygons (for instance, administrative boundaries) and the spatial aggregation function is MEAN. The grouping predicates as well as the aggregation functions might require additional specific input parameters to the aggregation process. For example, the spatial grouping predicates need to be defined by a set of polygons.

There are two possible sources of errors in spatio-temporal aggregation: (1) The input data of an aggregation process might be uncertain and cause errors in the aggregates or (2) the aggregation functions used for computing the values introduce uncertainties. An example of (2) is the aggregation of spatio-temporally distributed information on rainfall and soil moisture over a catchment during and after a rainfall event, which is aggregated to river discharge in order to predict floods. In this case, the aggregation function is a distributed hydrological model that can only approximate the true catchment response to a rainfall event, due to a simplified representation of the true aggregation process.

In this work, we focus on uncertainties in input data (1) that are aggregated using simple aggregation functions, e.g. mean, sum, or max, as illustrated in Fig. 1b. In case of spatio-temporal data, the uncertain input may be provided as a spatio-temporal random field $Y(q)$, where q is a spatio-temporal location. This random field is usually assumed to be normally distributed, i.e. $Y(q) \sim N(\mu(q), \Sigma)$, with $\mu(q)$ the mean vector for the locations and Σ the covariance matrix. In case an

aggregation function f is a non-linear function, e.g. computation of the maximum value, the expected value of the aggregates will typically differ from the aggregated expected values: $E[f(Y(q))] \neq f(E[Y(q)])$. Hence, aggregating the parameters of the input distributions in order to compute the probability distributions for the aggregates may introduce a bias. To avoid this, a Monte Carlo simulation approach is adopted to propagate the uncertainties in the aggregates (Heuvelink and Pebesma 1999). In case the inputs are provided as probability distribution functions (PDF) for each measurement value, realisations are generated from the input distributions and the aggregation process is run for each set of realisations resulting in a set of realisations for each output region that in turn approximates the target PDF.

The pseudocode for applying a Monte Carlo simulation is shown in Algorithm 1. The function y_i returns the i-th realisation value of a spatio-temporal random field $Y(q)$ at spatio-temporal location q within the target region R representing the grouping predicate. The realisation values per region are then used by the aggregation function f as inputs to compute the aggregate, for example, computing the sum. The actual aggregates for each spatio-temporal region R and i-th realisation r_i are returned by $\check{y}$. As an option, instead of returning all realisations of aggregates for each spatio-temporal region, summary statistics for the realisations, such as mean or the 95 %, may be computed by a function g as illustrated in Algorithm 2.

Algorithm 1 Spatio-temporal aggregation for multiple realisations

1: **for all** realisations r_i, i=1,...,n **do**
2: **for all** spatio-temporal regions R_j, j=1,...,m **do**
3: $\check{y}(R_j, r_i) = f(y_i(q_1), y_i(q_2), ..., y_i(q_p))$ *with* $\{q_1, ..., q_p\} \in R_j$
4: **end for**
5: **end for**

Algorithm 2 Aggregation with statistics computed from spatio-temporally aggregated realisations

1: **for all** spatio-temporal regions R_j, j=1,...,m **do**
2: $\bar{y}(R_j) = g(\check{y}(R_j, r_i), \check{y}(R_j, r_2), ..., \check{y}(R_j, r_n))$
3: **end for**

Besides allowing to propagate uncertainties with non-linear aggregation functions, the Monte Carlo simulation approach also allows for more flexibility than an analytical approach that is usually bound to a specific aggregation process (Heuvelink 1998). It also allows to consider input uncertainties for already existing deterministic aggregation processes without the need to change the underlying models of the aggregation processes. Spatio-temporal aggregation also provides a mean to control the uncertainty in model workflows: Given that there is variability in the data within the aggregation regions, aggregating the data to the mean of an

area, may reduce variability. However, the degree of variability reduction depends on the spatio-temporal autocorrelation (Gerharz and Pebesma 2012). The use case shown below illustrates that it also depends on the aggregation function used.

3 Error-Aware Aggregation in the Model Web

After introducing the general approach for an error-aware aggregation, the question remains how error-aware aggregation processes can be provided in the Model Web. Firstly, the input data needs to be provided with uncertainties and these need to be encoded in a standardized format (Sect. 3.1). Secondly, a common approach for providing and utilizing aggregation functionality in the Model Web needs to be defined that explicitly considers uncertainties (Sect. 3.2).

3.1 Formats for Spatio-Temporal Data with Uncertainties

In order to enable an error-aware spatio-temporal aggregation, the input data needs to contain uncertainty information. Up to now, if present at all, the uncertainty information is given in proprietary formats hindering a common approach and implementation of an error-aware aggregation. Hence, there is a need to provide common models and encodings for spatio-temporal data explicitly containing uncertainties. The Uncertainty Markup Language (UncertML) (Williams et al. 2009) has been developed as a common model and exchange format for probabilistic uncertainties. It allows to encode uncertainties as distributions, descriptive statistics or as a set of realisations. As UncertML does not explicitly define how to add spatial and temporal references to the uncertainties, there is a need for spatio-temporal models and exchange formats that support uncertainties.

To exchange uncertain spatio-temporal data in the Model Web, two common formats are defined. For vector data, the Uncertainty-enabled Observations & Measurements (U-O&M) format integrates UncertML with Observations & Measurements (O&M) (ISO 2010; Stasch et al. 2012), a common format for spatio-temporal observations and model results. Uncertainty can either be provided as additional metadata or as the result of an observation. U-O&M can be serialized in different formats such as XML, JSON, or plain text, such as comma separated values (csv), and hence be used for exchanging uncertain spatio-temporal data in the Model Web. The O&M format also allows to be used across different spatio-temporal aggregation levels of observations. In this work, we use observations with uncertain results as shown in Fig. 2 in XML format. The uncertainty is encoded as UncertML realisations in the result of the observation. While O&M is well suited for vector-based spatial data, NetCDF is a well-established format for gridded/raster data. NetCDF-U (Bigagli and Nativi 2011) has been defined to encode

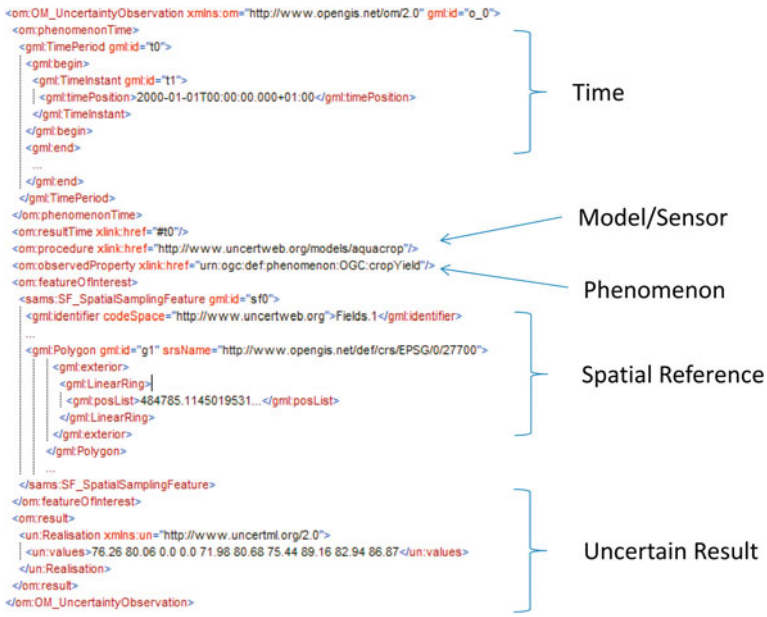


Fig. 2 Encoding of an observation that contains realisations of yield predictions for a field polygon in the year 2000

uncertainties in NetCDF (Domenico 2011) and is utilized in the Spatio-temporal Aggregation Service (STAS) for aggregating uncertain gridded data.

3.2 Error-Aware Spatio-Temporal Aggregation Service

To provide spatio-temporal aggregation functionality in the Model Web, we are extending the STAS that has been introduced by Stasch et. al. as an aggregation service for the Sensor Web (Stasch et al. 2012). As the Sensor Web envisions the tasking of sensors and the exchange of sensor data in the Web (Bröring et al. 2011), the Model Web may be seen as an extension that allows the discovery, access, and execution of environmental models and not just sensors. The overall concept of the STAS for the Model Web is illustrated in Fig. 3. The STAS is defined as a profile of the OGC Web Processing Service (WPS) (Schut 2007) and can be utilized as a mediator that transforms data from one resolution to another. The input data can be provided as output of model services, data sources, or as resources on a Web server. The STAS itself can then be invoked by end-users, model services, or orchestration engines and the aggregated data can in turn be directly published to model or data services or stored as a resource on a Web server.

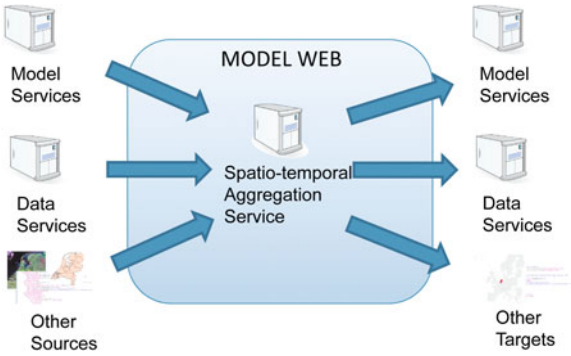


Fig. 3 Role of the spatio-temporal aggregation service in the model web. It acts as a mediator between data and model services in the model web, if an aggregation is required to fit outputs to other inputs. The aggregated data can be published via data services again

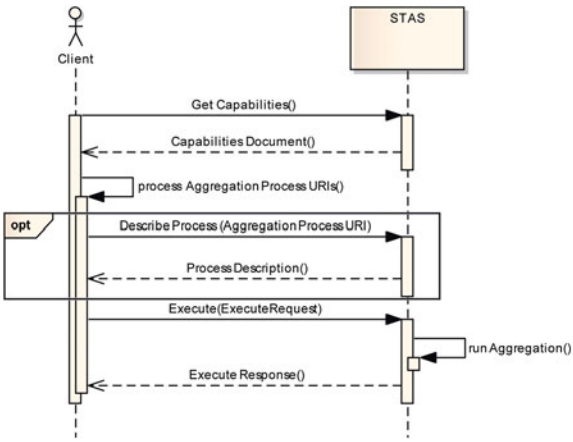


Fig. 4 UML sequence diagram showing the usual interaction pattern between a client and the STAS

The usual interaction pattern between clients and the aggregation service is shown as sequence diagram in the Unified Modeling Language (UML) in Fig. 4. The grouping predicates and aggregation functions of an aggregation process are identified in Unified Resource Identifiers (URI) of the aggregation processes. Therefore, we are utilizing the URI scheme as defined in (Stasch et al. 2012). All URIs of the available aggregation processes are listed in the service description (`Capabilities` document) that can be retrieved by the `GetCapabilities` operation. If a detailed description of a specific aggregation process including all input and output parameters is needed by the client, it can be retrieved using the `DescribeProcess` operation. To actually run an aggregation, the `Execute` operation needs to be invoked by passing an `ExecuteRequest` to the service. The request contains all necessary input parameters such as the input data (or a pointer to the data), parameters of the grouping predicates or of the aggregation functions. After aggregation, the `ExecuteResponse` can directly return the aggregated data in a requested format to the client or pass a reference, in case the aggregated data is inserted in another data service or stored on a server.

Table 1 Common input parameters of aggregation processes provided by the STAS for the model web

Input parameter Name	Cardinality	WPS input Type	Description
Identifier	1	URI	Identifier of the aggregation processes that should be run; defines the grouping predicates and aggregation functions
Variable	0...*	LiteralData	Name of variables (e.g. air temperature) that should be aggregated in case the input data contains several variables
InputData	1	ComplexData	Data that should be aggregated
SpatialFirst	0...1	Boolean	Indicates whether spatial aggregation should be done first (true) or not (false) in case of non-linear aggregation functions for space and/or time
TargetServer	0...1	LiteralData	Endpoint of the server, to which aggregated data should be written
TargetServerType	0...1	LiteralData	Type of server to which the aggregated data should be written

Common input parameters are defined for all aggregation processes in the Model Web as listed in Table 1. Depending on the grouping predicates and aggregation functions, additional parameters can be defined for particular aggregation processes. For example, the process introduced in Sect. 2 requires the additional parameter `FeatureCollection` that contains the polygons for the spatial grouping predicates and `duration` that defines the duration (24 h) for the temporal grouping predicates. The spatial and temporal references of the aggregates are defined by the parameters of the grouping predicates. The result of an aggregation execution not only provides the aggregated data, but also additional provenance information by pointing to an instance of a specific aggregation process description. This aggregation process description includes information about the predicates and aggregation functions used. In addition, the aggregated data points to the original data from which the aggregates have been computed.

A Monte Carlo simulation approach is used to propagate uncertainties in the STAS. The interaction pattern of an error-aware aggregation process in the STAS is shown in Fig. 5. Clients can indicate, whether a Monte Carlo simulation for the aggregation process should be run or not by passing the optional `NumberOf-Realisations` in an `Execute` request. If this parameter is present, the uncertain data has to be provided in the `InputData` parameter of the `Execute` request either as PDFs or as realisations. In case the uncertain inputs are PDFs, realisations are taken from the PDFs using the Uncertainty Transformation Service (UTS). The UTS is an external Web service for transforming uncertainties from one representation into another (Pross et al. 2012). For example, the UTS allows to convert from a normal distribution to a set of realisations. Then, for each Monte Carlo realisation, the aggregation is executed using an aggregation engine, e.g. the R software (R Development Core Team 2011), resulting in a set of samples of

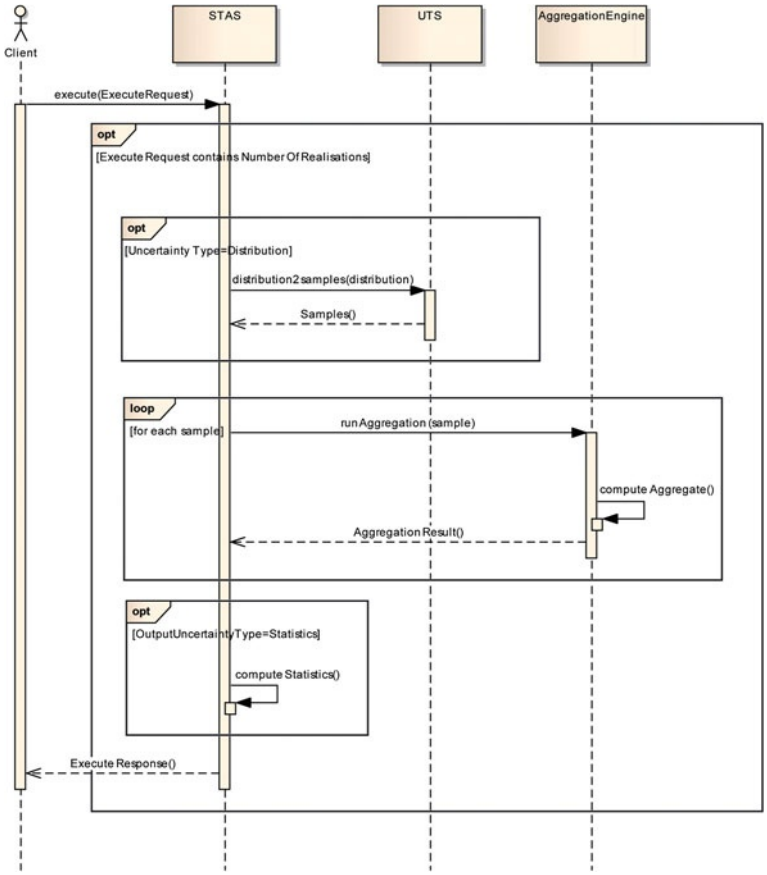


Fig. 5 UML sequence diagram showing the interaction pattern of an error-aware aggregation process between client, spatio-temporal aggregation service (STAS), uncertainty transformation service (UTS) and an aggregation engine

aggregated data. From the Web service, the aggregated data can either be retrieved as the full set of spatio-temporally aggregated realisations or as summary statistics of the realisations. Table 2 shows the additional input parameters for all error-aware aggregation processes. The `NumberOfRealisations` parameter is needed by all processes and defines the number of Monte Carlo simulation runs. In addition, the `OutputUncertaintyType` can define additional summary statistics that should be returned for the set of realisations of aggregated data. The `OutputUncertaintyType` parameter has to use the identifiers (URLs) of UncertML for the different statistics. The `InputData` parameter that is inherited from the common input paramters of the aggregation processes (Table 1) has the restriction to contain either realisations or distributions defined for UncertML in its inputs. In order to avoid errors, when clients are sending other uncertainty types or data without uncertainties to the service, the additional metadata element

Table 2 Additional parameters of error-aware aggregation processes

Input parameter name	Cardinality	WPS input type	Description
NumberOfRealisations	0..1	LiteralData	Number of Monte Carlo simulation runs
OutputUncertaintyType	0..*	LiteralData	The types of uncertainties as defined by UncertML in which the aggregated outputs should be provided. Per default, the aggregated data is provided as realisations, but also descriptive statistics of the realisations such as mean or standard deviation can be requested

`variable-uncertainty-types`, defined in the metadata conventions of the UncertWeb project, is nested in the `ows:Metadata` element of the `InputData` parameter in an aggregation process description is defined. The `variable-uncertainty-types` shall contain URLs of the UncertML dictionary for the supported uncertainty types. Besides the tag `variable-uncertainty-types`, several additional metadata tags, for example for the resolution of raster data, have been defined in the UncertWeb project and can also be used with other Web services than those defined by the OGC as described in Jones et al. (2012).

4 Case Study

This section describes a case study in which our approach is applied. The section starts with a description of the application scenario (Sect. 4.1) followed by a description of the Web service implementation (Sect. 4.2).

4.1 Application Scenario

The Food and Environment Research Agency[1] of the UK has established an environmental model workflow that is used to estimate land-use and crop yield responses to climatic and economic change. This model workflow has been extended to consider uncertainties and has been deployed via Web services in the internet (Jones et al. 2012). The model workflow estimates yields per field parcel for certain crop types, e.g. wheat or potatoes. The uncertainty in the yield predictions is propagated by running the yield model a number of times resulting in a number of yield realisations for each field per year.

[1] http://www.fera.defra.gov.uk/

Fig. 6 Excerpt from an overlay of fields and regions. Fields are shown in *white colour* and regions are shown in *grey colours*. The fields may be contained in several regions and some parts of the regions are not covered by fields

For privacy reasons and to provide an overview on a larger scale, the field estimates need to be aggregated to spatial regions. Thereby, the fields might be contained in several regions and some places in the regions might not be a field, e.g. urban areas or forests, as shown in Fig. 6.

Following our definition of spatio-temporal aggregation, the spatial grouping predicates are spatial regions and the spatial aggregation function f is defined as follows:

$$f(R) = \sum_{i=1}^{N_R} (x_i \times A_{iR}) \tag{1}$$

with R a spatial region over which we aggregate, N_R the number of field parcels intersecting R, x_i the estimated yield per hectar, and A_{iR} the spatial intersection area of field parcel i and region R. The yield prediction per hectar is multiplied with the area that intersects a region and for each region, the results are summed resulting in the total yield for each region. The data used in the case study consists of one thousand realisations of yield for 24 field parcels for the year 2012 that are aggregated to eight regions. For privacy reasons, not the real field parcel data are used, but artificially created parcels and regions were generated by a random process. As the data does only represent the year 2012, it does not need to be grouped temporally before executing the aggregation, though this is supported by the service implementation.

4.2 Web Service Implementation

The STAS is implemented as an extension of the 52° North WPS.[2] Therefore, an `AbstractAggregationProcess` class has been implemented that provides utility methods for accessing the common parameters of all aggregation processes. Several aggregation processes realizing the `AbstractAggregationProcesses` are implemented for vector and/or raster data. Three classes have been defined for the implementation of our error-aware aggregation approach as shown in Fig. 7. The `AbstractUncertainAggregationProcess` provides utility methods for the common inputs of error-aware aggregation processes and defines an additional abstract method `runMonteCarlo` that needs to be implemented by every subclass.[3]

To enable an aggregation as described in the previous Sect. 4.1, the class `Polygon2PolygonWeightedSum` has been implemented that extends the `AbstractUncertainAggregationProcess`. In addition, a `Polygon2PolygonMean` class is available to compute the arithmetic mean of the yields per region. For implementing the aggregation, a hash-based approach as described in Jeong et al. (2004) has been implemented in Java using the JTS library as follows: First, the input data is grouped by time and for each time, the input collection is stored in a hash map. Afterwards, the spatial grouping predicate (spatial intersection of the input features and the target regions) is checked. If there is an intersection, the realisations of intersecting features are cached with the intersection areas as weights for each target region using a hash map again. Then, the aggregation is executed for each target region several times until the number of realisations is reached and, depending on the requested uncertainty types, different statistics of the realisations per target region are computed.

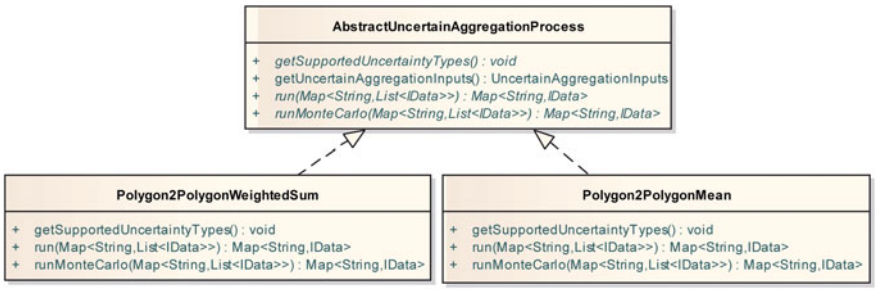


Fig. 7 UML class diagram of the two additional classes that are implemented for the error-aware aggregation service

[2] http://52north.org/communities/geoprocessing/wps/

[3] The classes are provided under the GNU General Public Licence (GPL) v2 licence as part of the STAS implementation at https://svn.52north.org/svn/geostatistics/main/uncertweb/stas/trunk

The uncertain yield predictions are provided in the U-O&M format with UncertML `ContinuousRealisations` as observation values for each field. The STAS runs the aggregation for each realisation of yield values per fields and hence produces 1,000 realisations for the aggregated yields per region. These are then returned again in the U-O&M format. The request/response encoding is automatically done by an additional component of the WPS framework developed in this work to support uncertain spatio-temporal inputs and outputs[4] (Sect. 3.1).

5 Results

Providing error-aware aggregation functionality in a standardized Web service allows for exchanging the aggregation methods in a flexible way in Web-based model workflows. In addition, the extension for Monte Carlo simulation allows for propagating the uncertainties during the aggregation.

Aggregation processes have been executed for computing the weighted sums and the mean of the yield values per region. Figure 8 shows the visualisation of the aggregated estimates in the UncertWeb visualisation client (Gerharz et al. 2012).

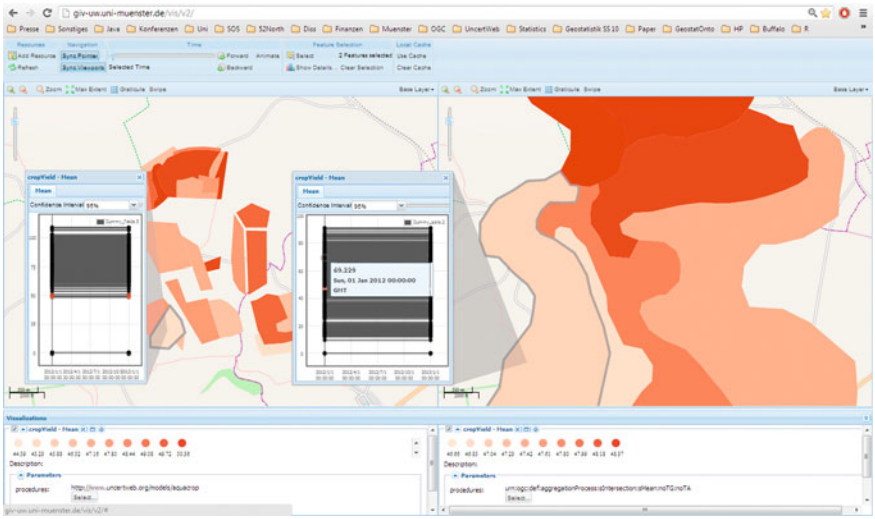


Fig. 8 Screenshot of the UncertWeb visualisation client visualising the non-aggregated (*left*) and aggregated yield estimates (*right*). Realisations of yield estimates can be visualised for specific fields and regions

[4] The input and output extension of the 52N WPS framework is acccessible as a separate package at https://svn.52north.org/svn/geostatistics/main/uncertweb/52n-wps-io-uncertweb/trunk and can also be used by other WPS implementations for uncertain data.

Table 3 Descriptive statistics of original data and aggregation results computed with mean as aggregation function

Description	Mean of realisation means	Standard deviations	Coefficient of variation
Non-aggregated yield estimates [in tonnes per hectar]	5.59	3.77	0.67
Aggregated yield estimates [in tonnes per hectar]	5.56	1.83	0.33

Table 4 Descriptive statistics of original data and aggregation results computed with weighted sum as aggregation function

Description	Mean of realisation means	Standard deviations	Coefficient of variation
Non-aggregated yield estimates [in tonnes]	44.04	29.69	0.67
Aggregated yield estimates [in tonnes]	155.54	49.94	0.32

For comparison, the non-aggregated as well as the aggregated estimates can be visualized. In this example, the realisation means of the fields as well as of the aggregates are shown in the map. For specific polygons all realisations are visualised in a popup. The information about the process that generated the data is given in the description of the layers. On the right side, for example, the URN shown in the procedures element is the identifier of the aggregation process.

Table 3 shows the descriptive statistics of the aggregation with mean as aggregation function. Firstly, means and variances of the yield realisations have been computed. Then, the means and variances have been derived from the realisation statistics. As expected, the mean yield per hectar is nearly the same for the fields as for the regions. The variability in the data is reduced by the mean aggregation, as the mean standard-deviation is reduced from 3.77 for the fields to 1.83 for the regions.

In order to compare the original field values with the weighted sum of the regions, the original values have been multiplied by the area of each field. While the aggregation to means of regions reduces the variability, the aggregation to weighted sums increases the variability as can be seen in Table 4. However, the coefficient of variation decreases in both cases.

6 Discussion

The approach of a Web based error-aware aggregation offers the following advantages: (1) a common way to communicate uncertain spatio-temporal data in the Web is defined, (2) the approach allows to change the resolution and

uncertainty in the data by aggregating them over space and time, (3) different error-aware aggregation processes can be accessed in a common way in the Web without the need to adopt workflow implementations for each aggregation process, (4) the standardized interface and the data formats allow for an integration in spatial information infrastructures, as they rely on standards used in these infrastructures.

The aggregation functionality provided by the aggregation service may be implemented, for example, in database systems (Jeong et al. 2004; Vega Lopez et al. 2005) or software systems such as R (Pebesma 2012) and then be provided by the aggregation service in the Web. While databases and other technologies usually only offer specific aspects needed in error-aware aggregation processes, the different technologies may be combined within the aggregation service. For instance, though there are databases for probabilistic data (Benjelloun et al. 2006), these databases do not provide spatio-temporal query functionality. An approach using Monte Carlo simulation for query evaluation is described in (Jampani et al. 2008), while they do not tackle the issue of aggregate queries on spatio-temporal entities. With our approach, the different technologies can be combined and provided in the Web via a common interface. In the case study, the input data has been transferred to the STAS. However, in case of big data, it may be more reasonable to tightly couple the aggregation functionality with the data sources as described for a coupling between the Sensor Observation Service and the STAS in our previous work (Stasch et al. 2012, p.117). The STAS interface can also be utilized in this case, but the parameter used for passing the inputs may then only identify data from the data source. This approach still allows to run the different aggregations in the Web.

Though the U-O&M format (Stasch et al. 2012) is defined for exchanging uncertain spatio-temporal vector data and NetCDF-U (Bigagli and Nativi 2011) is used to exchange uncertain spatio-temporal rater data, there is currently no generic standard for (uncertain) spatio-temporal data that is widely adopted. Hence, the (uncertain) spatio-temporal data needs to be converted, before it can be published in the Web and aggregated with the aggregation service. It needs to be explored, how well U-O&M and NetCDF-U map to other formats that are already in use and how uncertainty can be incoporated in such formats.

Another question is to which degree the process that generates the data relates to a sensor or to a model, or whether both concepts should be treated seperately. One would probably agree that a complex aggregation procedure such as a hydrological forecast model that aggregates measurements in space and time would not be considered as a sensor. However, observations such as discharge measurements usually have undergone a modelling procedure (Beven et al. 2012) and simple aggregations (and underlying models) are always part of technical sensors where the aggregation is done on a low abstraction level. Hence, there is still a need to clarify the semantics of the different concepts and, in a second step, to formalize them in order to be used in the Web for semantic interoperability (Sheth et al. 2008; Balazinska et al. 2007).

As our use case is based on realisations, the spatial autocorrelation does not need to be addressed in the Monte Carlo simulation. However, data formats defined in our previous work (Graeler and Stasch 2012), can be utilised to represent spatio-temporal random fields. Web services can then use the formats and provide spatio-temporal sampling procedures that consider the autocorrelation.

The current approach allows for the propagation of quantified uncertainties either provided as realisations or probability distributions. Further investigation is needed to check whether the current approach might be utilized in a scenario with categorical data. Furthermore, our approach requires that uncertainty information is available in the input data. Though there are already methods how to assess the uncertainty in measurements (Taylor 1997), assessing the uncertainty per observation or model output remains a challenging statistical and operational problem. In addition, most sensor data providers do not yet provide uncertainty information with the data. Hence, incentives have to be explored how to motivate data providers to make the uncertainties in their data explicit.

The implementation described in this chapter only includes aggregation processes, though the service interface of the STAS can also be used to provide disaggregation processes as described in (Bierkens et al. 2000). Finally, once there are more error-aware aggregation processes available in the Web that can be easily combined with model and sensor services, the question remains how to find those services and how to indicate that these aggregation services support uncertainty propagation. While we have introduced a description format for aggregation processes in our previous work, it has to be explored how these descriptions may be integrated in common approaches for sensor discovery (Jirka et al. 2009) and model web services (Nativi and Bigagli 2009).

7 Conclusion

This chapter presents an approach for an error-aware aggregation in the Model Web. For error propagation, Monte Carlo simulation is utilized. The uncertainty in the non-aggregated input data is provided as probability distributions, from which samples are taken, or is directly provided as samples. Then, for each sample an aggregation is carried out. Thus, the aggregation output consists of a set of realisations that in turn approximate the probability distribution of the aggregates. To deploy error-aware aggregation functionality in the Web, common data formats for uncertain spatio-temporal vector and raster data, U-O&M and NetCDF-U, are used and a Web service interface is defined as a profile of the OGC Web Processing Service. The application of the approach in a Web based model workflow for estimating crop yields in the UK shows that the approach allows for a flexible integration of aggregation processes and to propagate the uncertainties during aggregation. The approach also allows to tune the uncertainties in the data, depending on the aggregation function that is used.

Acknowledgments The research leading to these results has received funding from the European Union Seventh Framework Programme [FP7/2007-2013] under grant agreement no 248488. We are thankful to Jill Johnson and Sarah Knight from the Food and Environment Research Agency and Richard Jones from Aston University for the support during the integration of our approach in the yield prediction workflow.

References

Balazinska M, Deshpande A, Franklin M, Gibbons P, Gray J, Nath S, Hansen M, Liebhold M, Szalay A, Tao V (2007) Data management in the worldwide sensor web. Pervasive Comput IEEE 6(2):30–40 (April–June 2007)

Bastin L, Cornford D, Jones R, Heuvelink GBM, Stasch C, Pebesma E, Nativi S, Mazzetti P, Williams M (2012) Managing uncertainty in integrated environmental modelling frameworks: the uncertweb framework. Environ Model Softw 39:116–134

Benjelloun O, Sarma AD, Halevy A, Widom J (2006) ULDBs: databases with uncertainty and lineage. In: Proceedings of the 32nd international conference on very large data bases. VLDB '06, VLDB Endowment, pp 953–964

Beven K, Buytaert W, Smith LA (2012) On virtual observatories and modelled realities (or why discharge must be treated as a virtual variable). Hydrol Process 26(12):1905–1908

Bierkens M, Finke P, De Willingen P (2000) Upscaling and downscaling methods for environmental research. Kluwer Academic Publishers

Bigagli L, Nativi S, (eds) (2011) NetCDF uncertainty conventions (NetCDF-U). OGC 11–163. Open geospatial consortium, Inc, pp 17 (accessed 24 July 2012)

Bröring A, Echterhoff J, Jirka S, Simonis I, Everding T, Stasch C, Liang S, Lemmens R (2011) New generation sensor web enablement. Sensors 11(3):2652–2699

Domenico B (2011) OGC network common data form (NetCDF) core encoding standard version 1.0. OGC 10–090r3. Open geospatial consortium, Inc, pp 21 (Accessed on 01 Nov 2012)

Geller G, Turner, W.: The model web: a concept for ecological forecasting. In: Geoscience and Remote Sensing Symposium, (2007) IGARSS 2007. IEEE International. 2007:2469–2472

Gerharz L, Autermann C, Hopmann H, Stasch C, Pebesma E (2012) Uncertainty visualisation in the model web. European Geosciences Union (EGU) General Assembly

Gerharz L, Pebesma E (2012) Using geostatistical simulation to disaggregate air quality model results for individual exposure estimation on GPS tracks. Stoch Env Res Risk Assess 27:223–234

Graeler B, Stasch C (2012) Flexible representation of spatio-temporal random fields in the model web. European Geosciences Union (EGU) General Assembly

Heuvelink G (1998) Error propagation in environmental modelling with GIS. Taylor & Francis

Heuvelink G, Pebesma E (1999) Spatial aggregation and soil process modelling. Geoderma 89:47–65

ISO/TC211: ISO/FDIS 19156:2010: geographic information—observations and measurements. ISO/TC 211 (2010)

Jampani R, Xu F, Wu M, Perez LL, Jermaine C, Haas PJ (2008) MCDB: a Monte Carlo approach to managing uncertain data. In: Proceedings of the (2008) ACM SIGMOD international conference on Management of data. SIGMOD '08. New York, NY, USA, ACM, pp 687–700

Jeong SH, Fernandes AAA, Paton NW, Griffiths T (2004) A generic algorithmic framework for aggregation of spatio-temporal data. In: SSDBM '04: proceedings of the 16th international conference on scientific and statistical database management, Washington, DC, USA, IEEE Computer Society, p 245

Jirka S, Bröring A, Stasch C (2009) Discovery mechanisms for the sensor web. Sensors 9(4):2661–2681

Jones R, Cornford D, Bastin L (2012) UncertWeb processing service: making models easier to access on the web. Trans GIS 14(6):921–939

Maue P, Stasch C, Athanasopoulos G, Gerharz L (2011) Geospatial standards for web-enabled environmental models. Int J Spatial Data Infrastruct Res 6:145–167

Nativi S, Bigagli L (2009) Discovery, mediation, and access services for earth observation data. IEEE J Sel Top Appl Earth Observ Rem Sens 2(4):233–240

Nativi S, Mazzetti P, Geller GN (2012) Environmental model access and interoperability: the GEO model web initiative. Environ Model Softw 39:214–228. doi:10.1016/j.envsoft.2012.03.007

Pebesma E (2012) Spacetime: spatio-temporal data in R. J Stat Softw 51(7):1–30

Pross B, Gerharz L, Stasch C, Pebesma E (2012) Tools for uncertainty propagation in the model web using Monte Carlo simulation. In: Seppelt R, Voinov A, Lange S, Bankamp D (eds) Proceedings of the iEMSs sixth Biennial meeting: Managing resources of a limited planet. International congress on environmental modelling and software (iEMS 2012), international environmental modelling and software society (iEMSs)

R Development Core Team: R (2011) A Language and environment for statistical computing. R Foundation for statistical computing, Vienna, Austria. ISBN 3-900051-07-0

Schut P (2007) OpenGIS web processing service. OGC 05–007r7. Open Geospatial Consortium, Inc., 87pp. (Accessed on 24 July 2012)

Sheth A, Henson C, Sahoo S (2008) Semantic sensor web. IEEE Int Comput, pp 78–83

Stasch C, Foerster T, Autermann C, Pebesma E (2012) Spatio-temporal aggregation of European air quality observations in the sensor web. Comput Geosci 47:111–118

Stasch C, Autermann C, Foerster T, Pebesma E (2011) Towards a spatiotemporal aggregation service in the sensor web. Poster presentation. In: The 14th AGILE international conference on geographic information, science

Stasch C, Jones R, Cornford D, Kiesow M, Williams M, Pebesma E (2012) Representing Uncertainties in the Sensor Web. In: Proceedings of Workshop Sensing A Changing World

Taylor JR (1997) An introduction to error analysis: the study of uncertainties in physical measurements. University Science Books

Vega Lopez IF, Snodgrass RT, Moon B (2005) Spatiotemporal aggregate computation: a survey. IEEE Trans Knowl Data. Engineering 17(2):271–286

Williams M, Conford D, Bastin L, Pebesma E (2009) Uncertainty markup language (UncertML) (OGC 08–122r2)

Privacy-Preserving Distributed Movement Data Aggregation

Anna Monreale, Wendy Hui Wang, Francesca Pratesi, Salvatore Rinzivillo, Dino Pedreschi, Gennady Andrienko and Natalia Andrienko

Abstract We propose a novel approach to privacy-preserving analytical processing within a distributed setting, and tackle the problem of obtaining aggregated information about vehicle traffic in a city from movement data collected by individual vehicles and shipped to a central server. Movement data are sensitive because people's whereabouts have the potential to reveal intimate personal traits, such as religious or sexual preferences, and may allow re-identification of individuals in a database. We provide a privacy-preserving framework for movement data aggregation based on trajectory generalization in a distributed environment. The proposed solution, based on the differential privacy model and on sketching techniques for efficient data compression, provides a formal data protection safeguard. Using real-life data, we demonstrate the effectiveness of our approach also in terms of data utility preserved by the data transformation.

A. Monreale (✉) · F. Pratesi · D. Pedreschi
University of Pisa, Pisa, Italy
e-mail: annam@di.unipi.it

F. Pratesi
e-mail: prafra@yahoo.it

D. Pedreschi
e-mail: pedre@di.unipi.it

W. H. Wang
Stevens Institute of Technology, Hoboken, NJ, USA
e-mail: Hui.Wang@stevens.edu

S. Rinzivillo
ISTI-CNR, Pisa, Italy
e-mail: salvatore.rinzivillo@isti.cnr.it

G. Andrienko · N. Andrienko
Fraunhofer IAIS, Sankt Augustin, Germany
e-mail: gennady.andrienko@iais.fraunhofer.de

N. Andrienko
e-mail: natalia.andrienko@iais.fraunhofer.de

D. Vandenbroucke et al. (eds.), *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography,
DOI: 10.1007/978-3-319-00615-4_13,

1 Introduction

The widespread availability of low cost GPS devices enables collecting data about movements of people and objects at a large scale. Understanding of the human mobility behavior in a city is important for improving the use of city space and accessibility of various places and utilities, managing the traffic network, and reducing traffic jams. Generalization and aggregation of individual movement data can provide an overall description of traffic flows in a given time interval and their variation over time. Chapter Andrienko and Andrienko (2011) proposes a method for generalization and aggregation of movement data that requires having all individual data in a central station. This centralized setting entails two important problems: (a) the amount of information to be collected and processed may exceed the capacity of the storage and computational resources, and (b) the raw data describe the mobility behavior of the individuals with great detail that could enable the inference of very sensitive information related to the personal private sphere.

In order to solve these problems, we propose a privacy-preserving distributed computation framework for the aggregation of movement data. We assume that on-board location devices in vehicles continuously trace the positions of the vehicles and can periodically send derived information about their movements to a central station, which stores it. The vehicles provide a statistical sample of the whole population, so that the information can be used to compute a summary of the traffic conditions on the whole territory. To protect individual privacy, we propose a data transformation method based on the well-known differential privacy model. To reduce the amount of information that each vehicle transmits to the central station, we propose to apply the sketch techniques to the differentially private data to obtain a compressed representation. The central station, that we call *coordinator*, is able to reconstruct the movement data represented by the sketched data that, although transformed for guaranteeing privacy, preserve some important properties of the original data that make them useful for mobility analysis.

The remainder of the chapter is organized as follows. Section 2 introduces background information and definitions. Section 3 describes the system architecture and states the problem. Section 4 presents our privacy-preserving framework. In Sect. 5, we discuss the privacy analysis. Experimental results from applying our method to real-world data are presented and discussed in Sect. 6. Section 7 discusses the related work and Sect. 8 concludes the chapter.

2 Preliminaries

2.1 Movement Data Representation

Definition 1 (*Trajectory*) A Trajectory or spatio-temporal sequence is a sequence of triplets $T = <l_1, t_1>, \ldots, <l_n, t_n>$, where t_i $(i = 1 \ldots n)$ denotes a timestamp such that $\forall_{1 \leq i < n}\ t_i < t_{i+1}$ and $l_i = (x_i, y_i)$ are points in $\mathbf{R^2}$.

Intuitively, each pair $\langle l_i, t_i \rangle$ indicates that the object is in the position $l_i = \langle x_i, y_i \rangle$ at time t_i.

We assume that the territory is subdivided in cells $C = \{c_1, c_2, \ldots, c_p\}$ which compose a partition of the territory. During a travel a user goes from a cell to another cell. We use g to denote the function that applies the spatial generalization to a trajectory. Given a trajectory T this function generates the generalized trajectory $g(T)$, i.e. a sequence of *moves* with temporal annotations, where a *move* is an pair (l_{c_i}, l_{c_j}), which indicates that the moving object moves from the cell c_i to the *adjacent* cell c_j. Note that, l_{c_i} denotes the pair of spatial coordinates representing the centroid of the cell c_i; in other words $l_{c_i} = \langle x_{c_i}, y_{c_i} \rangle$. The *temporal annotated move* is the quadruple $(l_{c_i}, l_{c_j}, t_{c_i}, t_{c_j})$ where l_{c_i} is the location of the origin, l_{c_j} is the location of the destination and the t_{c_i}, t_{c_j} are the time information associate to l_{c_i} and l_{c_j}. As a consequence, we define a generalized trajectory as follows.

Definition 2 (*Generalized Trajectory*) Let $T = \langle l_1, t_1 \rangle, \ldots, \langle l_n, t_n \rangle$ be a trajectory. Let $C = \{c_1, c_2, \ldots, c_p\}$ be the set of the cells that compose the territory partition. A generalized version of T is a sequence of temporal annotated moves

$$T_g = (l_{c_1}, l_{c_2}, t_{c_1}, t_{c_2})(l_{c_2}, l_{c_3}, t_{c_2}, t_{c_3}) \ldots (l_{c_{m-1}}, l_{c_m}, t_{c_{m-1}}, t_{c_m})$$

where $m \leq n$.

Now, we show how a generalized trajectory can be represented by a frequency distribution vector. First, we define the function *Move Frequency MF* that given a generalized trajectory T_g, a move (l_{c_i}, l_{c_j}) and a time interval computes how many times the move appears in T_g by considering the temporal constraint. More formally, we define the *Move Frequency* function as follows.

Definition 3 (*Move Frequency*) Let T_g be a generalized trajectory and let (l_{c_i}, l_{c_j}) be a move. Let τ be a temporal interval. The *Move Frequency* function is defined as:

$$MF(T_g, (l_{c_i}, l_{c_j}), \tau) = |\{(l_{c_i}, l_{c_j}, t_i, t_j) \in T_g | t_i \in \tau \wedge t_j \in \tau\}|.$$

This function can be easily extended for taking into consideration a set of generalized trajectories $\mathscr{T}^{\mathscr{G}}$. In this case, the information computed by the function represents the total number of movements from the cell c_i to the cell c_j in a time interval in the set of trajectories.

Definition 4 (*Global Move Frequency*) Let $\mathscr{T}^{\mathscr{G}}$ be a set of generalized trajectories and let (c_i, c_j) be a move. Let τ be a time interval. The *Global Move Frequency* function is defined as:

$$GMF(\mathscr{T}^{\mathscr{G}}, (c_i, c_j), \tau) = \sum_{\forall T_g \in \mathscr{T}^{\mathscr{G}}} MF(T_g, (c_i, c_j), \tau).$$

The number of movements between two cells computed by either the function *MF* or *GMF* describes the amount of traffic flow between the two cells in a specific time interval. This information can be represented by a frequency distribution vector.

Definition 5 (*Vector of Moves*) Let $C = \{c_1, c_2, \ldots, c_p\}$ be the set of the cells that compose the territory partition. The *vector of moves* M with $s = |\{(c_i, c_j) | c_i \textit{ is adjacent to } c_j\}|$ positions is the vector containing all possible moves. The element $M[i] = (l_{c_i}, l_{c_j})$ is the *move* from the cell c_i to the adjacent cell c_j.

Definition 6 (*Frequency Vector*) Let $C = \{c_1, c_2, \ldots, c_p\}$ be the of the cells that compose the territory partition and let M be the vector of moves. Given a set of generalized trajectories in a time interval τ $\mathscr{T}^{\mathscr{G}}$. The corresponding *frequency vector* is the vector f with size $s = |\{(c_i, c_j) | c_i \textit{ is adjacent to } c_j\}|$ where each $f[i] = GMF(\mathscr{T}^{\mathscr{G}}, M[i], \tau)$.

The definition of *frequency vector of a trajectory set* is straightforward; it requires to compute the function *GMF* instead of *MF*.

Note that the above definitions are based on the assumption that consecutive locations can be contained in the same cell or in adjacent cells. In some cases (for example, because of GPS problems) this fact would not be true. In order to avoid illegal moves (i.e., moves that are not present in the Frequency Vector) a reasonable solution is to reconstruct the missing part of the trajectories, e.g. by interpolation.

2.2 Differential Privacy

Differential privacy implies that adding or deleting a single record does not significantly affect the result of any analysis. Intuitively, differential privacy can be understood as follows. Let a database D include a private data record d_i about an individual i. By querying the database, it is possible to obtain certain information I(D) about all data and information I(D $-$ d_i) about the data without the record d_i. The difference between I(D) and I(D $-$ d_i) may enable inferring some private information about the individual i. Hence, it must be guaranteed that I(D) and I(D $-$ d_i) do not significantly differ for any individual i.

The formal definition (Dwork et al. 2006) is the following. Here the parameter, ε, specifies the level of privacy guaranteed.

Definition 7 (*ε-differential privacy*) A privacy mechanism A gives ε-differential privacy if for any dataset D_1 and D_2 differing on at most one record, and for any possible output D' of A we have

$$Pr[A(D_1) = D'] \leq e^{\varepsilon} \times Pr[A(D_2) = D']$$

where the probability is taken over the randomness of A.

Two principal techniques for achieving differential privacy have appeared in the literature, one for real-valued outputs (Dwork et al. 2006) and the other for outputs of arbitrary types (McSherry and Talwar 2007). A fundamental concept of both techniques is the global sensitivity of a function mapping underlying datasets to (vectors of) reals.

Definition 8 (*Global Sensitivity*) For any function $f : D \rightarrow R^d$, the sensitivity of f is

$$\Delta f = max_{D_1,D_2} ||f(D_1) - f(D_2)||_1$$

for all D_1, D_2 differing in at most one record.

For the analysis whose outputs are real, a standard mechanism to achieve differential privacy is to add Laplace noise to the true output of a function. Dwork et al. (2006) propose the Laplace mechanism which takes as inputs a dataset D, a function f, and the privacy parameter ε. The magnitude of the noise added conforms to a Laplace distribution with the probability density function $p(x|\lambda) = \frac{1}{2\lambda} e^{-|x|/\lambda}$, where λ is determined by both the global sensitivity of f and the desired privacy level ε.

Theorem 1 (Dwork et al. 2006) *For any function $f : D \rightarrow R^d$ over an arbitrary domain D, the mechanism A $A(D) = f(D) + Laplace(\Delta f/\varepsilon)$ gives ε-differential privacy.*

A relaxed version of differential privacy discussed in Michael and Sebastian (2012) allows claiming the same privacy level as Definition 7 in the case there is a small amount of privacy loss due to a variation in the output distribution for the privacy mechanism A is as follows.

Definition 9 [(ε, δ)-*differential privacy*] A privacy mechanism A gives (ε, δ)-differential privacy if for any dataset D_1 and D_2 differing on at most one record, and for any possible output D' of A we have

$$Pr[A(D_1) = D'] \leq e^{\varepsilon} \times Pr[A(D_2) = D'] + \delta$$

where the probability is taken over the randomness of A.

Note that, if $\delta = 0$, $(\varepsilon; 0)$-differential privacy is ε-differential privacy. In the remaining of this chapter we will refer to this last version of differential privacy.

3 Problem Definition

3.1 System Architecture

We consider a system architecture as that one in described in Cormode and Garofalakis (2008). In particular, we assume a distributed-computing environment comprising a collection of k (trusted) remote sites (nodes) and a designated coordinator site. Streams of data updates arrive continuously at remote sites, while the coordinator site is responsible for generating approximate answers to periodic user queries posed over the unions of remotely-observed streams across all sites. Each remote site exchanges messages only with the coordinator, providing it with state information on its (locally observed) streams. There is no communication between remote sites.

In our scenario, the coordinator is responsible for computing the aggregation of movement data on a territory by combining the information received by each node. In order to obtain the aggregation of the movement data in the centralized setting we need to generalize all the trajectories by using the cells of a partition of the territory. In our distributed setting we assume that the partition of the territory, i.e., the set of cells $C = \{c_1, \ldots, c_p\}$ useful for the generalization, is known by both all the nodes and the coordinator. Each node, that represents a vehicle that moves in this territory, in a given time interval observes a sequence of spatio-temporal points (trajectory), generalizes it and contributes to the computation of the global vector.

Formally, each remote site $j \in \{1, \ldots, k\}$ observes local update streams that incrementally render a collection of (up to) s distinct frequency distribution vectors (equivalently, multi-sets) $f_{1,j}, \ldots, f_{s,j}$ over data elements from corresponding integer domains $[U_i] = \{0, \ldots, U_{i_1}\}$, for $i = 1, \ldots, s$; that is, $f_{i,j}[v]$ denotes the frequency of element $v \in [U_i]$ observed locally at remote site j.

The coordinator for each $i \in \{1, \ldots, s\}$ computes the *global frequency distribution vector* $f_i = \sum_{j=1}^{k} f_{i,j}$.

3.2 Privacy Model

In our setting we assume that each node in our system is secure; in other words we do not consider attacks at node level. Instead, we take into consideration possible attacks from any intruder between the node and the coordinator (i.e., attacks during the communications), and any intruder at coordinator site, so our privacy preserving technique has to guarantee privacy even against a malicious behavior of

the coordinator. We consider sensitive information as any information from which the typical mobility behavior of a user may be inferred. This information is considered sensitive for two main reasons: (1) typical movements can be used to identify the drivers who drive specific vehicles even when a simple de-identification of the individual in the system is applied; and (2) the places visited by a driver could identify specific sensitive areas such as clinics, hospitals, the user's home.

Therefore, we need to find effective privacy mechanisms on the real count associate to each move, in order to generate uncertainty. As a consequence, the goal of our framework is to compute a distributed aggregation of movement data for a comprehensive exploration of them while preserving privacy.

Definition 10 (*Problem Definition*)

Given a set of cells $C = \{c_1, \ldots, c_p\}$ and a set $V = \{V_1, \ldots, V_k\}$ of vehicles, the *privacy-preserving distributed movement data aggregation problem* (DMAP) consists in computing, in a specific time interval τ the $f^{\tau}_{DMAP}(V) = [f_1, f_2, \ldots, f_s]$, where each $f_i = GMF(\mathscr{T}^{\mathscr{G}}, M[i], \tau)$ and $s = |\{(c_i, c_j) | c_i \textit{ is adjacent to } c_j\}|$, while preserving privacy. Here, $\mathscr{T}^{\mathscr{G}}$ is the set of generalized trajectories related to the k vehicles V in the time interval τ and M is the vector of moves defined on the set of cells C.

4 Our Solution

Clearly, in order to guarantee the privacy within this framework we may apply many privacy-preserving techniques depending on the privacy attack model and the background knowledge of the adversary that we want to consider in this scenario. In this chapter, we provide a solution based on the differential privacy, that is a very strong privacy model independent on the the background knowledge of an adversary. In this section, we describe the details of our solution, including the computation of each node and the coordinator in the system.

The pseudo code of our algorithm is shown in Algorithm 1. Each node represents a vehicle that moves in a specific territory and this vehicle in a given time interval observes sequences of spatio-temporal points (trajectories) and computes the corresponding frequency vector to be sent to the coordinator. The computation of the frequency vector requires four steps described in Algorithm 1: (a) trajectory generalization; (b) frequency vector construction; (c) frequency vector transformation for privacy guarantees and (d) vector sketching for compressing the information to be transmitted.

Algorithm 1: NodeComputation(ε, τ, M, T^G, w, d)

Input: A privacy budget ε, a time interval τ, the vector of the moves M, a set of trajectories T^G, the dimension of the sketch w and d.

Output: The sketch vector representing the privacy-preserving frequency vector $sk(\tilde{f}^{V_j})$.

foreach *observed trajectory* T **do**
  // Trajectory Generalization (Sec. 4.1)
  $T_g = TrajectoryGeneralization(M, T)$;
  // Update of the Frequency Vector f^{V_j} **(Sec. 4.2)**
  foreach *move* $(l_{c_i}, l_{c_j}) \in T_g$ **do**
    $n = MF(T_g, (l_{c_i}, l_{c_j}), \tau)$;
    $f^{V_j}[(l_{c_i}, l_{c_j})] += n$;
// Privacy Transformation (Sec. 4.3)
foreach *vector element* $f^{V_j}[i]$ **do**
  $noise = Laplace(0, \frac{1}{\varepsilon})$;
  if $noise > f^{V_j}[i]$ **then**
    $noise = f^{V_j}[i]$;
  if $noise < -f^{V_j}[i]$ **then**
    $noise = -f^{V_j}[i]$;
  $\tilde{f}^{V_j}[i] = f^{V_j}[i] + noise$;
// Generation of Sketch Vector (Sec. 4.4)
$sk(\tilde{f}^{V_j}) = CountMin(\tilde{f}^{V_j})$;
return $sk(\tilde{f}^{V_j}, w, d)$

4.1 Trajectory Generalization

Given a specific division of the territory, a trajectory is generalized in the following way. We apply place-based division of the trajectory into segments. The area c_1 containing its first point l_1 is found. Then, the second and following points of the trajectory are checked for being inside c_1 until finding a point l_i not contained in c_1. For this point l_i, the containing area c_2 is found.

The trajectory segment from the first point to the i-th point is represented by the vector (c_1, c_2). Then, the procedure is repeated: the points starting from l_{i+1} are checked for containment in c_2 until finding a point l_k outside c_2, the area c_3 containing l_k is found, and so forth up to the last point of the trajectory.

In the result, the trajectory is represented by the sequence of moves $(c_1, c_2, t_1, t_2)(c_2, c_3, t_2, t_3) \ldots (c_{m-1}, c_m, t_{m-1}, t_m)$. Here, in a specific quadruple t_i is the time moment of the last position in c_i and t_j is the time moment of the last position in c_j. There may be also a case when all points of a trajectory are contained in one and the same area c_1. Then, the whole trajectory is represented by the sequence $\{c_1\}$. Since, globally we want to compute aggregation of moves we discard this kind of trajectories. Moreover, as most of the methods for analysis of trajectories are suited to work with positions specified as points, the areas $\{c_1, c_2, \ldots, c_m\}$ are replaced, for practical purposes, by the sequence $l_{c_1}, l_{c_2}, \ldots, l_{c_m}$ consisting of the centroids of the areas $\{c_1, c_2, \ldots, c_m\}$.

4.2 Frequency Vector Construction

After the generalization of a trajectory, the node computes the *Move Frequency* function for each move (l_{c_i}, l_{c_j}) in that trajectory and updates its frequency vector f^{V_j} associated to the current time interval τ. Intuitively, the vehicle populates the frequency vector f^{V_j} according the generalized trajectory observed. So, at the end of the time interval τ the element $f^{V_j}[i]$ contains the number of times that the vehicle V_j moved from m to n in that time interval, if $M[i] = (m, n)$.

4.3 Vector Transformation for Achieving Privacy

As we stated in Sect. 3.2, if a node sends the frequency vector without any data transformation any intruder may infer the typical movements of the vehicle represented by the node. As an example, he could learn his most frequent move; this information can be considered very sensitive because the cells of this move usually correspond to user's home and his work place. Clearly, the generalization step can help the privacy user but it depends on the density of the area; specifically, if the area is not so dense it could identify few places and in that case the privacy is at risk. *How can we hide the event that the user moved from a location a to a location b in the time interval* τ? We propose a solution based on a very strong privacy model called ε-differential privacy (Sect. 2.2). As explained above the key point of this model is the definition of the sensitivity. Given a move (a, b) its sensitivity is straightforward: releasing its frequency have sensitivity 1 as adding or removing a single flow can affect its frequency by at most 1. Thus adding noise according to $Lap(\frac{1}{\varepsilon})$ to the frequency of each of the moves in the frequency vector satisfies ε-differential privacy. As a consequence, at the end of the time interval τ, before sending the vector to the coordinator, for each position of the vector (i.e., for each move) has to generate the noise by the Laplace distribution with zero mean and scale $\frac{1}{\varepsilon}$ and then has to add it to the value in that position of the vector. At the end of this step the node transforms f^{V_j} into $\tilde{f}^{V_j}$.

Differential privacy must be applied with caution because in some context it could lead to the destruction of the data utility because of the added noise that, although with small probability, can reach values of arbitrary magnitude. Moreover, adding noise drawn from the Laplace distribution could generate negative values for the flow in a move and negative flows does not make sense. To prevent this two problems we decided to draw the noise from a cutting version of the Laplace distribution. In particular, for each value x of the vector f^{V_j} we draw the noise from $Lap(\frac{1}{\varepsilon})$ bounding the noise value to the interval $[-x, x]$. In other words, if we have the original flow $f^{V_j}[i] = x$ in the perturbed version we obtain a flow value in the interval $[0, 2x]$. The use of a truncated version of the Laplace distribution can lead to privacy leaks and in Sect. 5 we show that our privacy mechanism satisfies (ε, δ)-differential privacy, where δ represents this privacy loss.

4.4 Vector Sketching for Compact Communications

In a system like ours an important issue to be considered is the amount of data to be communicated. In fact, real life systems usually involve 1,000 vehicles (nodes) that are located in any place of the territory. Each vehicle has to send to the coordinator the information contained in its frequency vector that has a size depending on the number of cells that represent the partition of the territory. The number of cells in a territory can be very huge and this can make each frequency vector too big. As an example, in the dataset of real-life trajectories used in our experiments we have about 4,200 vehicles and a frequency vector with about 15,900 positions. Therefore, the system has to be able to handle both a very large number of nodes and a huge amount of the global information to be communicated. These considerations make the reduction of the information communicated necessary. We propose the application of a sketching method (Cormode et al. 2012a) that allows us to apply a good compression of the information to be communicated. In particular, we propose the application of *Count-Min* sketch algorithm (Cormode and Muthukrishnan 2005). In general, this algorithm maps the frequency vector onto a more compressed vector. In particular, the sketch consists of an array C of $d \times w$ counters and for each of d rows a pairwise independent hash functions h_j, that maps items onto $[w]$. Each item is mapped onto d entries in the array, by adding to the previous value the new item. Given a sketch representation of a vector we can estimate the original value of each component of the vector by the following function $f[i] = min_{1 \leq j \leq d} C[j, h_j(i)]$. The estimation of each component j is affected by an error, but it is showed that the overestimate is less than n/w, where n is the number of components. So, setting $d = \log \frac{1}{\gamma}$ and $w = O(\frac{1}{\alpha})$ ensures that the estimation of $f[i]$ has error at most αn with probability at least $1 - \gamma$. Here, α indicates the accuracy (i.e. the approximation error), and γ represents the probability of exceeding the accuracy bounds.

4.5 Coordinator Computation

The computation of the coordinator is composed of two main phases: (1) computation of the set of moves and (2) computation of the aggregation of global movements.

Move Vector Computation. The coordinator in an initial setting phase has to send to the nodes the *vector of moves* (Definition 5). The computation of this vector depends on the set of cells that represent the partition of the territory. This partition can be a simple grid or a more sophisticated territory subdivision such as the Voronoi tessellation. The sharing of vector of moves is a requirement of the whole process because each node has to use the same data structure for allowing the coordinator the correct computation of the global flows.

Global Flow Computation. The coordinator has to compute the global vector that corresponds to the global aggregation of movement data in a given time interval τ by composing all the local frequency vectors. It receives the sketch vector $sk(\tilde{f}^{V_j})$ from each node; then it reconstructs each frequency vector from the sketch vector, by using the estimation described in Sect. 4.4. Finally, the coordinator computes the global frequency vector by summing the estimate vectors component by component. Clearly the estimate global vector is an approximated version of the global vector obtained by summing the local frequency vectors after the only privacy transformation.

5 Privacy Analysis

As pointed out in Kifer and Machanavajjhala (2011) differential privacy must be applied with caution. The privacy protection provided by differential privacy relates to the data generating mechanism and deterministic aggregate level background knowledge. As in our problem the trajectories in the raw database are independent of each other, and no deterministic statistics of the raw database will ever be released, we are ready to show that Algorithm 1 satisfies (ε, δ)-differential privacy.

Let F and F' be the frequency distribution before and after adding Laplace noise. We observe that bounding the Laplace noise will lead to some privacy leakage on some values. For instance, from the noisy frequency values that are large, the attacker can infer that these values should not be transformed from small ones. To analyze the privacy leakage of our bound-noise approach, we first explain the concept of *statistical distance*. Statistical distance is defined in Michael and Sebastian (2012). Formally, given two distributions X and Y, the *statistical distance* between X and Y over a set U is defined as

$$d(X, Y) = max_{S \in U}(Pr[X \in S] - Pr[Y \in S]).$$

Michael and Sebastian (2012) also shows the relationship between (ε, δ)-differential privacy and the statistical distance.

Lemma1 (Michael and Sebastian 2012) *Given two probabilistic functions F and G with the same input domain, where* F *is* (ε, δ_1)*-differentially private. If for all possible inputs x we have that the statistical distance on the output distributions of F and G is*:

$$d(F(x), G(x)) \leq \delta_2,$$

then G *is* $(\varepsilon, \delta_1 + (e^{\varepsilon} + 1)\delta_2)$*-differentially private*.

Let F and F' be the frequency distribution before and after adding Laplace noise. We can show that the statistical distance between F and F' can be bounded as follows:

Lemma 2 (Michael and Sebastian 2012) *Given an* (ε, δ)-*differentially private function* F *with* $F(x) = f(x) + R$ *for a deterministic function* f *and a random variable* R. *Then for all* x, *the statistical distance between* F *and its throughput-respecting variant* F' *with the bound* b *on* R *is at most*

$$d(F(x) - F'(x)) \leq Pr[|R| > b].$$

Michael and Sebastian (2012) has the following lemma to bound the probability $Pr[|R| > b]$.

Lemma 3 (Michael and Sebastian 2012) *Given a function* F *with* $F(x) = f(x) + Lap(\frac{\Delta f}{\varepsilon})$ *for a deterministic function* f, *the probability that the Laplacian noise* $Lap(\frac{\Delta f}{\varepsilon})$ *applied to* f *is larger than* b *is bounded by*:

$$Pr(Lap(\frac{\Delta f}{\varepsilon}) > b) \leq \frac{2(\Delta f)^2}{b^2\varepsilon^2}.$$

We stress that in our approach, the bound b of each frequency value x is not fixed. Indeed, $b = x$. Therefore, each frequency value x has different amounts of privacy leakage. Our approach thus achieves different degree of (ε, δ)-differentially privacy guarantee on each frequency value x. Theorem 2 shows more details.

Theorem 2 *Given the total privacy budget* ε, *for each frequency value* x, *Algorithm* 1 *ensures* $(\varepsilon, (e^{\varepsilon} + 1)\frac{2}{x^2\varepsilon^2})$-d*ifferential privacy*.

Proof Algorithm 1 consists of four steps, namely *TrajectoryGeneralization*, *FrequencyVectorUpdate*, *PrivacyTransformation*, and *SketchVectorGeneration*. The steps of *TrajectoryGeneralization* and *FrequencyVectorUpdate* mainly prepare the frequency vectors for privacy transformation. Hence we focus on the privacy guarantee of *PrivacyTransformation* and *SketchVectorGeneration* steps. For each frequency value x, the *PrivacyTransformation* step can achieve $(\varepsilon, (e^{\varepsilon} + 1)\frac{2(\Delta f)^2}{x^2\varepsilon^2}$-differentially privacy. This can be easily proven by Lemma 1 and Lemma 3. Note that the the frequency vectors with Laplace noise (without truncation) satisfies $(\varepsilon, 0)$-differentially privacy. In our approach, $\Delta f = 1$. Thus the *PrivacyTransformation* step can achieve $(\varepsilon, (e^{\varepsilon} + 1)\frac{2}{x^2\varepsilon^2})$-differentially privacy. For the *SketchVectorGeneration* step, it only accesses a differentially private frequency vector, not the underlying database. As proven by Michael et al. (2010), a post-processing of differentially private results remains differentially private. Therefore, Algorithm 1 as a whole maintains $(e^{\varepsilon} + 1)\frac{2}{x^2\varepsilon^2})$-differentially privacy. □

6 Experiments

6.1 Dataset

For our experiments we used a large dataset of GPS vehicles traces, collected in a period from 1st May to 31st May 2011. In our simulation, the coordinator collects the FV from all the vehicles to determine the Global Frequency Vector (GFV), i.e. the sum all the trajectories crossing any link, at the end of each day, so we defined a series of time intervals τ_i, where each τ_i spans over a single day. In the following we show the resulting GFV for the 25th May 2011, but similar accuracy is observed also for the other days. The GPS traces were collected in the geographical areas around Pisa, in central Italy, and it counts for around 4,200 vehicles, generating around 15,700 trips.

6.2 Space Tessellation

The generalization and aggregation of movement data is based on space partitioning. Arbitrary territory divisions, such as administrative districts or regular grids, do not reflect the spatial distribution of the data. The resulting aggregations may not convey the essential spatial and quantitative properties of the traffic flows over the territory. Our method for territory partitioning extends the data-driven method suggested in chapter Andrienko and Andrienko (2011). Using a given sample of points (which may be, for example, randomly selected from a historical set of movement data), the original method finds spatial clusters of points that can be enclosed by circles with a user-chosen radius. The centroids of the clusters are then taken as generating seeds for Voronoi tessellation of the territory. We have modified the method so that dense point clusters can be subdivided into smaller clusters, so that the sizes of the resulting Voronoi polygons vary depending on the point density: large polygons in data-sparse areas and small polygons in data-dense areas. The method requires the user to set 3 parameters: maximal radius R, minimal radius r, and minimal number of points N allowing a cluster to be subdivided. In our experiments, we used a tessellation with 2,681 polygons obtained with R = 10 km, r = 500 m, N = 80.

6.3 Utility Evaluation

In the proposed framework, the coordinator collects the Frequency Vectors from all the vehicles in the time interval τ and aggregate them to obtain the resulting GFV, representing the flow values for each link of the spatial tessellation. Each FV received from the vehicles is perturbed by means of a two-step transformation:

Table 1 Reduced sizes of the FV for different values of α and γ

	α	γ	Columns (w)	Rows (d)	$w \times d$
CM_{5k}	0.0008	0.1	2,500	2	5,000
CM_{7k}	0.00078	0.05	2,564	3	7,692
CM_{10k}	0.00057	0.05	3,508	3	10,524

privacy transformation—with the objective of protecting sensitive information–, and sketches summarization—to reduce the volume of communication to be sent. These two transformations are regulated by two set of parameters: ε for the differential privacy transformation, and α and γ for the Count-Min Sketch summarization. When ε tends to 1 very little perturbation is introduced and this yields a low privacy protection. On the contrary, better privacy guarantees are obtained when ε tends to zero. The two parameters α and γ regulate the compression of the FV to be sent to the coordinator. Table 1 shows how the choice of these two parameters influences the final size of the FV. For example, for $\alpha = 0.0008$ and $\gamma = 0.1$ the original FV of 16k entries is reduced to a vector of 5k cells.

Since the two transformations operate on the entries of the FV, and hence on the flows, we compare two measures: (1) the *flow per link* (*fpl*), i.e. the directed volume of traffic between two adjacent zones; (2) the *flow per zone* (*fpz*), i.e. the sum of the incoming and outgoing flows in a zone. Figure 1 shows the resulting distributions of different privacy transformation with $\varepsilon = 0.9, 0.5, 0.3$. Figure 1 (*left*) shows the reconstructed flows per link: fixed a value of flow (x) we count the number of links (y) that have that flow. Figure 1 (*right*) shows the distribution of sum of flows passing for each zone: given a flow value (x) it shows how many zones (y) present that total flow.

From the distribution we can notice how the privacy transformation preserves very well the distribution of the original flows, even for more restrictive values of the parameter ε.

When we consider several flows together, like those incident to a given zone [Fig. 1 (*right*)], the distribution curves present several local variations, however the general shape is preserved for all the privacy transformations. Since the global distributions are comparable, we choose a value 0.3 for ε for the following discussions, in order to obtain a better privacy protection.

Fixed the privacy transformation parameter, we can evaluate the error introduced by the Count-Min sketch summarization. In Fig. 2 we can appreciate how a large compression of the FV yields a precise reconstruction of the transformed flows. In fact, we can observe that the general shape of the distribution curves are also preserved after the application of sketching techniques.

To maintain the data utility for mobility density analysis, we want to preserve the relative density distribution over the zones, i.e. it is desirable that former zones with low (high) traffic still present low (high) traffic after the transformations. To check this property, we show in Fig. 3 the correlation plots to compare the original flows with the transformed ones. From the charts we can notice how the

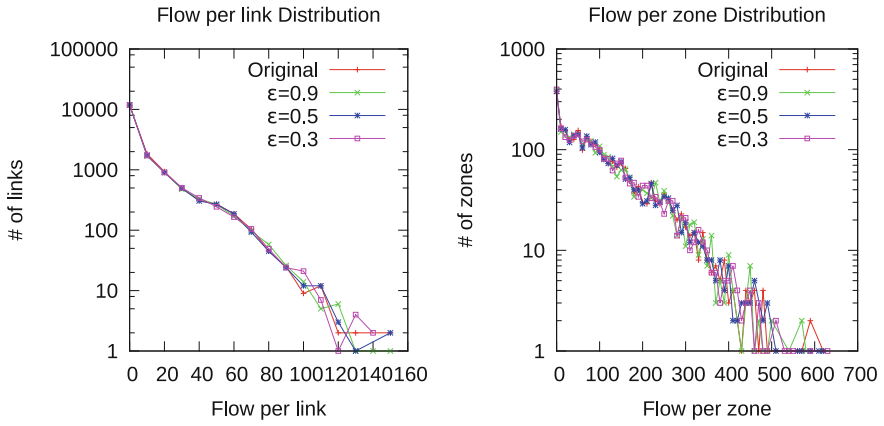


Fig. 1 Distribution of flow per link (*left*) and flow per zone (*right*)

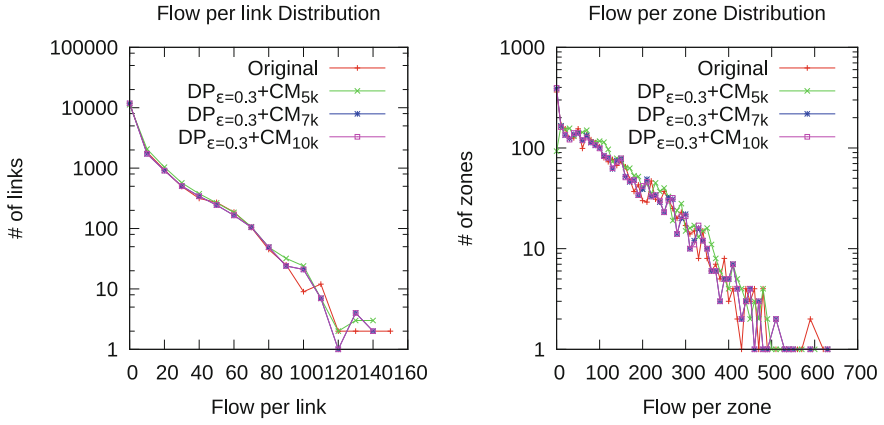


Fig. 2 Distribution of flow per link (*left*) and flow per zone (*right*) after the Count-Min sketch transformation

transformed flows maintain a very strong correlation with the original ones, enabling relative flows comparisons also in the transformed data.

Qualitatively, Fig. 4 shows a visually comparison of each Sketch summarization with the original flows. Each flow is draw with arrows with thickness proportional to the volume of trajectories observed on a link. From the figure it is evident how the relevant flows are preserved in all the transformed GFV, revealing the major highways and urban centers.

Similarly, the flow per zone is also preserved, as it is shown in Fig. 5, where the flow per each cell is rendered with a circle of radius proportional to the difference from the median value of each GFV. The maps allow us to recognize the dense areas (red circles, above the median) separated by sparse areas (blue circle below

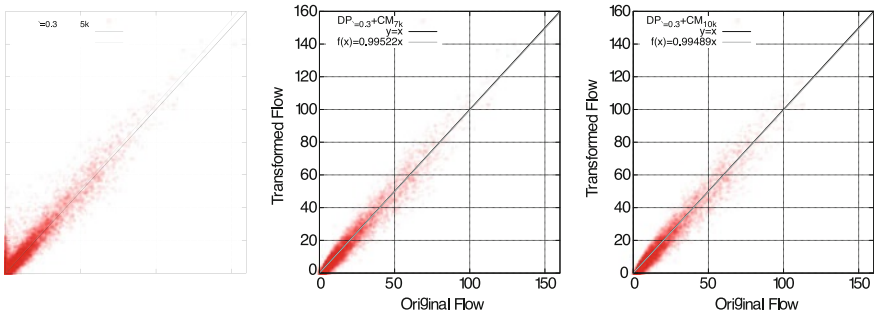


Fig. 3 Correlation between original flows and transformed flows with DP $\varepsilon = 0.3$ and CM_{5k} (*first*), CM_{7k} (*second*), CM_{10k} (*third*)

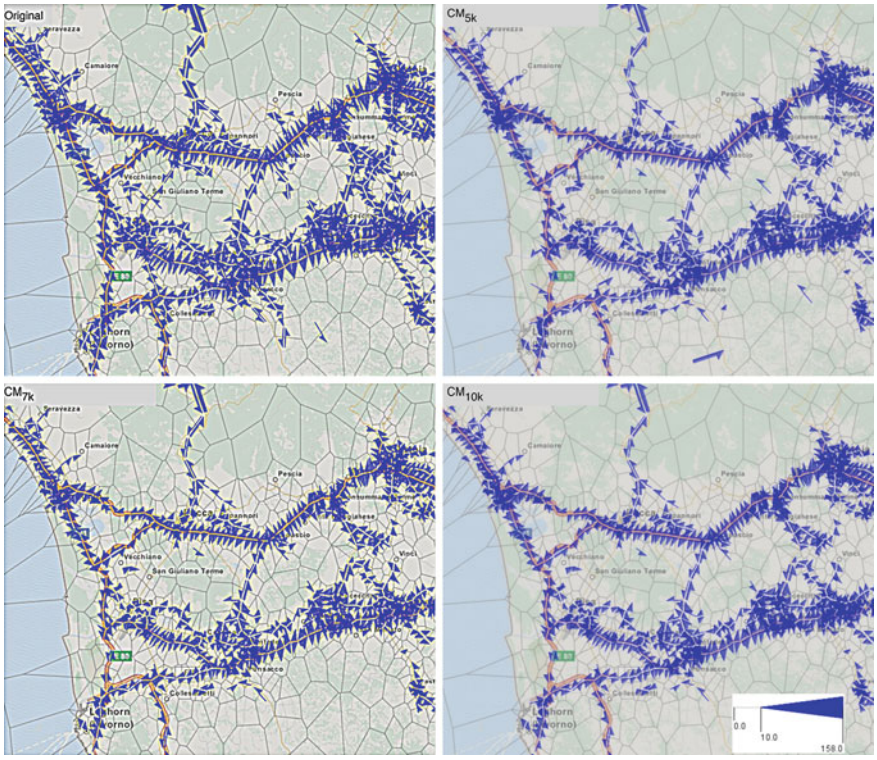


Fig. 4 Comparison of the original flows (**a**) with the GMF obtained with $\varepsilon = 0.3$ and CM_{5k} (**b**), CM_{7k} (**c**), and CM_{10k} (**d**)

the median). The high density traffic zones follow the highways and the major city centers along their routes.

The two comparisons proposed above give the intuition that, while the transformations protect individual sensitive information, the utility of data is preserved.

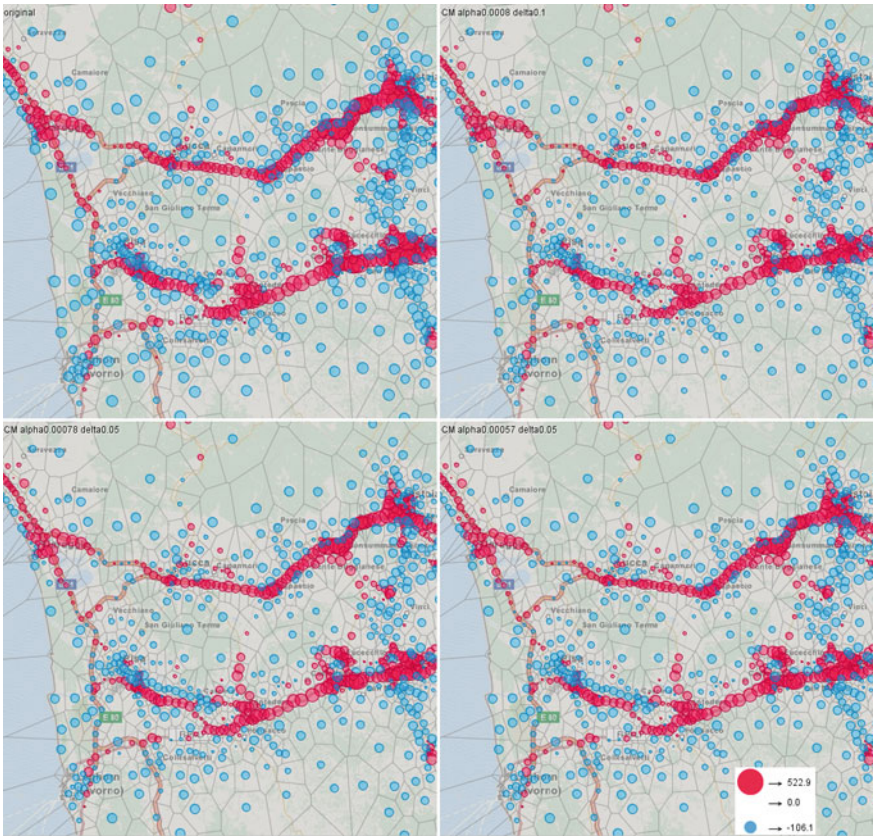


Fig. 5 Traffic flow per zone drawn with *circles* proportional to the difference from the median for each transformation with privacy transformations with different parameters (**a**), without privacy transformation; $\varepsilon = 0.3$ and CM_{5k} (**b**), CM_{7k} (**c**), and CM_{10k} (**d**)

7 Related Work

The existing methods of privacy-preserving publishing of trajectories can be categorized into two classes: (1) generalization/suppression based data perturbation, and (2) differential privacy.

Generalization/suppression based data perturbation techniques. There have been some recent works on privacy-preserving publishing of spatio-temporal moving points by using the generalization/suppression techniques. The mostly widely used privacy model of these work is adapted from what so called k-anonymity (Samarati and Sweeney 1998a, b), which requires that an individual should not be identifiable from a group of size smaller than k based on their quasi-identifies (QIDs), i.e., a set of attributes that can be used to uniquely identify the individuals. Abul et al. (2008) proposes the (k, δ)-anonymity model that exploits

the inherent uncertainty of the moving object's whereabouts, where δ represents possible location imprecision. Terrovitis and Mamoulis (2008) assume that different adversaries own different, disjoint parts of the trajectories. Their anonymization technique is based on *suppression* of the dangerous observations from each trajectory. Yarovoy et al. (2009) consider timestamps as the quasi-identifiers, and define a method based on *k -anonymity* to defend against an attack called *attack graphs*. Monreale et al. (2010) propose a spatial generalization approach to achieve *k*-anonymity. A general problem of these *k*-anonymity based privacy preserving techniques is that these techniques assume a certain level of background knowledge of the attackers, which may not be available to the data owner in practice.

Differential privacy. The recently proposed concept of *differential privacy* (DP) (Dwork et al. 2006) addresses the above issue. There are two popular mechanisms to achieve differential privacy, *Laplace* mechanism that supports queries whose outputs are numerical (Dwork et al. 2006) and *exponential mechanism* that works for any queries whose output spaces are discrete (McSherry and Talwar 2007). The basic idea of the Laplace mechanism is to add noise to aggregate queries (e.g., counts) or queries that can be reduced to simple aggregates. The Laplace mechanism has been widely adopted in many existing work for various data applications. For instance, Xiaokui et al. (2011), Cormode et al. (2012b) present methods for minimizing the worst-case error of count queries; Barak et al. (2007), Ding et al. (2011) consider the publication of data cubes; Michael et al. (2010), Xu et al. (2012) focus on publishing histograms; and Mohammed et al. (2011), Ninghui et al. (2012) propose the methods of releasing data in a differential private way for data mining. On the other hand, for the analysis whose outputs are not real or make no sense after adding noise, the exponential mechanism selects an output from the output domain, $r \in R$, by taking into consideration its score of a given utility function q in a differentially private manner. It has been applied for the publication of audition results (McSherry and Talwar 2007), coresets (Feldman et al. 2009), frequent patterns (Bhaskar et al. 2010) and decision trees (Friedman and Schuster 2010).

Regarding publishing differentially private spatial data, Chen et al. (2012) propose to release a prefix tree of trajectories with injected Laplace noise. Each node in the prefix tree contains a doublet in the form of $< tr(v), c(v) >$, where $tr(v)$ is the set of trajectories of the prefix v, and $c(v)$ is a version of $|tr(v)|$ with Laplace noise. Compared with our work, the prefix tree in Chen et al. (2012) is *data-dependent*, i.e., it should have a different structure when the underlying database changes. In our work, the frequency vector is *data-independent*. Cormode et al. (2012b) present a solution to publish differentially private spatial index (e.g., quadtrees and kd-trees) to provide a private description of the data distribution. Its main utility concern is the accuracy of multi-dimensional range queries (e.g., how many individuals fall within a given region). Therefore, the spatial index only stores the count of a specific spatial decomposition. It does not store the movement information (e.g., how many individuals move from location i to location j) as in

our work. In another chapter, Cormode et al. (2012c) proposes to publish a contingency table of trajectory data. The contingency table can be indexed by specific locations so that each cell in the table contains the number of people who commute from the given source to the given destination. The contingency table is very similar to our frequency vector structure. However, Cormode et al. (2012c) has a different focus from ours: we investigate how to publish the frequency vector in a differential privacy way, while Cormode et al. (2012c) address the sparsity issue of the contingency table and presents a method of releasing a compact summary of the contingency table with Laplace noise.

There are some work on publishing time-series data with differential privacy guarantee (McSherry and Mahajan 2010; Rastogi and Nath 2010). Since we only consider spatial data, these work are complement to our work.

8 Conclusion

In this chapter, we have studied the problem of computing movement data aggregation based on trajectory generalization in a distributed system while preserving privacy. We have proposed a method based on the well-known notion of differential privacy that provides very nice data protection guarantees. In particular, in our framework each vehicle, before sending the information about its movements within a time interval, applies to the data a transformation for achieving privacy and then, creates a summarization of the private data (by using a sketching algorithm) for reducing the amount of information to be communicated. The results obtained in our experiments show that the proposed method preserves some important properties of the original data allowing the analyst to use them for important mobility data analysis.

Future investigations could be directed to explore other methods for achieving differential privacy; as an example, it would be interesting to understand the impact of the use of the geometric mechanism instead of the Laplace one for achieving differential privacy.

Acknowledgments This work has been partially supported by EU FET-Open project LIFT (FP7-ICT-2009-C n. 255951) and EU FET-Open project DATA SIM (FP7-ICT 270833)

References

Abul O, Bonchi F, Nanni M (2008) Never walk alone: uncertainty for anonymity in moving objects databases. In: Proceedings of the 2008 IEEE 24th international conference on data engineering (ICDE), pp 376–385

Andrienko N, Andrienko G (2011) Spatial generalization and aggregation of massive movement data. IEEE Trans Visual Comput Graphics 17:205–219

Backes M, Meiser S (2012) Differentially private smart metering with battery recharging. IACR cryptology ePrint archive, p 183

Barak B, Chaudhuri K, Dwork C, Kale S, McSherry F, Talwar K (2007) Privacy, accuracy, and consistency too: a holsistic solution to contingency table release. In: Proceedings of the 26th ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems (PODS), pp 273–282

Bhaskar R, Laxman S, Smith A, Thakurta A (2010) Discovering frequent patterns in sensitive data. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining (KDD), pp 503–512

Chen R, Fung BCM, Desai BC, Sossou NM (2012) Differentially private transit data publication: a case study on the montreal transportation system. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining (KDD), pp 213–221

Cormode G, Muthukrishnan S (2005) An improved data stream summary: the count-min sketch and its applications. J Algorithms 55(1):58–75

Cormode G, Garofalakis MN (2008) Approximate continuous querying over distributed streams. ACM Trans Database Syst 33(2)

Cormode G, Garofalakis MN, Haas PJ, Jermaine C (2012a) Synopses for massive data: samples, histograms, wavelets, sketches. Found Trends Databases 4(1–3):1–294

Cormode G, Procopiuc CM, Srivastava D, Shen E, Yu T (2012b) Differentially private spatial decompositions. In: ICDE, pp 20–31

Cormode G, Procopiuc CM, Srivastava D, Tran TTL (2012c) Differentially private summaries for sparse data. In: ICDT, pp 299–311

Ding B, Winslett M, Han J, Li Z (2011) Differentially private data cubes: optimizing noise sources and consistency. In: Proceedings of the 2011 ACM SIGMOD international conference on management of data, pp 217–228

Dwork C, McSherry F, Nissim K, Smith A (2006) Calibrating noise to sensitivity in private data analysis. In: Proceedings of the 3rd conference on theory of cryptography (TCC), pp 265–284

Feldman D, Fiat A, Kaplan H, Nissim K (2009) Private coresets. In: Proceedings of the 41st annual ACM symposium on theory of computing (STOC), pp 361–370

Friedman A, Schuster A (2010) Data mining with differential privacy. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining, pp 493–502

Hay M, Rastogi V, Miklau G, Suciu D (Sep 2010) Boosting the accuracy of differentially private histograms through consistency. Proc VLDB Endow 3(1–2):1021–1032

Kifer D, Machanavajjhala A (2011) No free lunch in data privacy. In: Sellis TK, Miller RJ, Kementsietsidis A, Velegrakis Y (eds) ACM-SIGMOD conference, pp 193–204

Li N, Qardaji WH, Su D, Cao J (2012) Privbasis: frequent itemset mining with differential privacy. PVLDB 5(11):1340–1351

McSherry F, Mahajan R (2010) Differentially-private network trace analysis. In: Proceedings of the ACM SIGCOMM 2010 conference, pp 123–134

McSherry F, Talwar K (2007) Mechanism design via differential privacy. In: Proceedings of the 48th annual IEEE symposium on foundations of computer science (FOCS), pp 94–103

Mohammed N, Chen R, Fung BCM, Yu PS (2011) Differentially private data release for data mining. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining

Monreale A, Andrienko GL, Andrienko NV, Giannotti F, Pedreschi D, Rinzivillo S, Wrobel S (2010) Movement data anonymity through generalization. Trans Data Priv 3(2):91–121

Rastogi V, Nath S (2010) Differentially private aggregation of distributed time-series with transformation and encryption. In: SIGMOD, pp 735–746

Samarati P, Sweeney L (1998a) Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppresion. In: Proceedings of the IEEE symposium on research in security and privacy, pp 384–393

Samarati P, Sweeney L (1998b) Generalizing data to provide anonymity when disclosing information(abstract). In: Proceedings of the 17th ACM symposium on principles of, database systems (PODS)

Terrovitis M, Mamoulis N (2008) Privacy preservation in the publication of trajectories. In: Proceedings of the 9th international conference on mobile data management (MDM)

Xiao X, Wang G, Gehrke J (Aug 2011) Differential privacy via wavelet transforms. IEEE Trans Knowl Data Eng 23(8):1200–1214

Xu J, Zhang Z, Xiao X, Yang Y, Yu G (2012) Differentially private histogram publication. In: ICDE, pp 32–43

Yarovoy R, Bonchi F, Lakshmanan LVS, Wang WH (2009) Anonymizing moving objects: how to hide a mob in a crowd? In: EDBT, pp 72–83

Moving and Calling: Mobile Phone Data Quality Measurements and Spatiotemporal Uncertainty in Human Mobility Studies

Corina Iovan, Ana-Maria Olteanu-Raimond, Thomas Couronné and Zbigniew Smoreda

Abstract In the past few years, mobile network data are considered as a useful complementary source of information for human mobility research. Mobile phone datasets contain massive amount of spatiotemporal localization of millions of users. The analyze of such huge amount of data for mobility studies reveals many issues such as time computation, users sampling, spatiotemporal heterogeneities, semantic incompleteness. In this chapter, two issues are addressed: (1) location sampling aiming at decreasing computation time without losing useful information on the one hand and to eliminate data considered as noise in the other hand and (2) users sampling whose goal is to select users having relevant information. For the first issue two measures allowing eliminating redundant information and ping-pong positions are proposed. The second issue requires the definition of a set of measures allowing estimating mobile phone data quality. New methods to qualify mobile phone data at local and global level are proposed. The methods are tested on one-day mobile phone data coming from technical mobile network probes.

1 Introduction

Human mobility analysis is an important issue in social sciences, and mobility data are among the most sought-after sources of information in economic forecasting, geography, transportation engineering and urban planning. Mobility studies have received increasing attention in the past few years, particularly the ones conducted on mobile phone location data (González et al. 2008; Song et al. 2010;

C. Iovan (✉) · A.-M. Olteanu-Raimond · T. Couronné · Z. Smoreda
Sociology and Economics of Networks and Services department,
Orange Labs R&D, Paris, France
e-mail: corina.iovan@orange.com

A.-M. Olteanu-Raimond
Laboratoire Cogit, Institut Géographique National, Saint-Mandé, France

D. Vandenbroucke et al. (eds.), *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography,
DOI: 10.1007/978-3-319-00615-4_14,

Andrienko et al. 2010; Onnela et al. 2011; Calabrese et al. 2011a, b; Olteanu Raimond et al. 2012; Becker et al. 2013). Based on mobile network data, different aspects can be tackled: collective spatial and temporal mobility patterns, patterns of travel behavior at the individual level, collective behavior at large scales, trajectory pattern mining or trajectory clustering (Song et al. 2010; Vieira et al. 2010). Human mobility models were successfully applied in various topics such as activities detection (Olteanu Raimond et al. 2012; Phithakkitnukoon et al. 2010), human mobility prediction (Song et al. 2010), tourism applications (Olteanu Raimond et al. 2011; Steenbruggen et al. 2011), traffic estimation (Phithakkitnukoon et al. 2010; Caceres et al. 2007) or commuting patterns (Zhang et al. 2010; Sevtsuk and Ratti 2010; Yuan et al. 2011).

Such datasets, daily collected by cellular service providers for billing and troubleshooting purposes and appropriately anonymized for privacy, contain massive amount of spatiotemporal localization of millions of users. The analysis of such a large amount of data is very costly from the time computation point of view. Moreover, they are often incomplete and/or have heterogeneous spatiotemporal resolutions. Thus, the use of mobile phone data for socioeconomic forecasting, in general, and human mobility studies, in particular, requires the sampling of data according to some selection criteria to create a subsample of statistically sound records, the definition of different assumptions to improve data by adding semantic information or the definition of adapted models to infer human behavior.

In this chapter, we focus on sampling issues. Two aspects are considered. The first concerns location sampling which consists of eliminating locations considered as redundant, on the one hand, and erroneous (i.e., locations produced by ping-pong phenomenon), on the other hand. Location sampling improves both computation time by eliminating redundant and erroneous locations but also mobile phone data quality by detecting and eliminating locations which could introduce a bias. The second aspect concerns user sampling, i.e., how to select users described by data having a minimal quality required to mobility analysis? We propose a user sampling approach based on the definition of a set of measures allowing estimating mobile phone data quality.

Mobile phone data quality estimation has already been studied in the literature. For example, mobile phone individual trajectories were compared with actual individual trajectories provided from GPS (Kang et al. 2012; Schulz et al. 2012) or with data access records (Ranjan et al. 2012; Zhao et al. 2011) in order to determine the bias and the characteristics of human mobility when mobile phone data are used. Measures allowing characterization of individual trajectory (e.g. length, travel distance, direction) (Andrienko et al. 2011; Hasan et al. 2012) or the territory of trajectories such as entropy (Song et al. 2010; Ranjan et al. 2012), eccentricity (Kang et al. 2012), radius of gyration (González et al. 2008; Song et al. 2010; Ranjan et al. 2012) and convex hull were proposed end tested (Csáji et al. 2012). Although the high number of measures proposed in the literature, the list is not exhaustive and these measures do not cover the temporal aspect of data, which is of great importance when studying location and user sampling.

The chapter is structured as follows. In the Sect. 2, mobile phone data are briefly described. Section 3 introduces the proposed measures to filter locations. In Sect. 4 user sampling is discussed. We first present the impact that user sampling methods could have on human mobility studies by analyzing the correlation between the communication and itinerancy events. Then, the new measures to qualify mobile phone data (local and global measures) and a decision process allowing to chose "representative" users (e.g. user sampling) are described. Section 5 concludes and suggests some directions for future work.

2 Mobile Phone Data

Each telecom operator collects and stores for a given period customers' mobile phone activities for billing or for technical measurement purposes. This type of collection is called "passive collection", since recordings are made automatically. There are three main types of mobile phone data collected through "passive collection": Call Detail Records (CDR) data representing cell phone billing records, probe data and Wi-Fi data. In this chapter, only probe data are briefly described, since these data are used to test the proposed measures. For more information about mobile phone data for human mobility studies, interested readers can refer to (Smoreda et al. 2013).

Probe data are issued from mobile network probes (MSC data), they are anonymous (each mobile phone SIM identifier is replaced with an identifier consisting in a unique integer) and contain both cell localized communication events (i.e. calls and SMS) and itinerancy events: handover (HO) and location area update (LAU). The location of mobile phone users is limited to the base station location. The base station is composed by at least three antennas, each antenna having a spatial coverage. Figure 1 shows an example of mobile phone localization according to different events that occur.

HO data are generated while an active communication is transferred from one cell of the mobile network to another, when a mobile device is on the move; LAU records are generated when a device changes location area, even if the user is not in communication (for Paris, a location area groups on average 150 cells).

In this chapter, anonymous data from a French cellular network operator are used. Spatially, it covers the Parisian region (12,012 km^2–4,638 m^2), and contains recordings of one weekday (Thursday, the 2nd of April 2009) of over 4 million mobile phone users (accounting for a total of 122,208,870 records).

3 Location Sampling

The goal of location sampling is to remove location points considered as "noise". Two such cases are considered here, duplicate location points (Sect. 3.1) and location points generated by the ping-pong phenomenon (Sect. 3.2).

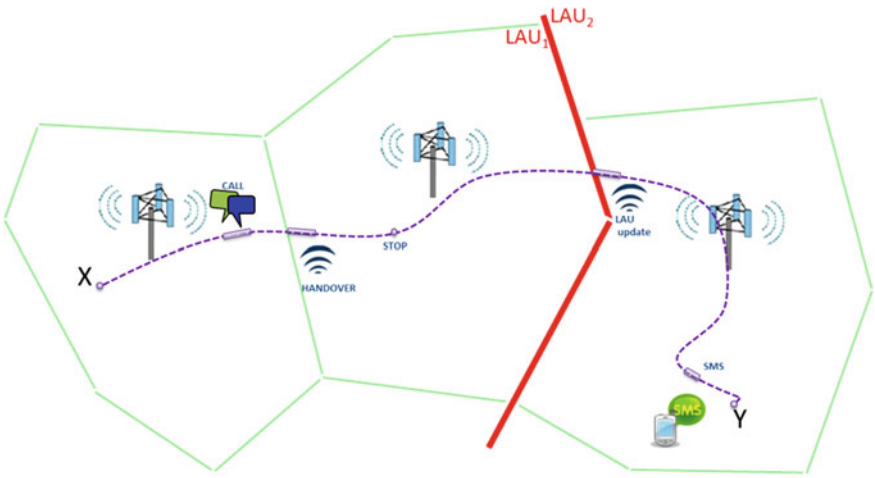


Fig. 1 An example of mobile phone network localization data types for one user travelling from X to Y

Location sampling is especially important when dealing with a huge number of records and consists in an efficient and lossless data reduction strategy. We propose to apply location sampling user by user, thus we use the concept of trajectory for a better understanding of the proposed measures and notations.

A user trajectory is defined as a set of locations in time and space. For each record, a geometry point in a geographic coordinate system is added. In the remainder of this article, by language abuse, the term point will be used to identify a user's location in cartographic coordinates.

Let t_j be the trajectory of user u_i defined as a sequence of points: $p_k \in P$, $P = \{p_1, p_2, \ldots, p_n\}$. A simplified illustration of a user's trajectory is given in Fig. 2 where a user's points p_k, defined by the location of the base station (latitude, longitude) and timestamp (T) serving an event (call, SMS, itinerancy), are sequentially connected to form a trajectory illustrated on a 2D plan.

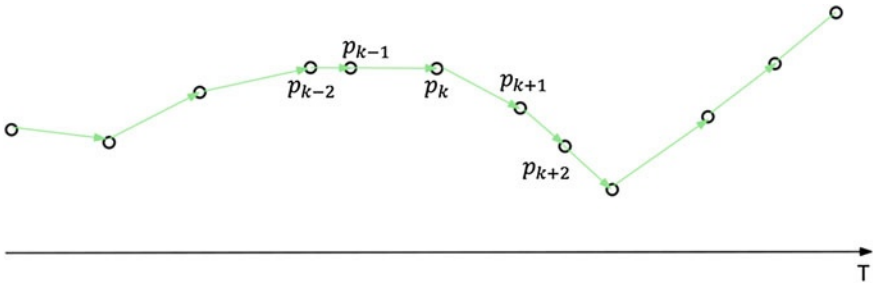


Fig. 2 Creating a user's trajectory t_j from successive location points p_k

3.1 Duplicated Location Points

Let's consider n consecutive locations, having the same geographic coordinates and recorded in a time interval below a threshold. We consider that these duplicate locations do not provide additional information. We agree that locations do not represent the absence of movement, but the lack of more detailed information about the movement. Our assumption is that by keeping one location point and eliminating the other (n − 1 location points) the temporal information about the presence at that geographic position is not lost and they are not essential to be used as an indicator of the a user's stationary state. This assumption is true, only and only if, the time interval is small. We propose a threshold equal to one minute. In this way, the computation process can be faster when individual approaches are carried out. The location is represented by a point defined by its geographic coordinates at a given time.

The set of points removed during this step is given by Eq. 1:

$$\{p_k | (T(p_k) - T(p_{k-1})) \leq 1 \text{ min } \& \; D(p_k - p_{k-1}) = 0 \text{ m}\} \tag{1}$$

where $T(p_k)$ is the timestamp recorded for point p_k and $D(p_k - p_{k-1})$ is the distance between two consecutive points. The percentage of duplicate points (dp) removed is thus given by Eq. 2:

$$dp = \frac{card(p_k)}{n} * 100 \tag{2}$$

where n stands for the total number of points and $card(p_k)$ stands for the cardinality of the set of points removed after the duplicate point filtering step. For one day of data consisting of a total of 122,208,870 points, 24 % of them are eliminated after this filtering step.

3.2 Ping-Pong Points

Since a mobile is connected to the cell providing the best coverage, a change in the cell to which the mobile connects occurs in time. When the mobile is located at cell edge or at the border between two location areas, it might give rise to the "ping-pong handover": within a short period of time (less than 10 s), the mobile switches back to the old cell, fluctuating between the two neighboring antennas. This phenomenon can also occur when an equal intensity strength signal is received from two or more base stations, when the mobile is located in a coverage hole or in areas shadowed by high buildings.

The ping-pong phenomenon has been studied by the cellular network research community, which proposed several approaches to tackle such events (Gudmundson 1991; Pollini 1996). Handover algorithms can be decomposed in two steps, initiation and decision. The first one consists in deciding when to request a handover while the

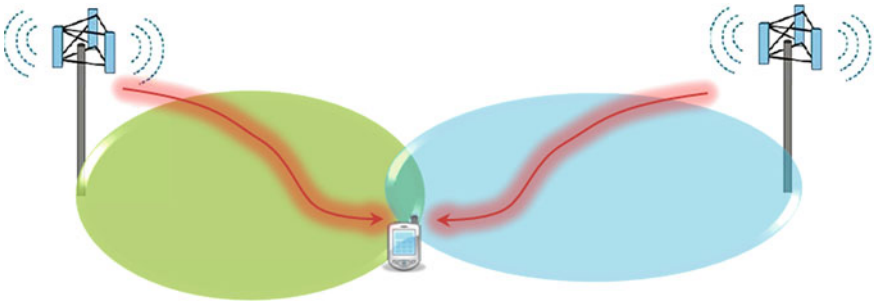


Fig. 3 The "ping-pong" handover phenomenon

latter one is based on signal strength comparison between the current and neighboring base station. Several handover types based on channel usage, microcellular and multilayered systems and network characteristics have been analyzed by state of the art researches [interested readers are invited to refer to (Ekiz et al. 2005)]. Most existing solutions focus on improving the decision in handover algorithms, which consists in comparing the differential signal power level between the serving and target base stations to a constant threshold value. This is mainly achieved by increasing this hysteresis threshold, designed to reduce the ping-pong effect during handover. Some methods propose handover algorithms performing at a sub-cell level, providing a more precise location of the mobile handset inside the cell (Feher et al. 2012). Other solutions take into account different types of location information to assist in the reduction of unnecessary handovers (He et al. 2010).

In human mobility studies based on cellular network traces, user location points issued from the ping-pong phenomenon are considered as noise. Since user trajectories are considered from a set of location points, noise should be filtered by a data pre-processing step, which operates on mobile handset communication logs.

Recently, in (Haoyi et al. 2012) such noise points were filtered out by mapping network cells to non-overlapping regions and identifying each region by the full set of cell towers covering the region. Then, user trajectories are considered by taking into account location points from the region having the longest hourly stay.

This produces anomalous points in a user trajectory as the users' handset is registered either with one LAU or the neighbor one, as illustrated in the simplified example in Fig. 3 which depicts a part of a network made of two base transceiver stations (BTS) each providing a coverage area (illustrated by oval shapes) for mobile stations. A mobile handset close to coverage borders of both BTS, is served by the left BTS (green cell) first, but it is then attached to the right BTS (blue cell), then switched back to the left BTS, and so on in a short time interval. The mobile device is stationary but in the data we observe erratic movements between the two positions.

In Fig. 4, the approach that allows detecting the ping-pong handover is illustrated. Points p_k (illustrated by circles) belonging to a user's trajectory (depicted by

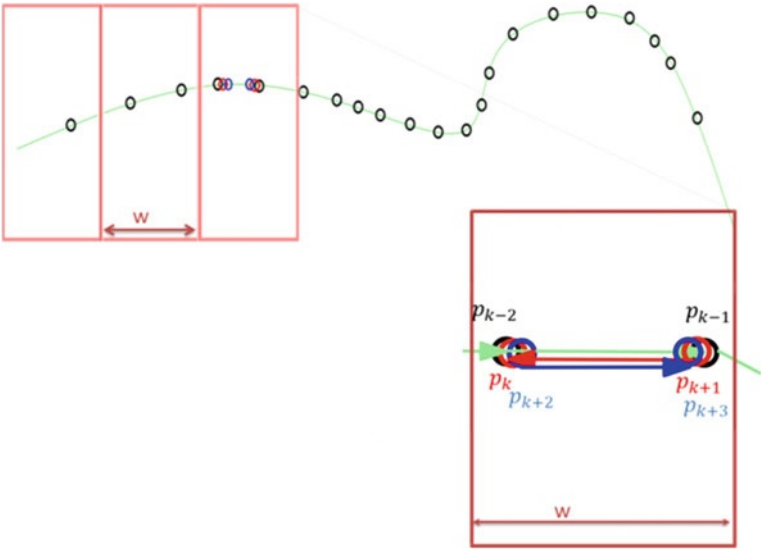


Fig. 4 The ping-pong handover detection

a continuous green line) are analyzed during a sliding spatial and temporal interval of width *w* (depicted by the rectangle of size w) in the upper left image and excerpt of successive points considered during the analysis window (in the lower right image). The ping-pong handover appears between points p_{k-2} and p_{k+3}, which are the points which should be considered in the user's trajectory as the first and the final points of the trajectory in w.

To identify the ping-pong handover (pp_{HO}) phenomenon, we define a sliding window of size *w* and analyze successive points of a trajectory belonging to the analysis window, denoted *W* in the following.

Between two consecutive points, p_{k-1}, and p_k, we compute the spatial distance $D(p_k - p_{k-1})$ and the temporal interval $T(p_k) - T(p_{k-1})$. Subsequently, the velocity of moving from a location point p_{k-1} to the following location point p_k is computed as follows:

$$v_{p_k} = \frac{D(p_k - p_{k-1})}{T(p_k) - T(p_{k-1})} \tag{3}$$

The heading direction (Zheng et al. 2008), $h_{p_{k-1}}$ is computed between the successive points, by considering North as the basis of the heading direction (cf. simplified illustration in Fig. 5). After computing pairwise heading direction between all points from a user's trajectory, we compute the heading change (Zheng et al. 2008) as:

$$hc_{p_k} = \left| h_{p_{k-1}} - h_{p_k} \right| \tag{4}$$

to identify location points having a heading change of 180°.

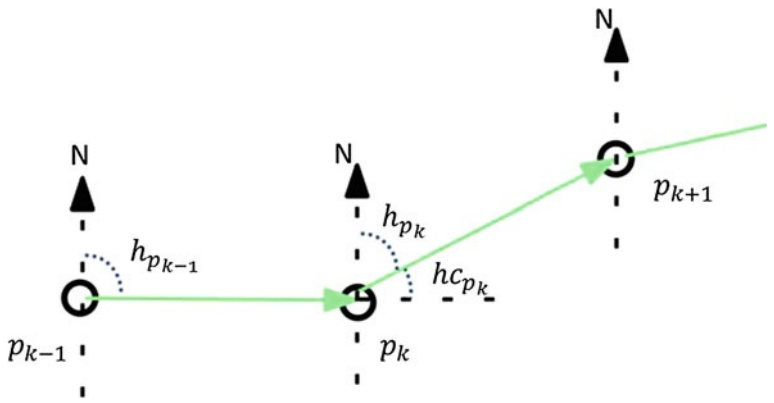


Fig. 5 Computing heading direction and heading change between three successive points from a user's trajectory

For all points inside the analysis window W, we pairwise compute velocity and heading change between successive points. The decision if a point is issued from a ping-pong phenomenon is taken based on the velocity between successive points and the heading change value. If the velocity is higher than a threshold, and the heading change is equal to 180°, the points are discarded as considered issued from a ping-pong handover phenomenon. The value of this threshold is empirically set at 200 km/h in our experiments.

The set of points issued from a ping-pong phenomenon is given by Eq. 5:

$$\{p_k | v_{p_k} > 200 \text{ km/h} \quad \& hc_{p_k} = 180^\circ\} \tag{5}$$

To assess the validity of the proposed approach for ping-pong point detection, we computed pairwise velocity, heading and heading change for a subset of 10 million points of the entire dataset of 122,208,870 points. Figure 6 below, shows the cumulative distribution function of points with a velocity between 0 and 500 km/h before (blue line) and after (red line) ping-pong filtering. It highlights the fact that over 90 % of the points belong to users moving with a speed lower than 100 km/h (which would correspond to vehicles or motorbikes usually used to move in urban areas). The effect of the ping-pong filtering is illustrated by the red line and shows that point velocity values belong after the filtering to lower bounds. Figure 7 illustrates point velocity plotted against the heading change value. The ping-pong phenomenon is shown in Fig. 7 through the straight peak at 180° which would correspond to the start of a handover phenomenon.

After the point filtering procedures a total of 40 % of points have been discarded: 24 % were due to duplicate points filtering and the remaining 16 % were points issued from ping-pong phenomenon.

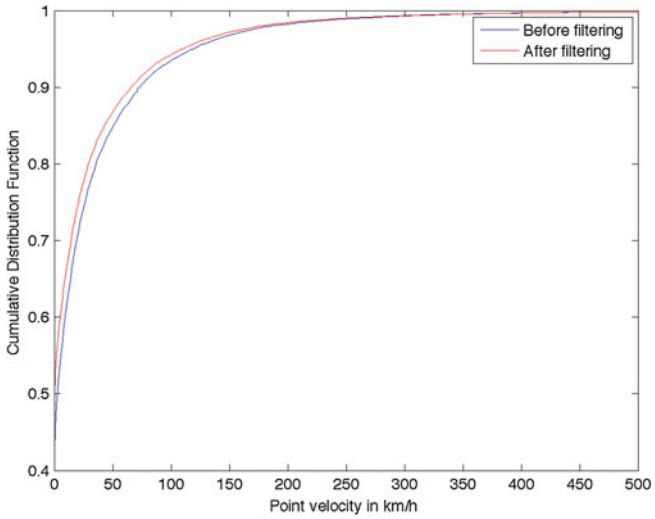


Fig. 6 Cumulative distribution function points velocity

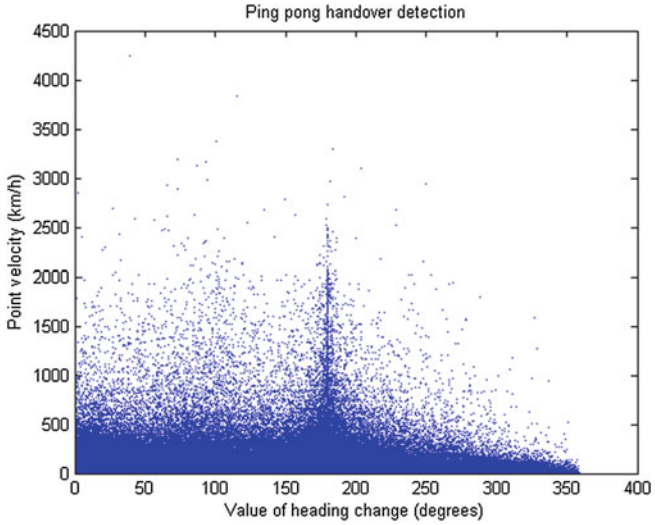


Fig. 7 Detecting ping-pong handover. A peak at 180° in the value of heading change marks the ping point phenomenon

4 User Sampling

The aim of this section is to introduce several quality measures for mobile phone data in order to sample users considered as reliable for a given application.

4.1 User Sampling Issue

Since, the set of observed locations for a user is dependent of his communication activity and/or the telecom network, the data describing the daily user activity are incomplete and heterogeneous. To overcome this drawback, some researchers (González et al. 2008) select randomly a sample of users from their dataset. Others try to optimize this method, by taking into account only users with a high number of recorded events (Song et al. 2010; Onnela et al. 2011). According to the desired application, the criteria could be to use only users having at least 0.5 calls per hour (Song et al. 2010), or users having records during each day for the study period. While this approach seems statistically sound (more user location points makes the analysis more precise), location points are generated whenever a communication or a LAU event is recorded. Thereby, user locations are biased, as they are depending of the users' calling frequency, mobility and the operator network. These evident shortcomings in estimating user mobility (Andrienko et al. 2012) have recently been raised in (Zhao et al. 2011; Couronné et al. 2011; Tiru and Ahas 2012).

In this section, we propose to estimate the correlation between the user's communication and itinerancy to better assess the bias of users sampling based on frequency activity. Using probe data, the relationship between communication and mobility patterns is studied. Analysis was conducted on two types of data: communication (voice calls, SMS) and itinerancy records. Figure 8 depicts communication frequency plotted against the median number of mobility records (LAU) for each user.

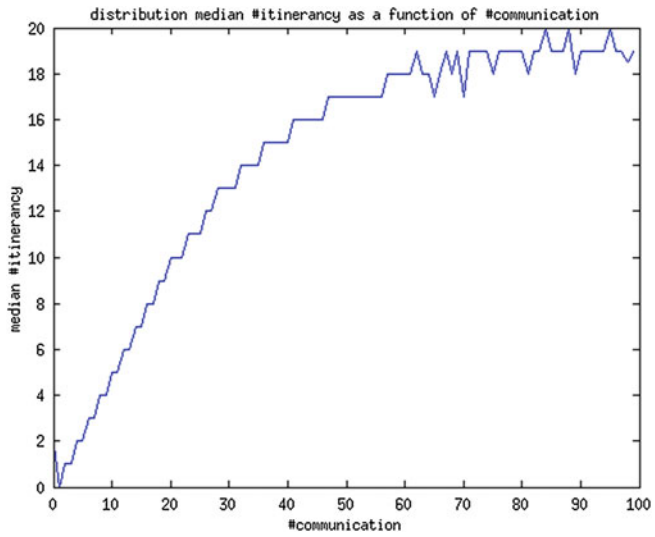


Fig. 8 Median number of local area change as a function of communication frequency

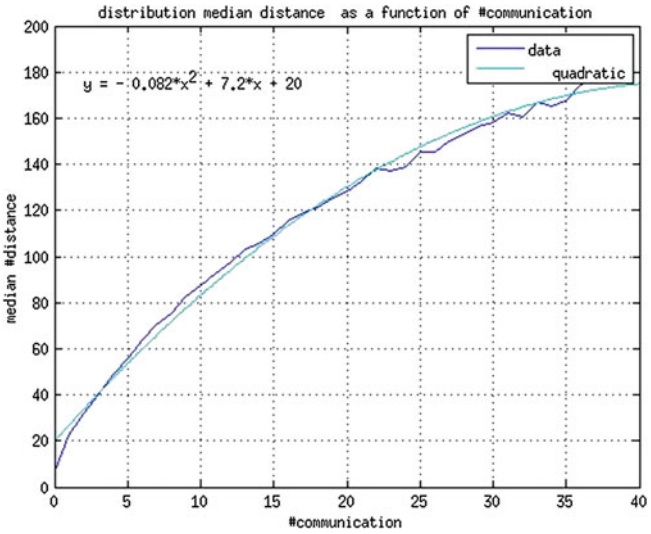


Fig. 9 Median daily distance traveled (km) as a function of communication events frequency

Ninety percent of users have less than 30 communication events (calls or SMS) during the observed day. For this group, we notice a clear, almost linear correlation between the frequency of communications and the median number of location area changes (daily mobility indicator).The curve reaches a plateau at about 50 communications per day and then the communication-itinerancy link disappears. People who communicate extremely frequently can no longer be distinguished by their median itinerancy.

To complete this study, the daily traveled distance (i.e. the Euclidian distance between locations defining the daily trajectory) per user was computed using all localized records (calls, SMS, HO, and LAU) and compared to communication events distribution (see Fig. 9).

We looked for a regression model to fit our data, where y is the median daily distance in km and x is the number of communication events (call, SMS). It appears that the best model is a quadratic function:

$$y = -0.082\,x^2 + 7.2\,x + 20 \tag{6}$$

This analysis confirms our observation showed in Fig. 9: the higher the number of communication events, the less the cumulative travelled distance increases.

A significant correlation between user mobility events and communication frequencies confirms our intuition that in mobile phone usages both phenomena are interrelated. A highly mobile person has in fact greater probability to use a mobile phone that someone who only commutes between a few places where s/he can also communicate *via* a landline telephone, VoIP, etc. In the same way, the higher mobility in the city context is frequently associated with a distant coordination via a mobile phone (Diminescu et al. 2009), and the mobile communication

is also linked to a management of the mobility itself: delays, traffic problems, last minute adjustments. Finally, correspondents of a highly mobile person learn with time which is the most adapted communication channel to reach this person, they will also contribute to reinforce the observed correlation.

From the point of view of human mobility analysis based on mobile phone data, the obtained results show that a careful examination of the sampling methods is of high importance. Selecting users with frequent communication traces, i.e., with many cell localizations, seems to introduce a clear bias because people having more mobile communications are also in a more mobile class of the general population.

In this context, the following questions arise: Which measures can be defined to obtain a statistically representative sampling of general human mobility patterns? How to choose the relevant measures for a specific application? Our approach is to use not only the frequency of records to sample representative users but also criteria that take into account data precision, accuracy or resolution. Moreover, data quality is a relative concept having different meanings to different consumers: data which is good enough for one user/application might not be of acceptable quality for another one. This is why, our approach consists in defining a set of different measures that qualify mobile phone data and then to use one or to combine different measures to sample users, depending upon the nature and objectives of the study.

4.2 Local and Global Measures

As stated in Sect. 3, individual trajectories are defined for each day and for each user. After performing location sampling, for each location-point composing a user's valid trajectory, local quality measures (Sect. 4.2.1) are first defined and further global ones (Sect. 4.2.2) are introduced. Finally, according to the application purposes these measures can be combined to estimate reliability of user's data.

4.2.1 Local Measures

In this section we define some measures that allow estimating position accuracy and precision for each user at local (location point) level. To do this, we propose to characterize pairwise consecutive points.

First, we define a speed index (named theta and noted θ) which describes user mobility between two sequential records, $n-1$ and n, considering laps-time $\Delta t = t(n) - t(n-1)$ and distance between the location of two sequential records far from each other from $d = \mathrm{Dist}(n-1,\ n)$. Note that measures defined in Eqs. 7 and 8 were proposed and tested in (Couronné et al. 2011), but they are explained here for better understanding.

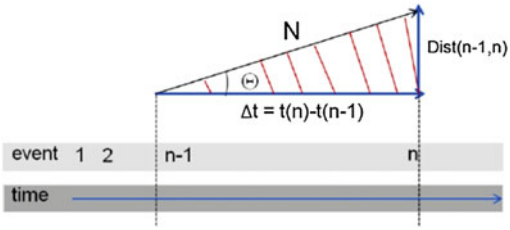


Fig. 10 Speed and uncertainty estimation

As we show in Fig. 10, θ is computed by the following equation:

$$\theta = \arctan \frac{d}{\Delta t} \tag{7}$$

The second index named "uncertainty" and noted U assesses whether θ estimation is confident or if this value has only a mathematical ground. Thus, the "uncertainty" reflects how confident the measured mobility state is; it is defined by the norm's vector having θ angle:

$$U = \frac{\Delta t}{\cos(\theta)} \tag{8}$$

The uncertainty estimation is related to the entropy concept. Thus, considering an entropy approach, we define a quality indicator, noted Q which measures the spatial accuracy. The more we are confident about our measurement, the less the entropy increases: the more probable it will be that we will find the user in a given location.

$$Q = e^{-\frac{\theta}{\frac{\pi}{2}} * U} \tag{9}$$

If uncertainty or theta increases then the quality Q of the measurement decreases: the more Q is close to one the better the confidence is. Q is defined as an exponential so that it spans between 0 and 1.

Q describes the probability distribution function to find a user in a spherical 3 dimensional space (2 geographical and 1 temporal), knowing two spatiotemporal measurements. The more the distance in space and or time is large, the larger the probability function will be widely distributed.

From the speed estimation and its confidence, we can derive the probable successions of position of the user during the time between two records; the quality of the user mobility increases with sampling frequency.

4.2.2 Global Measures

While local measures qualify data at point-level for each user, global measures addressed in this section are meant to qualify data at a user's trajectory level. A trajectory is hence made of successive points (locations) visited by a user during a

time interval. The aim of the measures proposed hereafter is to evaluate the confidence in a users' trajectory created from mobility/communication mobile phone data. These measures allow selection of users according to the applications.

(a) **Number of events**

As a user's trajectory is made of location points, one of the most intuitive ways to qualify a user's trajectory is to count the number of events generating the trajectory. A threshold can be applied in order to select users having the number of points greater than the threshold. As we discussed in Sect. 4.1, this selection can introduce some biases in final results when the threshold is important. We notice that this measure has meaning if the threshold is small, for example less than three. Indeed, users having less than three points (per day or during the time period of study) are not considered as relevant.

(b) **Temporal activity spread**

During a day, a user's communication or mobility habits gives more or less mobile network records (see Fig. 11). The aim of this second measure is to find the period of the day when the user is most active in terms of number of mobility/communication records. This measure is of highest importance as it gives a hint on the temporal resolution related to a user's mobility: a user having a uniformly distributed temporal spread is more reliable when analyzing its trajectory than one having few recordings (for which the loss of information is high).

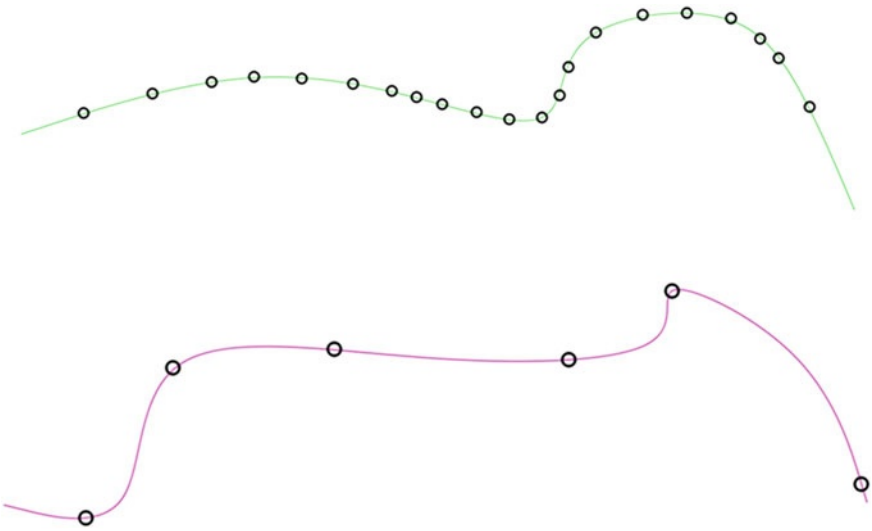


Fig. 11 Types of user trajectories. The *upper image* illustrates a reliable trajectory for mobility analysis as it is built upon a high number of points p_k (*illustrated by circles*) belonging to a user's trajectory. On the contrary, the trajectory depicted in the *lower image* is less reliable as it is built upon fewer points

We define the temporal activity spread of a user as the number of communication/mobility events recorded during the time span between the first and the last activity events generated by the user.

The number of recordings for a user is either given by communication or mobility events. To qualify a user's trajectory, it is crucial to establish if the recordings are given by the communication or itinerancy events. We introduce the event-rate (denoted R_e) as the number of events recorded for a user during the observation period. Let e be the number of events recorder for a user and Δt be the time interval during the first and the last recording hours considered for each user. Then the event-rate is given by the Eq. 10:

$$R_e = \frac{e}{\Delta t} \tag{10}$$

Similarly, we also define the voice activity rate (R_{va}) and the itinerancy rate (R_{vi}). The R_{va} is defined as the number of user-generated activity events (voice calls, SMS) during the time interval when the user is active while the R_{vi} is computed as the number of user-itinerancy events during the time span when the user is observed.

(c) **User based spatio temporal entropy measure**

Using the local measure defined in Sect. 4.1, we define a global one by computing for each user based on Shannon entropy and named "user based spatio temporal entropy" (UBSTE).

Let n be the number of records of the user u_i. The entropy $h(u_i)$ of the user u_i on a given number n of records is computed as follows:

$$h(u_i) = -\frac{1}{n}\sum_{j=1}^{n-1} Q_j * \log(Q_j) \tag{11}$$

where Q_j is the quality indicator, normalized w.r.t. $\sum_{j=1}^{n-1} Q_j = 1$.

The entropy describes the quantity of information we have about the state of the observed system. It increases when the quality of measurement decreases.

This indicator can be used as a filter to exclude sequences of records (user trajectories) having high entropy, so weak global spatiotemporal measurement quality. Traditionally, to reject users, the mean and the standard deviation of the entropy computed for the total number of users are used.

Thus, using this measure a user u_i is rejected if the next condition is true:

$$h(u_i) > \bar{h} + \sqrt{\frac{1}{N}\sum_{k=1}^{N} (h(u_k) - \bar{h})^2} \tag{12}$$

where $\bar{h} = \frac{1}{N}\sum_{k=1}^{N} h(u_k)$ represents the mean value of the entropy, and N represents the total number of users.

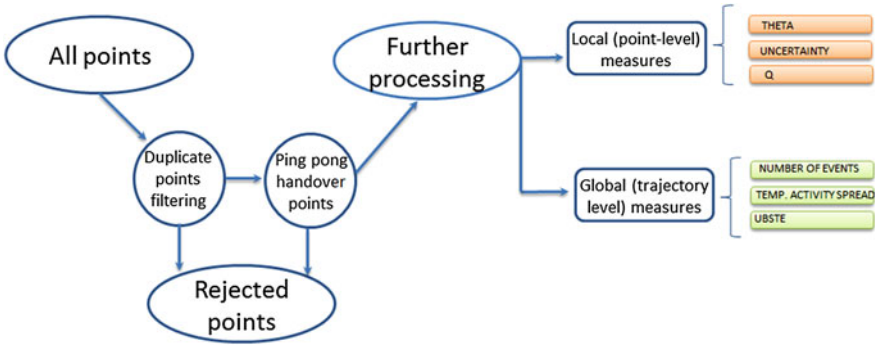


Fig. 12 Decision system for points filtering and user's selection

4.3 Decision Making

As the use of mobile network data reveals caveats for human mobility analysis, the measures proposed in this article can be handy when it comes to quantifying the bias they introduce. Such measures can be combined in cascade of decision systems taking into account the proposed measures depending on the nature and objectives of the studies considered, as illustrated in Fig. 12.

5 Conclusion

This chapter addresses location sampling and user sampling in mobile phone data. We propose two methods to handle location sampling: the first one consists of removing redundant location points (generally caused by rejected communications or SMS exchange in a very small time interval) and the second one allows the detection prior to discarding of erroneous locations points (caused by the ping-pong effect). Furthermore, we propose global and local quality measures meant to qualify mobile phone data for user sampling, a process which depends upon the nature and objectives of the study. Results of our study show a correlation between mobility span and communication frequency. This means that when working with CDR data, user selection has to be carefully performed, as random sampling is not efficient. Moreover, the mobility behavior seems to be associated with increased mobile communication, and billing data are generated only when a communication event is available. Thus, the data of more active users are of better quality (more points), but those users' mobility is also different. This obviously can cause serious problems in the mobile phone data based analysis of human mobility.

The measures proposed here should be seen as complementary measures to other existing state of the art measures. These measures can be integrated in a decision system aiming at defining the most exhaustive set of measures.

Future work in this area first includes testing our methods on different mobile phone datasets (more than one day, and covering different spatial areas such as dense urban areas, less dense urban areas, rural areas). To validate the proposed methods we would like to experimentally test these methods on a long period of time for different applications (O/D matrix, tourists behavior analysis, detection of the mean of transportation) and compare the obtained results to previous results (without using location and user sampling).

Second we wish to explore more ways to combine such measures in order to get the most reliable information out of the data. One of the improvements could be the design of a reliable decision system including the proposed measures.

Acknowledgments We would like to thank our colleague, Cezary Ziemlicki, who preprocessed data and has discussed with us many technical issues related to this chapter.

References

Andrienko G, Andrienko N, Bak P, Bremm S et al (2010) A framework for using self-organising maps to analyse spatio-temporal patterns, exemplified by analysis of mobile phone usage. J Locat Based Serv 4(3–4):200–221

Andrienko G, Andrienko N, Bak P, Keim D, Kisilevich S, Wrobel S (2011) A conceptual framework and taxonomy of techniques for analyzing movement. J Visual Lang Comput 22(3):213–232

Andrienko G, Andrienko N, Hurter C, Rinzivillo S, Wrobel S (2012) Scalable analysis of movement data for extracting and exploring significant places. In: Proceedings of IEEE transactions on visualization and computer graphics

Becker R, Cáceres R, Hanson H, Isaacman S, Loh JM et al (2013) Anonymous location data from cellular phone networks sheds light on how people move around on a large scale. Commun ACM 56(1):74–82

Caceres N, Wideberg J, Benitez F (2007) Deriving origin-destination data from a mobile phone network. Intel Trans Syst IET 1(1):15–26

Calabrese F, Di Lorenzo G, Liu L, Ratti C (2011a) Estimating origin-destination flows using opportunistically collected mobile phone location data from one million users in Boston metropolitan area. IEEE Pervasive Comput 10(4):36–44

Calabrese F, Smoreda Z, Blondel V, Ratti C (2011b) Interplay between telecommunications and face-to-face interactions—a study using mobile phone data. PLoS One 6(7):e208814

Couronné T, Smoreda Z, Olteanu AM (2011a) Chatty mobiles: individual mobility and communication patterns. NetMob, Boston

Couronné T, Olteanu AM, Smoreda Z (2011) Urban mobility: velocity and uncertainty in mobile phone data. In: Proceedings of Privacy, security, risk and trust (passat), 2011 IEEE third international conference on and 2011 IEEE third international conference on social computing (socialcom), pp 1425–30

Csáji BC, Browet A, Traag VA, Delvenne JC, Huens E et al (2012) Exploring the mobility of mobile phone users. Phys A 392(6):1459–1473

Diminescu D, Licoppe C, Smoreda Z, Ziemlicki C (2009) Tailing untethered mobile user: studying urban mobilities and communication practices. In: Ling R, Campbell SW (eds) The reconstruction of space and time. Mobile communication practices. Transaction Publishers, New Brunswick, NJ, pp 17–37

Ekiz N, Salih T, Kucukoner S, Fidanboylu K (2005) An overview of handoff techniques in cellular networks. Int J Inf Technol 2(3):132–136

Feher Z, Veres A, Heszberger Z (2012) Ping-pong reduction using sub cell movement detection. In: Proceedings of vehicular technology conference (VTC Spring)

González MC, Hidalgo CA, Barabási AL (2008) Understanding individual human mobility patterns. Nature 453:779–782

Gudmundson M (1991) Analysis of handover algorithms. In: Proceedings of 4th IEEE vehicular technology conference, gateway to the future technology in motion

Haoyi X, Zhang D, Zhang D, Gauthier V (2012) Predicting mobile phone user locations by exploiting collective behavioral patterns. In: Proceedings of the 9th IEEE conference on ubiquitous intelligence and computing (UIC'12), Fukuoka, Japan, 2012

Hasan S, Schneider CM, Ukkusuri SV, González MC (2012) Spatiotemporal patterns of urban human mobility. J Stat Phys 151:304–318

He D, Chi C, Chan S, Chen C, Bu J, Yin M (2010) A simple and robust vertical handoff algorithm for heterogeneous wireless mobile networks. Wireless Pers Commun 59(2):361–373

Kang C, Liu Y, Mei Y, Xu L (2012) Evaluating the representativeness of mobile positioning data for human mobility patterns. GIScience, Columbus

Olteanu Raimond AM, Trasarti R, Couronne T, Giannotti F, Nanni M et al.(2011) GSM data analysis for tourism application. In: Proceedings of 7th international symposium on spatial accuracy assessment in natural resources and environmental sciences

Olteanu Raimond AM, Couronne T, Fen-Chong J, Smoreda Z (2012) Le Paris des visiteurs, qu'en disent les téléphones mobiles? Inférence des pratiques spatiales et fréquentations des sites touristiques en Ile-de-France. Revue Internationale de la Géomantique 3:413–437

Onnela JP, Arbesman S, González MC, Barabási AL, Christakis NA (2011) Geographic constraints on social network groups. PLoS One 6(4):e16939

Phithakkitnukoon S, Horanont T, Di Lorenzo G, Shibasaki R, Ratti C (2010) Activity-aware map: identifying human daily activity pattern using mobile phone data. In: Proceedings of international conference on pattern recognition, Workshop on human behavior understanding, pp 14–25

Pollini GP (1996) Trends in handover design. IEEE Commun Mag 34(3):82–90

Ranjan G, Zang H, Zhang Z, Bolot J (2012) Are call detail records biased for sampling human mobility? ACM SIGMOBILE Mob Comput Commun Rev 16(3):33–44

Schulz D, Bothe S, Körner C (2012) Human mobility from GSM data-a valid alternative to GPS? Mobile data challenge 2012 workshop, June 18–19, Newcastle, UK

Sevtsuk A, Ratti C (2010) Does urban mobility have a daily routine? Learning from the aggregate data of mobile networks. J Urban Technol 17(1):41–60

Smoreda Z, Olteanu-Raimond AM, Couronné T (2013) Spatiotemporal data from mobile phones for personal mobility assessment. In: Zmud J, Lee-Gosselin M, Carrasco JA, Munizaga MA (eds) Transport survey methods: best practice for decision making. Emerald Group Publishing, London

Song C, Qu Z, Blumm N, Barabási AL (2010) Limits of predictability in human mobility. Science 327:1018–1021

Steenbruggen J, Borzacchiello MT, Nijkamp P, Scholten H (2011) Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities. GeoJournal 78:223–243

Tiru M, Ahas R (2012) Passive anonymous mobile positioning data for tourism statistics. In: Proceedings of 11th global forum on tourism statistics, Iceland

Vieira MR, Frias-Martinez E, Bakalov P, Frias-Martinez V, Tsotras VJ (2010) Querying spatio-temporal patterns in mobile phone-call databases. In: Proceedings of eleventh international conference on mobile data management (MDM), pp 239–248

Yuan Y, Raubal M, Liu Y (2011) Correlating mobile phone usage and travel behavior—a case study of Harbin, China. Comput Environ Urban Syst 36(2):118–130

Zhang Y, Qin X, Dong S, Ran B (2010) Daily O-D matrix estimation using cellular probe data. In: Proceedings of 89th annual meeting transportation research board

Zhao N, Huang W, Song G, Xie K (2011) Discrete trajectory prediction on mobile data. In: APWeb'11 Proceedings of the 13th Asia-Pacific web conference on web technologies and applications, pp 77–88

Zheng Y, Li Q, Chen Y, Xie X, Ma WY (2008) Understanding mobility based on GPS data. In: Proceedings of the 10th international conference on ubiquitous computing

Spatial Accuracy Evaluation of Population Density Grid Disaggregations with Corine Landcover

Johannes Scholz, Michael Andorfer and Manfred Mittlboeck

Abstract The article elaborates on the spatial disaggregation approach of the 1 km population density grid created by the European Forum for Geostatistics in a defined study area where accurate population reference data are available. The chapter presents an approach to disaggregate the population grid to target resolution of 100 and 500 m respectively and describes the evaluation methodology. The resulting population grids are evaluated with respect to the reference population dataset of the Austrian Bureau of Statistics. In addition, the results are evaluated regarding their correlation to the reference or a random population dataset. The results indicate that there is evidence that the disaggregated population grid with 500 m resolution is more accurate than the 100 m population grid. In addition, the 100 m disaggregated population raster shows more correlation with the random population grid. Furthermore, the chapter shows that densely populated zones are estimated with higher accuracy than medium and sparsely populated areas.

1 Introduction

The European Union provides population data, which is comprised of the national census data sets, of which a few are available to the public. The available population dataset created by the GEOSTAT 1A project of the European Forum for Geostatistics (EFGS) is a 1 km population grid that is hosted by EUROSTAT. Due to the fact that the GEOSTAT 1A population grid is very coarse for detailed simulation activities, a downscaling or disaggregation process is necessary in order to obtain population density data on a finer granularity level.

J. Scholz (✉) · M. Andorfer · M. Mittlboeck
Studio iSPACE, Research Studios Austria, Schillerstrasse 25, Salzburg, Austria
e-mail: johannes.scholz@researchstudio.at

D. Vandenbroucke et al. (eds.), *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography,
DOI: 10.1007/978-3-319-00615-4_15,

Downscaling is a well-known term in environmental studies, which describes the process of generating fine granular data from coarse base data (Bierkens et al. 2000; Reibel and Agrawal 2007). In order to generate a dataset of finer spatial granularity, auxiliary data are necessary that provide additional information on the spatial phenomena to be disaggregated. In this chapter, Corine Landcover data are employed to detect population densities as complementary source to the GEOSTAT 1A population grid. As 1 km population grid cells may contain parts with varying population density—e.g. dense urban zones or urban areas with parks having no population. By population grid cells which span over different population "density zones" the representation of those zones are blurred.

Corine Landcover data are chosen as auxiliary data due to their availability over Europe and the consistent semantics over Europe. This is of interest for the research project that forms the organizational frame for this chapter and research work. The research project aims at modeling and simulating socio-economical phenomena on a detailed level. The results of the study should be applicable in all member states of the European Union. Due to the fact that fine granular population data are not available for all European countries in a consistent manner, the population raster with 1 km resolution is employed as harmonized population dataset.

Application fields of fine granular population data are found in the assessment of natural disasters like floods or hailstorms (Tralli et al. 2005; Chen et al. 2004). Thiecken et al. (2006) used disaggregated population data in order to evaluate the population affected by the flooding in Germany of 1999 and 2002. Such population grids help to estimate the impact of noise on people living around airports (Vinkx and Visée 2008).

In literature downscaling methods have been discussed in depth. A number of publications elaborate on the disaggregation of data from a zonal system—i.e. districts, communes—to smaller zones. The process is supported by ancillary data, usually land cover data. These approaches assign a population density to each land cover type in a certain zone of the study area. A number of methods belonging to the zonal family use a regression model to improve and obtain the population density for each land cover class (Yuan et al. 1997; Briggs et al. 2007). Mennis (2009) replaces the regression with average densities determined from a sample of zones having a single land cover type. Gallego (2010) evaluates the Expectation-Maximum likelihood algorithm (Flowerdew et al. 1991) that is able to substitute the regression step.

Eicher and Brewer (2001) report three different downscaling approaches:

- Binary method (Langford and Unwin 1994): This method assigns the population to a single land cover class.
- Three-class method: This method allocates some population density to forestry and agricultural classes.
- Limiting variable method: This approach starts with a homogeneous population density for all land cover classes per administrative unit. The population density is then refined through thresholds applied to each land cover class and a redistribution of the "leftover" population to other land cover classes.

Eicher and Brewer (2001) conclude in their chapter, that the limiting variable method performs best. Other publications related to the group of limit-based methods described by Eicher and Brewer (2001) are e.g. Reibel and Bufalino (2005) or Mrozinsky and Cromley (1999).

Gallego (2010) describes four methods to generate dasymetric population density grids based on population data on a commune level and Corine Landcover. This chapter evaluates the disaggregation of four different methods, namely CLC-iterative, CLC-Lucas, CLC-Lucas logit and the EM algorithm. The results of the disaggregation processes are compared with the GEOSTAT 1A population grid. In this chapter the CLC-Lucas logit method performed best, but according to Gallego (2010) are the differences between the approaches moderate.

An approach to disaggregate population data to a resolution of 100 m is presented in Gallego et al. (2011). The chapter evaluates six methods to disaggregate population data based on the commune population data and Corine Landcover. Other ancillary data sources are EUROSTAT point survey and the land use/cover frame survey. The disaggregated data were evaluated using parts of the GEOSTAT 1A population grid, and best results were obtained with a modified version of the limiting variable method.

The aim of this chapter is to evaluate a spatial disaggregation approach of the GEOSTAT 1A population density grid in a defined study area with accurate reference data. The article elaborates on the disaggregation of the GEOSTAT 1A population grid having 1 km resolution with two distinct target granularities—100 and 500 m. An evaluation of the correlation of the resulting population grids with reference data and a weighted random population grid, results in the "similarity" of the disaggregated, reference and random population grid.

This chapter is organized as follows. In Sect. 2 the study area and data used in this chapter are explained. The disaggregation methodology and evaluation approach is described in Sect. 3. The results are given in Sect. 4. Section 5 comments on the results obtained. A conclusion and outlook is given in Sect. 6.

2 Study Area and Data

The study area is located in the northern part of the province of Salzburg and some parts of Upper Austria, Austria. Special focus is set on the northern part of the province of Salzburg and the northern parts of the City of Salzburg and surrounding areas. Hence, the study area covers densely populated as well as rural areas. The data used in this chapter are originating from the Austrian Bureau of Statistics, EUROSTAT and European Environmental Agency. The following chapter elaborates on the study area followed by a description of the spatial data used in the experiment.

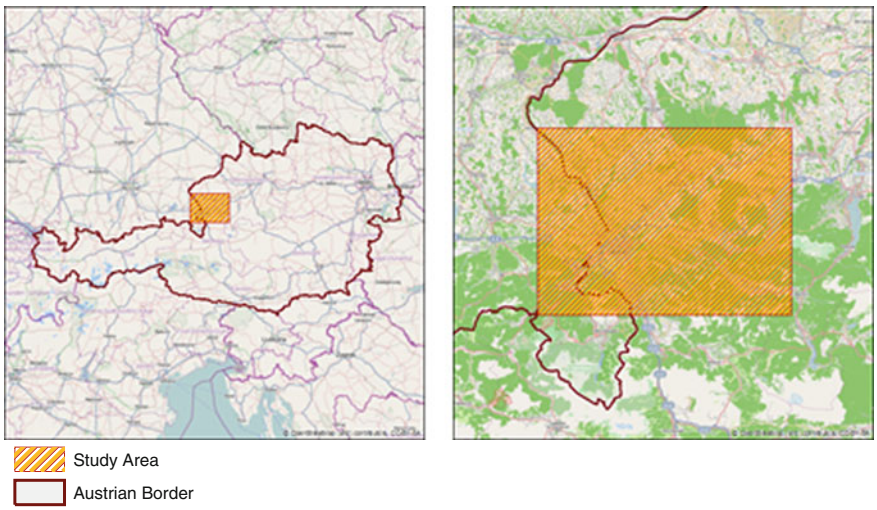


Fig. 1 Location of the study area in Austria (*left*), and the detailed map of the study are (*right*)

2.1 Study Area: Northern part of Salzburg

The study of this chapter is conducted in the northern part of the Province of Salzburg and the western parts of Upper Austria, Austria. In order to have areas with varying population density represented in the study, the area of interest comprises of urban and rural areas.

Figure 1 shows the location of the study area in this chapter. In addition, the base Corine Landcover (CLC 2006) classes are depicted in order to underpin the varying population density in the different land cover classes. Important for this chapter are densely populated areas of the City of Salzburg, that are covered by urban fabric and the outskirts of the city that show urban sprawl. To the north of the city of Salzburg large areas dominated by agriculture and forestry can be found, that are only sparsely populated.

2.1.1 Used Source Datasets

In order to conduct an evaluation of the quality of downscaling methods, several datasets are necessary that originate from official statistical sources. In this chapter data of the Austrian Bureau of Statistics are used for ground truth information, and population datasets of the EUROSTAT provide one ingredient for spatial downscaling. In addition, the European Environmental Agency provides data on the Corine Landcover Classification.

The Austrian Bureau of Statistics collects the population census and provides aggregated census data—i.e. population numbers in a regular grid—with a resolution of 100 and 500 m compatible to the European Reference raster in Lambert azimuthal equal area projection (ETRSA89-LAEA). Hence, the spatial resolutions of the population rasters of the Austrian Bureau of Statistics fit to each other and to the European Reference raster. The 100 m population raster dataset serves as "ground truth" for validating of the disaggregation results, due to the underlying accurate census data. The reference year of the Austrian census data is 2010.

The population raster to be disaggregated is provided by EUROSTAT, and was created by the GEOSTAT 1A project of the EFGS. The resolution of the population density grid is 1 km and the population data are based on the reference year 2006. The data sources used to generate the GEOSTAT 1A population raster are listed in EFGS (2012). Hence, they are not mentioned in this chapter due to the minor relevance for this work.

In order to spatially disaggregate population grids, ancillary data are necessary that provide additional information on where population lives. On a European level Corine Land Cover 2006 (CLC) is an appropriate dataset that maps the land cover in a 100 m resolution grid. CLC is produced by applying common interpretation rules to SPOT-4 and IRS P6 satellite images (EEA-ETC/TE 2002). The results of the CLC are land cover datasets representing the land cover in a 1 ha resolution raster with a minimum mapping unit of 25 ha. The CLC nomenclature consists of 44 classes, which are hierarchically organized. If a polygon cannot be clearly assigned to one dominant land cover type the area is denoted as "heterogeneous". Gallego (2010) reports that smaller urban areas are not represented due to their small patch size smaller than the minimum mapping unit of 25 ha. The CLC data for the study area is given in Fig. 2.

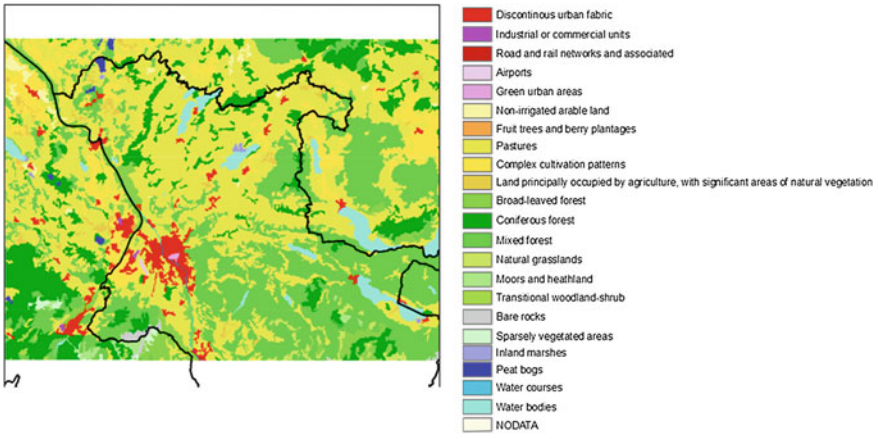


Fig. 2 Corine Landcover data of the study area in the northern part of Salzburg. The biggest continuous urban fabric agglomeration in the middle of the map is the city of Salzburg

3 Spatial Disaggregation of Geostat 1A Population Raster and Evaluation of Disaggregated Population Raster

This section describes the spatial disaggregation method used in this chapter to generate a 100 m population raster from the original 1 km Geostat 1A European population grid. The method used in this chapter is strongly related to the approach presented by Gallego and Peedell (2001) and Gallego (2010). In addition, this publication evaluates the results of the disaggregation of the Geostat 1A population raster, by comparing it with accurate census data of the Austrian Statistical Bureau. Besides, a detailed evaluation of the disaggregation accuracy of different CLC classes provides strengths and weaknesses of the disaggregation approach as such.

3.1 Spatial Disaggregation of Geostat 1A Population Raster

Spatial disaggregation of datasets refers to a process that creates high resolution datasets based on low resolution information with auxiliary data. For the case of population density, CLC data employed as ancillary information on where population is living. The disaggregation approach used here is similar to Gallego and Peedell (2001) and Gallego (2010). The spatial disaggregation methodology is described in detail in this section.

In order to spatially disaggregate the population data of the Geostat 1A raster with 1 km resolution CLC data are employed and integrated similar to the CLC-iterative method described by Gallego (2010), Gallego and Peedell (2001) and Thieke et al. (2006). Gallego (2010) as well as Gallego and Peedell (2001) describe downscaling of population data based on population data per commune (EU LAU 2 level) and CLC—which is an approach that is based on different spatial resolution. This is underpinned by the fact that the spatial extent of communes varies to a certain extent. This is reflected by the statistical evaluation of the surface area of LAU 2 entities based on EUROSTAT (2012) which is depicted in Table 1. In this table the mean surface area of a LAU 2 entity—i.e. a community—and the standard deviation and the skewness is given. The standard deviation given in Table 1 indicates that the spatial extent of the communities varies, which results in different spatial resolutions. In this chapter, we disaggregate based on one

Table 1 Statistical evaluation of LAU2 entity area data for the EU27 (except Denmark and Germany) based on EUROSTAT (2012)

Statistical metric	LAU2 entity area (km^2)
Mean	37,64
Median	145,20
Standard deviation	211,67
Skewness	47,56

homogeneous spatial resolution which is determined by the resolution of the Geostat 1A population grid—1 km.

The disaggregation of the population data—i.e. the Geostat 1A population raster is done using the CLC-iterative method. This approach assumes that base data to be disaggregated is available as polygons representing communes (Gallego 2011; Gallego 2010; Gallego and Peedell 2001). In this chapter this approach is altered, due to the fact that the dataset to be disaggregated is a raster data model. The methodology presented here does not follow the CLC-iterative method presented in Gallego and Peedell (2001) completely. Hence communes cannot be stratified within each NUTS2 region into dense, intermediate and sparse population communes. NUTS is an abbreviation for the Nomenclature of territorial units for statistics, a hierarchical system of territorial units in the European Union according to EC Regulation No. 1069/2003. Nevertheless, the model presumes a fixed-ratio:

$$Y_{cm} = U_c W_m \tag{1}$$

In Eq. 1 m denotes a raster cell in the original, coarse Geostat 1A grid. In this model Y_{cm} represents the population density for land cover class c in the raster cell m of the Geostat 1A grid. In addition, U_c denotes the relative population density for each land cover class c. W_m is a number that ensures the pycnophylactic constraint (Tobler 1979) for each raster cell m after the estimation of U_c. In order to calculate U_c and Y_{cm} the following equations (Eq. 2) are necessary:

$$\begin{aligned} X_m = \sum_c S_{cm} Y_{cm} \Rightarrow W_m &= \frac{X_m}{\sum_c S_{cm} U_c} \\ \Rightarrow Y_{cm} &= U_c \frac{X_m}{\sum_c S_{cm} U_c} \end{aligned} \tag{2}$$

X_m denotes the population in raster cell m, and S_{cm} is the space that is covered by land cover class c in raster cell m. The disaggregation process starts with a parameter U_c using Eq. 3. The population data for each target raster cell m'—where m' denotes a target raster cell having finer resolution than m—are disaggregated with the coefficients U_c and S_{cm} (Eq. 3). Furthermore, $X_{m'}$ denotes the population in raster cell m', and $S_{cm'}$ is the space that is covered by land cover class c in raster cell m'

$$Y_{cm'} = U_c \frac{X_{m'}}{\sum_c S_{cm'} U_c} \tag{3}$$

Consecutively, the population attributed to raster cell m is estimated by using Eq. 4. Furthermore, the known population in the raster cell X_m—the original coarse raster cell m—is compared with the estimated population X_m^* in order to calculate disagreement indicators given in Eq. 5.

$$X_m^* = \sum\nolimits_{m' \in m} \sum\nolimits_c S_{cm'} Y_{cm'} \tag{4}$$

$$\psi_m = \frac{X_m^*}{X_m} \quad \delta_m = \sum \left| X_m^* - X_m \right| \tag{5}$$

Due to the fact that the estimation of the disaggregated population density is dependent on the population resident in the coarse raster cell m denoted as X_m the authors omit the iterative calculation of U_c. In addition, through Eq. 5 an evaluation of the population density of Geostat 1A and the estimated population of the disaggregated population raster is possible.

3.2 Evaluation of Disaggregated Population Raster

The evaluation of the disaggregated population raster is done by utilizing aggregated accurate census data provided by Austrian Statistical Bureau, having a spatial resolution of 100 and 500 m. In addition, the disaggregated population raster data are compared with a guided—i.e. weighted—random distribution, in order to evaluate if the results of the disaggregation process show similarities to a random distribution. The evaluation approach is depicted in Fig. 3. In general, the Geostat 1A population raster is disaggregated and compared with the reference dataset of the Austrian Statistical Bureau and a weighted random population grid, resulting in several deviation numbers and statistics. The following sections elaborate on the evaluation method in detail.

In order to evaluate the accuracy of the disaggregated population grids the disagreement between the population of disaggregated raster cells X_m^* and the reference raster cells X_m^R is calculated by using Eq. 6. Due to the same extent and origin of the two grids a direct comparison is possible. This is done for two given spatial resolutions, 100 and 500 m respectively—resulting in disagreement $D_{Da,R,100}$ and $D_{Da,R,500}$. Similar to Eq. 6 the absolute disagreement between reference and weighted random data (see Eq. 7)—$D_{Da,Rnd,100}$ and $D_{Da,Rnd,500}$—as

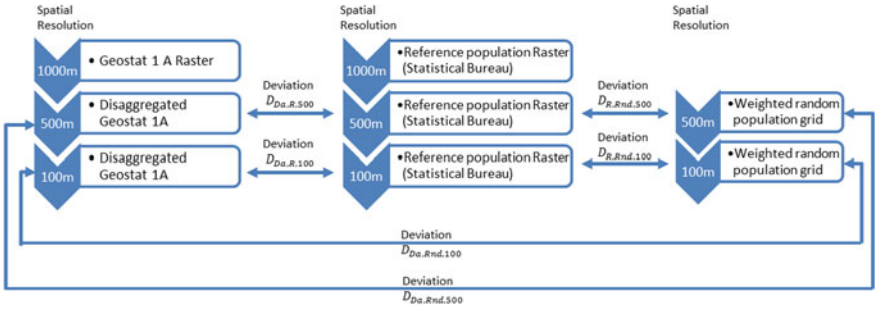


Fig. 3 Evaluation approach of the disaggregated population raster. The approach emphasizes on the deviations between the disaggregated population raster, reference population raster from Austrian Statistical Bureau and weighted random population grid

well as the disagreement between disaggregated and weighted random grid (see Eq. 8)—$D_{Da,Rnd,100}$ and $D_{Da,Rnd,500}$—is calculated.

$$D_{Da,R} = \sum \left| X^{*}_{m'} - X^{R}_{m'} \right| \tag{6}$$

$$D_{Da,Rnd} = \sum \left| X^{R}_{m'} - X^{Rnd}_{m'} \right| \tag{7}$$

$$D_{Da,Rnd} = \sum \left| X^{*}_{m'} - X^{Rnd}_{m'} \right| \tag{8}$$

In addition, the disagreements between the population grids depicted in Fig. 1 are represented as difference grids, which can be processed in any GIS. Hence, the evaluation of the disagreement grids contains a comparison of the statistical parameters for each disagreement grid. In detail, there are the following disagreement grids used in this chapter:

- $G_{Da,R,500}$: difference between disaggregated and reference population grid, 500 m resolution
- $G_{Da,R,100}$: difference between disaggregated and reference population grid, 100 m, resolution
- $G_{R,Rnd,500}$: difference between reference and random population grid, 500 m, resolution
- $G_{R,Rnd,100}$: difference between reference and random population grid, 100 m, resolution
- $G_{Da,Rnd,500}$: difference between disaggregated and random population grid, 500 m, resolution
- $G_{Da,Rnd,100}$: difference between disaggregated and random population grid, 100 m, resolution

The disagreement grids are analyzed regarding their statistics in order to analyze the "behavior". Thus, the comparison with a weighted random grid is done in order to evaluate if the disaggregation on the two examined granularity levels becomes more similar to a random distribution. Hence, the following statistical parameters are computed for each disagreement grid: maximum absolute difference, standard deviation σ, skewness, and kurtosis. In addition, the calculation of the correlation matrix between the population rasters is done in order to evaluate the similarity between them. Correlation matrices express the similarity of raster layers (Snedecor and Cochran 1968).

The weighted random distribution, used to evaluate the nature of the disaggregated population raster, is a function that creates random point distribution—i.e. population—with respect to a given probability. Due to the fact, that the basic correlation between CLC and population density is known in advance, we used that information in order to create a probability surface for the placement of random points. Hence, the random point generator placed the number of "humans" in the study area that is defined by the census data of the Austrian Statistical Bureau. The random points are placed such that raster cells with larger

Table 2 Population density classes and respective Corine Landcover classes (Gallego and Peedell 2001)

Population density	Corine Landcover classes
Dense	112
Medium	211, 222, 231, 242, 243
Sparse	121, 122, 123, 141, 311, 312, 313, 321, 322

values are more likely to have a point placed in them—which is defined by Corine Landcover. For that reason the CLC classes were divided into three categories (Gallego and Peedell 2001): densely populated, medium populated and sparsely populated. The respective Corine Landcover Classes are displayed in Table 2. Subsequently, the weighted random points are aggregated into regular grids having 100 and 500 m resolution sharing the extent of the reference grid of the Austrian Statistical Bureau.

In addition, an evaluation of the population grids for three population density classes, given in Table 2, is conducted. The population grids are divided into population density classes. The correlation coefficient between the population grids and the three population density classes is calculated and evaluated. Furthermore, the variance—as a sign of disagreement—is computed based on the density classes and the generated disagreement grids.

4 Results

This section elaborates on the numerical results of the 500 and 100 m disaggregation of the Geostat 1A population grid and the evaluation thereof. First, the chapter highlights the results of the disaggregation methodology and presents some graphical outcomes of the disaggregation process. Secondly, the evaluation of the disaggregated Geostat 1A raster with 100 and 500 m resolution is presented.

The disaggregation process as described in Sect. 3.1 results in population datasets that have finer granularity than the original Geostat 1A grid. The resulting disaggregated population raster with a resolution of 100 m is given in Fig. 4 *right*, denoted with *Disaggregated Population* 100. The reference dataset originating from the Austrian Statistical Bureau is depicted in Fig. 4 *left*, denoted with *Population* 100.

The resulting disaggregated population raster with a resolution of 500 m is given in Fig. 5 *right*, denoted with *Disaggregated Population* 500. The reference dataset originating from the Austrian Statistical Bureau is depicted in Fig. 5 *left*, denoted with *Population* 500.

The evaluation of the disaggregated datasets follows the approach described in Sect. 3.2. This comprises the calculation of the disagreements between population grids based on Eqs. 6–8 and Fig. 3. In addition, the disagreement grids are created

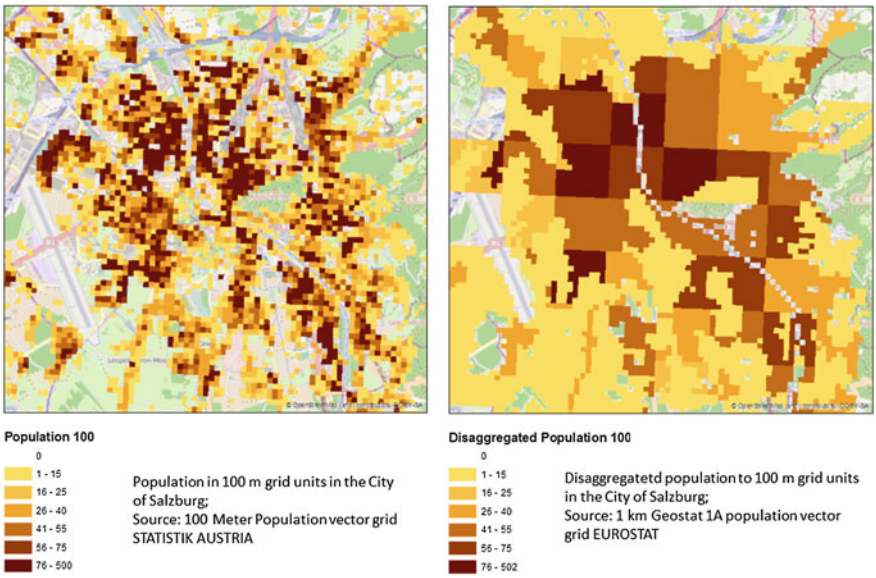


Fig. 4 Visual comparison of the reference population grid with 100 m resolution of the Austrian Statistical Bureau (*left*), and the disaggregated population raster with 100 m resolution (*right*). Both grid datasets show the city of Salzburg and their outskirts

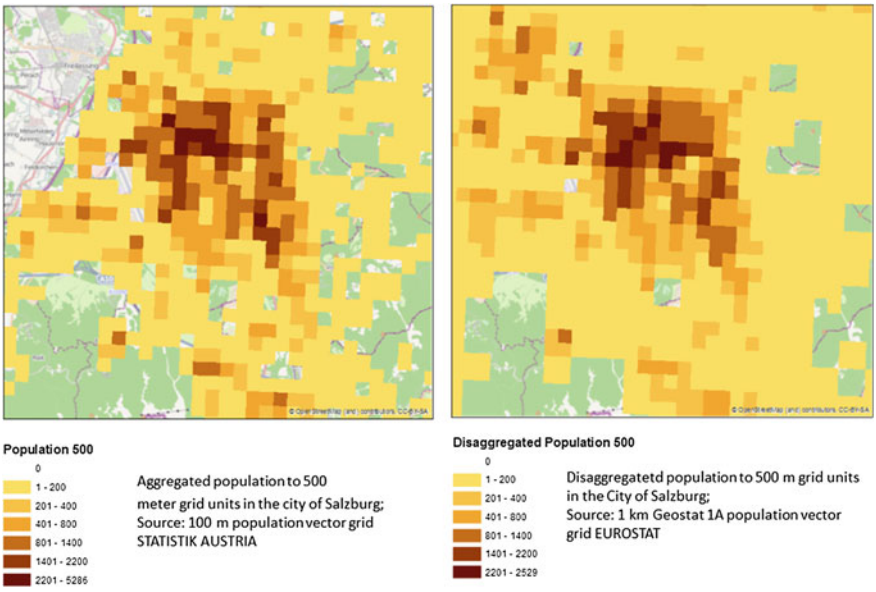


Fig. 5 Visual comparison of the reference population grid with 500 m resolution of the Austrian Statistical Bureau (*left*), and the disaggregated population raster with 500 m resolution (*right*). Both grid datasets show the city of Salzburg and their outskirts

Table 3 Statistical evaluation of the disagreement grid $G_{Da,R}$ for 100 and 500 m resolution

Reference grid versus disaggregated grid	100 m resolution $G_{Da,R,100}$	500 m resolution $G_{Da,R,500}$
$D_{Da,R}$	493239	210593
Maximum absolute difference	708	2828
Standard deviation σ	10,21	86,47
Skewness	15,63	13,8
Kurtosis	478,01	319,1

Table 4 Statistical evaluation of the disagreement grid $G_{R,Rnd}$ for 100 and 500 m resolution

Disaggregated grid versus random grid	100 m resolution $G_{Da,Rnd,100}$	500 m resolution $G_{Da,Rnd,500}$
$D_{Da,Rnd}$	353290	358170
Maximum absolute difference	143	2383
Standard deviation σ	5,6	120,05
Skewness	10,9	10,53
Kurtosis	160	141,8

Table 5 Statistical evaluation of the disagreement grid $G_{R,Rnd}$ for 100 and 500 m resolution

Reference grid versus random grid	100 m resolution $G_{R,Rnd,100}$	500 m resolution $G_{R,Rnd,500}$
$D_{R,Rnd}$	519523	406322
Maximum absolute difference	134	4730
Standard deviation σ	10,27	135,26
Skewness	18,5	13,4
Kurtosis	604	291,73

Table 6 Correlation coefficient of population grids

Correlation coefficient	100 m resolution	500 m resolution
Disaggregated population versus reference population	0,47	0,87
Random population versus reference population	0,34	0,74
Disaggregated population versus random population	0,56	0,80

with Map Algebra, and the statistics of these grids are calculated respectively. The results are presented in Tables 3, 4 and 5.

To evaluate the correlation between the population grids the correlation and covariance matrices are calculated in a pairwise manner. These results show the correlation between the reference, disaggregated and random population grids on two levels of detail (100 and 500 m). The results are presented in Table 6.

An evaluation of the standard deviation of the density classes of the disagreement grids gives completes the accuracy results of the population grids (see Table 7). In Table 8 the absolute deviation of the disaggregated population raster with respect to the population density classes is given, which is calculated based on Eqs. 6–8.

Table 7 Standard deviation of the disagreement grids and their population density classes

Standard deviation	Densely populated	Medium populated	Sparsely populated
Disaggregated population minus. reference population $G_{Da,R}$			
100 m resolution	44,2	7,5	2,4
500 m resolution	236	45,5	12,5
Random population minus reference population $G_{R,Rnd}$			
100 m resolution	49,1	7,2	2,3
500 m resolution	443,6	62,7	19,1
Disaggregated population minus. random population $G_{Da,Rnd}$			
100 m resolution	25,26	2,5	1,0
500 m resolution	376	45,5	16,1

Table 8 Absolute disagreement of the disaggregated population grid with respect to population density classes, and reference population numbers in density classes

	Densely populated	Medium populated	Sparsely populated
Absolute disagreement			
100 m resolution	161265	298394	29882
500 m resolution	50952	127526	23095
Relative disagreement			
100 m resolution	101 %	146 %	216 %
500 m resolution	37 %	61 %	83 %
Reference population			
100 m resolution	160112	204503	13833
500 m resolution	138571	208370	27955

5 Discussion of the Results

The results of the disaggregation process and the evaluation are given in Sect. 5. Based on the numerical results given, a discussion thereof is conducted. This section focuses on the interpretation of the results achieved, and comments critically on the numbers. The section highlights the disaggregated population grids and the comparison with the reference population raster data. In addition, the evaluation of the weighted random population distribution in combination with the reference and disaggregated grid should elaborate on the "difference" between disaggregation and randomly generated datasets.

First the chapter elaborates on the visual difference between disaggregated and reference population grid. The disaggregated and reference population grid with 100 m resolution are depicted in Fig. 4. Noticeable are the visual differences that are observable between the reference and the disaggregated population dataset. In comparison to the latter, the disaggregated and reference population grids with 500 m resolution (see Fig. 5) show less visual differences. This underpins the

assumption that the disaggregated dataset with 100 m resolution has lower accuracy than the 500 m population raster.

A numerical evaluation of the disagreement between the disaggregated reference population grid for 100 and 500 m shows lower disagreement $D_{Da,R}$ at the 500 m level (see Table 3). In addition, the authors look at the "distance" between the reference grid to the disaggregated and random population grid. In order to have a metric for that, the authors look at the correlation coefficients of the grids and the standard deviation of the disagreement rasters. Tables 4, 5 show that the standard deviation of the disagreement grid $G_{Da,Rnd}$ and $G_{R,Rnd}$ for 100 and 500 m resolution respectively. For 100 m resolution $G_{R,Rnd}$ shows a standard deviation of 10,27 whereas the $G_{Da,Rnd}$ has a deviation of 5,6. For 500 m resolution $G_{R,Rnd}$ shows a standard deviation of 135,26 whereas the $G_{Da,Rnd}$ has a deviation of 120,05. This gives evidence, that the disaggregated population raster at 100 m is "closer" to the random data than the reference grid. For 500 m resolution the standard deviation of $G_{Da,Rnd}$ is lower than the one of $G_{R,Rnd}$, but the proportion between them is lower than at 100 m level. Hence, the authors assume that the 500 m disaggregation results differ from random population grid with a similar "distance" as the reference grid. In addition, the standard deviation of the $G_{Da,R}$ is lower than $G_{Da,Rnd}$ which shows that the disaggregated grid is "closer" to the reference grid, which is supported by the disagreement numbers $D_{Da,R,500} < D_{Da,Rnd,500}$. Furthermore, the disagreement of the disaggregated grid to the random grid $D_{Da,Rnd}$ at 100 m resolution is lower than $D_{Da,R,100}$, and the standard deviation shows a similar behavior. This indicates that the disaggregated population raster at 100 m level is closer to the random population grid. In general, the facts support the argument that the disaggregated 500 m population raster shows fewer inaccuracies than the one with 100 m resolution, which is closer to the weighted random population grid.

In order to evaluate on the similarity of the disaggregated population data with the reference population grid, the chapter highlights the correlation between the raster data sets at different levels of detail accordingly (see Table 6). Generally speaking, the correlation between disaggregated and reference grid shows a correlation coefficient of 0,87 at the 500 m level. Compared to the value of 0,47 at 100 m level the authors conclude that the 500 m disaggregation result is similar to the reference population, as the correlation of the disaggregated to the random population at 500 m resolution is slightly lower. For 100 m resolution the situation is different, as the correlation between disaggregated and random population is higher than the correlation between disaggregated and reference population. Nevertheless, the correlation coefficients at the 100 m level are low in comparison to the 500 m resolution. Generally, the correlation coefficient between random and reference population is lowest at both levels of detail, whereas the relative distance to the correlation coefficient of disaggregated and reference is lower at the 500 m resolution level. The evaluation of the correlation coefficients of the population density grids shows that at 100 m resolution level the disaggregated population raster shares most similarities with the withed random population grid, whereas at

500 m resolution the correlation coefficient between disaggregated and reference grid shows the highest value.

In addition, the standard deviation of the disagreement grids with respect to population density zones, given in Table 2, is analyzed. The results are given in Table 7, where $G_{Da,Rnd}$ shows the lowest standard deviation in all population density classes of the 100 m resolution. For 500 m resolution $G_{Da,R}$ shows the lowest standard deviation for all population density classes, except for the medium populated areas. For medium populated areas $G_{Da,R,500}$ and $G_{Da,Rnd,500}$ share a standard deviation of 45,5 which indicates that the distance between disaggregated to reference and disaggregated to random population grid are equal, and the disaggregation at the 500 m level for medium populated areas shows accuracy deficits. In addition, $G_{Da,R}$ of sparsely populated areas with 500 m resolution shows a slightly lower standard deviation than $G_{Da,Rnd}$. Hence, the distance between disaggregated and reference population grid and disaggregated and random population grid is comparable small when looking at the relative difference. Hence, the authors assume that the disaggregation for sparsely populated areas shows inaccuracies.

The absolute differences for the population density classes in different levels of detail, given in Table 8, indicate that for the 500 m resolution the absolute disagreement in densely populated areas is lowest in comparison to the reference population. For sparse and medium populated areas the absolute disagreement is generally higher in comparison to the reference population, whereas population density grid the 500 m resolution shows lower disagreement than at 100 m resolution.

6 Conclusion and Outlook

The article elaborates on the disaggregation of the GEOSTAT 1A population grid for a study area located in the northern part of the province of Salzburg and some parts of Upper Austria in Austria. The chapter describes an approach to downscale a population raster of 1 km resolution towards a target resolution of 500 and 100 m respectively, by using Corine Landcover as ancillary data. In order to evaluate the results achieved the authors employ an accurate reference dataset originating from the Austrian Bureau of Statistics representing the census in Austria.

The results achieved with the methodology described show that disaggregating of population grids with Corine Landcover is possible. The accuracy evaluation indicates that the results achieved at 100 m resolution show more correlation with a weighted random population distribution than with the reference population data. For 500 m resolution the disaggregated grid is slightly more correlated with the reference dataset. In addition, the total absolute disagreement is lower for the 500 m grid. These numerical results give evidence, that the 500 m disaggregation of the GEOSTAT 1A raster is more accurate than the 100 m disaggregation.

In detail, the results for population density zones show that densely populated areas can be estimated quite well, whereas disaggregation results in medium and

sparsely populated areas tend to be more inaccurate. This fact is mentioned in Gallego et al. (2011) and Gallegeo (2010) as well.

Further research in this area could include the investigation of the effect of further ancillary data and other disaggregation approaches mentioned in literature. Interesting for the authors are crowd sourced auxiliary data, like open street map data, that have at least some additional inherent land cover information. In addition, an evaluation of the obtained accuracy in the context of the intended application area is pending.

Acknowledgments The present work has been funded by the European Commission under Framework Programme for RTD 7 through the project "Modeling and Simulation of the Impact of Public Policies on SMEs (MOSIPS)"—Grant agreement no.: 288833.

References

Bierkens MFP, Finke PA, de Willigen P (2000) Upscaling and downscaling methods for environmental research. Kluwer, Dordrecht, p 190

Briggs DJ, Gulliver J, Fecht D, Vienneau DM (2007) Dasymetric modelling of small-area population distribution using land cover and light emissions data. Remote Sens Environ 108:451–466

Chen K, McAneney J, Blong R, Leigh R, Hunter L, Magill C (2004) Defining area at risk and its effect in catastrophe loss estimation: a dasymetric mapping approach. Appl Geogr 24:97–117

Eicher C, Brewer C (2001) Dasymetric mapping and areal interpolation: implementation and evaluation. Cartography Geogr Inf Sci 28:125–138

EEA-ETC/TE (2002) Corine land cover update. Technical guidelines. Web: http://www.eea.europa.eu/publications/technical_report_2002_89

EUROSTAT (2012) Local administrative units. Web: http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/local_administrative_units. Accessed 29 Dec 2012

Flowerdew R, Green M, Kehris E (1991) Using areal interpolation methods in GIS. Pap Reg Sci 70(3):303–315

Gallego FJ (2010) A population density grid of the European Union. Popul Environ 31(6):460–473

Gallego FJ, Batista F, Rocha C, Mubareka S (2011) Disaggregating population density of the European Union with CORINE land cover. Int J Geogr Inf Sci 25(12):2051–2069

Gallego J, Peedell S (2001) Using CORINE land cover to map population density. Towards agri-environmental indicators, topic report 6/2001 European environment agency, Copenhagen, http://reports.eea.eu.int/topic_report_2001_06/en. pp 92–103

Langford M, Unwin DJ (1994) Generating and mapping population density surfaces within a geographical information system. Cartogr J 31(1):21–26

Mennis J (2009) Dasymetric mapping for estimating population in small areas. Geogr Compass 3(2):727–745

Mrozinski R, Cromley R (1999) Singly- and doubly-constrained methods of areal interpolation for vector-based GIS. Trans Geogr Inf Syst 3:285–301

Reibel M, Agrawal A (2007) Areal interpolation of population counts using pre-classified land cover data. Popul Res Policy Rev 26(5–6):619–633

Reibel M, Bufalino M (2005) Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems. Environ Plann A 37(1):127–139

Snedecor GW, Cochran WG (1968) Statistical methods, 6th edn. The Iowa State University Press, Ames, Iowa

Thieken A, Mueller M, Kleist L, Seifert I, Borst D, Werner U (2006) Regionalisation of asset values for risk analyses. Nat Hazards Earth Syst Sci 6:167–178

Tobler WR (1979) Smooth pycnophylatic interpolation for geographical regions. J Am Stat Assoc 74:519–530

Tralli DM, Blom RG, Zlotnicki V, Donnellan A, Evans DL (2005) Satellite remote sensing of earthquake, volcano, flood, landslide and coastal inundation hazards. ISPRS J Photogr Remote Sens 59(4):185–198

Vinkx K, Visee T (2008) Usefulness of population files for estimation of noise hindrance effects. In: Proceedings of ICAO committee on aviation environmental protection. CAEP/8 modelling and database task force (MODTF), 4th meeting. Sunnyvale, USA, February 2008, pp 20–22

Yuan Y, Smith RM, Limp WF (1997) Remodeling census population with spatial information from Landsat TM imagery. Comput Environ Urban Syst 21(3–4):245–258

Tailoring Trajectories and their Moving Patterns to Contexts

Monica Wachowicz, Rebecca Ong and Chiara Renso

Abstract Nowadays heterogeneous mobile data sources are producing an enormous amount of contextual information that can improve our interpretation of discovered mobility patterns. Because both an entity and the data sources can be mobile, what context is and how it can be used to interpret mobility patters may vary anyplace at anytime. This chapter describes an approach for tailoring mobility patterns based on the synergy of trajectory and mobility pattern annotation techniques, where contexts are represented as dynamic semantic views. These views are obtained after the classification of context variables that are selected based on the classification criteria previously proposed for a taxonomy of collective phenomena. An experiment is used to illustrate the proposed approach for tailoring moving flock patterns to contexts of visitors in a recreational area.

1 Introduction

Due to the latest advancements in telecommunication, wireless and location technologies, context information of any entity is now accessible from mobile devices that come with technologies such as GSM, UMTS, Bluetooth and Wi-Fi. The word 'context' itself is derived from the Latin con (with or together) and texere (to connect). In previous research work, context information has been used to adapt interfaces (Hariri et al. 2008), tailor the set of application-relevant data (Bolchini et al. 2011), increase the precision of information retrieval (Crestani and Ruthven 2007), discover services (Rasch et al. 2011), make the user interaction

M. Wachowicz (✉)
University of New Brunswick, New Brunswick, Canada
e-mail: monicaw@unb.ca

R. Ong · C. Renso
CNR—KDD Lab, Pisa, Italy

D. Vandenbroucke et al. (eds.), *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography,
DOI: 10.1007/978-3-319-00615-4_16,

implicit (Schmidt 2000), build smart environments (Schmidt 2011) and support a semantic enrichment knowledge discovery process (Baglioni et al. 2009). Most of this research has been focused on the technical issues associated with context, and the syntactic relationships between different application domains (e.g. environmental context, spatio-temporal context, social context, and user context).

In this chapter, we define context as a dynamic process where mobility patterns are tailored to contexts based on the annotation and classification of context variables belonging to discovered mobility patterns and their individual trajectories.

This chapter is organised as follows: Sect. 2 provides an overview of the related literature on the existing definitions of a context. Section 3 describes the current techniques developed for context acquisition. Section 4 proposes a context building process based on annotation and classification techniques. Section 5 provides a discussion of the performed experiment and the results obtained from implementing our approach for understanding moving flock patterns of visitors in a recreational park. Finally, Sect. 5 summarises the contributions of this work, the conclusions obtained from the experiment, and possible extensions of the current work.

2 What is a Context?

In order to use context effectively, we must understand what context is and how it can be used. In this section, we will review previous conceptualisations of context in order to lay down the foundations for proposing our adopted definition. The well known definition from Dey (2001) stated that context is any information that can be used to characterise the situation of an entity. According to Kofod-Petersen and Mikalsen (2005), two main categories are found in the literature. First, context has been defined by specific entities, such as location or object. For example, Schilit and Theimer (1994) have first coined the term context-aware, having in mind location, identity of nearby people and objects as the primary sources of context. They have classified context into three categories: computing context (e.g. communication bandwidth, network connectivity, nearby resources such as printers and computers); user context (e.g. location, user's profile, other users' location); and physical context (e.g. noise levels and weather conditions). Ryan et al. (1998) have further this idea of context from an archeological domain perspective by adding the time and identity as other categories (See Chen and Kotz 2000 for a survey of context-aware mobile computing research). All these views are focused on the assumption that context is a particular type of information.

In contrast, context has been also defined in terms of existing relationships and structures of context categories. Brézillon and Pomerol take the view that there is no special type of knowledge that can objectively be called context, they argue that context is in the eye of the beholder, or in their own words: "… knowledge that can be qualified as "contextual" depends on the context!" (Brézillon and Pomerol 1999). More recently, Göker and Myrhaug (2002) advocate the standardisation of

a user context in the belief that users want to be mobile, and therefore, they propose a generic user context divided into five-subcategories: task context, social context, personal context (physiological and mental sub-contexts), spatio-temporal context, and environmental context. (Details are found in Göker and Myrhaug 2002).

Given the complexity, there is a risk of ending up in a situation where everything can be a context. To remedy this, we propose the definition that a context is a process where dynamic semantic views are connected to mobility patterns. In order to achieve that, we propose annotation techniques of trajectories and their respective mobility patterns. This will be explained further in the next section.

3 Context Acquisition

Annotations are tied to situational factors as contexts, and as a result, they are important to a context-building process. Marshall (1998) points out that annotation can be constructed in several ways: as link making, as path building, as commentary, as marking in or around existing text, as a de-centering of authority, as a record of reading and interpretation, or as community memory. In this chapter, annotation refers to selecting context variables with useful value for building a semantic view of a mobility pattern. Moreover, these context variables provide answers to "who", "what", "when", "where", and "how" questions. While semantic annotation for the Web has been a well studied topic as demonstrated in the survey provided in Uren et al. (2006), the set of annotation approaches for mobility patterns is quite new and hence, not as rich. Semantic annotation enriches the unstructured or semi-structured data with a context that is further linked to the structured knowledge of a domain. We have distinguished two types of annotation techniques which have been previously developed for mobility data sets. They are: trajectory and pattern annotation levels.

3.1 Trajectory Annotation

Existing works on trajectory annotation mostly add context information about stops and moves, episodes and trips.

Stops and Moves Annotation. Stop and moves of trajectory (SMoT) is an algorithm that converts each trajectory to a corresponding list of stops and moves (Alvares et al. 2007a). It accepts the list of interesting places, along with the typical time duration spent in these places in order to be considered a stop, as input. The authors indicate that interesting places are defined by the users. Subsequently, it checks whether each point of the trajectory intersects the region of any candidate stop according to the list of stops and moves. If the duration of the

stop is at least equal to the associated minimum duration of the region, the place is considered as a stop. A move is then recorded between the previous stop and the recently marked stop. Cluster-based stop and moves of trajectory (CB-SMoT) is an extension of SMoT (Alvares et al. 2007b) based on a clustering technique that does not assume user's knowledge on all the interesting places and the typical time duration spent in each place. The algorithm itself automatically identifies places that may be relevant according to a given context.

Aside from SMoT and CB-SMoT, Yan and Spaccapietra (2009) and Yan (2010) also describe other approaches for identifying and annotating stops. Moreover, Yan et al. (2011) introduced a system called Semantic Middleware for Trajectories (SeMiTri), which semantically annotates trajectories by exploiting both geometric properties and domain knowledge. Its first phase is called the trajectory computation layer, which performs some pre-processing steps, such as data cleaning and episode identification. The second phase uses three annotation layers including semantic region, line, and point annotation layers in order to annotate trajectory with the regions and roads that it passes as well as the probable activities carried on during the stops. In the third phase, the annotation computation is stored into the semantic trajectory store.

Episode, Trip and Trip Purpose Annotation. Guc et al. (2008) proposed the use of GPS tracks to facilitate manual semantic annotation without having the need for neither manual interview nor manual mobility records. They proposed a conceptual annotation model that includes two annotation elements, which are episodes and trips. Episodes were defined by Mountain and Raper (2001) as time periods in which the user's movement behaviour was relatively homogeneous while trips are sequences of episodes that are concerned with a common aim. The homogeneity of episodes in Guc et al. (2008) depends on the purpose of an action and the mode of transportation though this may be extended further depending on the context.

Using this model, they have implemented an annotation tool developed in the Java environment. The architecture of the software includes three layers: data handling for the storage of the trajectory and annotation data, program control for the program, and user interface for the graphical user interface (GUI) components. The tool includes interface functionality for visualisation of the GPS tracks, display of temporal trajectory aspects through a timeline bar, trajectory animation for visualising slow and fast movements in certain time periods and the direction of movement as well, and place marks allowing the user to specify his/her favourite places.

Wolf (2000) and Wolf et al. (2001) have studied the feasibility of replacing travel diaries, which require a manual recording and retrieval process, with automatic extraction of trips and trip purposes from GPS tracks. Trips are automatically extracted by checking the part of trajectories wherein there is no movement detected. Once the end of trips and other relevant information are derived, the next step involves automatic extraction of trip purposes. This, however, requires a manual process of combining land use information and other geographic information in the case that the land use information is not enough.

The land use data are linked with a set of purposes, which includes a primary purpose and may include a secondary or even a tertiary purpose. These combined properties are used to determine the purpose of a moving entity by matching the land use with the identified purposes based on the starting and the ending positions of the trips, and the temporal component of the trips made by the entity. Axhausen et al. (2003) proposes a similar approach that uses personal information about peoples' home and work addresses aside from the land use information.

3.2 Mobility Pattern Annotation

While trajectory annotation techniques enrich individual trajectories, mobility pattern annotation techniques enrich the mobility patterns themselves. This section provides some examples of the latter technique.

Mei et al. (2006) proposes an approach for automatically generating semantic annotations for frequent patterns. This is realised by building a context model, extracting representative transactions, and finding semantically similar patterns for each frequent pattern. The context model is built by selecting a set of informative context indicators, which is made up of context units that have the strongest weights with respect to the currently considered frequent pattern. Each transaction is modelled as a vector and representative transactions are selected by finding the top-ranking transactions based on the cosine similarity. Finally, the set of similar patterns are selected by computing the similarity between the context models of the frequent pattern with that of the candidate patterns.

Another example is found in Baglioni et al. (2008, 2009), wherein frequent patterns based on the discovered stops are post-processed in order to classify patterns according to a predefined context defined by using a domain ontology. The ontology represents the concepts, rules and assumptions present in the considered application domain for conceptual representation and deductive reasoning of trajectory patterns obtained from GPS tracks.

Our approach differs from the previous research since it exploits the synergy of both trajectory and mobility pattern annotation.

3.3 Our Context Building Process

There is no pragmatic context acquisition that allows us to efficiently rule out information that is not context, especially in the case of mobility patterns. For example, during a pre-processing task, developers can ask the question: is this information relevant for understanding the mobility patterns? If not, does it mean that the information should be discarded as not being context? But what would happen if in a later stage, the same information might prove to be relevant for the understanding the discovered mobility patterns. Therefore, we propose the synergy of trajectory and mobility pattern annotation techniques for building a context based process. In this process, context is represented by a collection of semantic

views that are obtained from the classification of context variables. These variables, in turn, have been previously selected using the classification criteria from the taxonomy for collective phenomena originally presented by Wood and Galton (2009).

Step 1: Mapping between the classification criteria and context variables

Since the number of properties can be large, their selection can be a difficult task. A simple random choice is possible, but this can lead to meaningless results when tailoring mobility patterns to contexts. As a guideline for selecting the context variables to be used in the context building process, we propose the use of the taxonomy for collective phenomena presented by Wood and Galton (2009). This taxonomy was proposed as a framework to represent a wide variety of collectives that exist (e.g. society, football team, the wheels of a car, or the atoms of a water molecule) and the different types of mobility patterns that they exhibit. Our research premise is that a collective may produce a set of mobility patterns over a period of time. These patterns endure over a time period, exist as a whole at each moment during that period, and possibly undergo various types of change (e.g. in location or membership).

Furthermore, this taxonomy offers a set of classification criteria that represents a wide variety of collectives. The use of these criteria helps us to know and understand how different collectives relate to each other. We have selected the membership and location criteria proposed in this taxonomy for tailoring mobility patterns to contexts. The other criteria (i.e. depth, cohesion and role) were not used in this research due to the lack of context information available for the experiment that could be used to evaluate the proposed context building process.

Membership. This criterion is concerned with both the identity and cardinality of the members (e.g. the individual trajectories) that belong to a movement pattern. For example, in the case of flock patterns, the main division is between those patterns which have the same members throughout their lifetime (*constant membership*) from those that will cease to exist when the cardinality drops below two. Sample context variables based on this criterion would include the number of members belonging to a discovered pattern, and the characteristics of each member (e.g. cars flocking on a highway access versus birds flocking in the sky).

Location. This criterion deals with three granularity levels: the location of a mobility pattern, the location of its members, and the relation between these two. We have identified four categories of contexts according to the location of a mobility pattern and its members. They are described as one of the following:

- Fixed Pattern and Member Locations: A series of mobility patterns that has a fixed location as well as its members;
- Fixed Pattern but Variable Member Locations: A series of mobility patterns that has a fixed location but the location of its members are variable;
- Variable Pattern and Member Locations: A series of mobility patterns that has a variable location as well as its members;

- Variable Pattern but Fixed Member Locations: A series of mobility patterns that has a variable location but the location of its members is fixed.

For the first two categories, a context represents not only a situation where the location of a mobility pattern is fixed, but it also includes situations amongst mobility patterns which cannot be assigned a location. For example, a postal address of a mobility pattern does not denote its location. Suspension patterns are an example of a series of mobility patterns that has a fixed location as well as its members. In transportation, they may indicate traffic congestions, traffic lights, and construction work. In contrast, a crowd crossing a busy street intersection is represented by a series of flow patterns that has a fixed location but the location of its members is variable. Each individual member might be moving in different random directions, but the location of the flow pattern is fixed (e.g. a zebra crossing), and the pattern will gradually disappears when all members have moved forward in one general direction in order to cross the street.

For the last two categories, a context represents a situation where all the members of a mobility pattern can be moving in essentially the same way or in a random way, but following the same or otherwise spatially related paths, either simultaneously or staggered over time. Synchronized flow patterns in highways are an example of a series of mobility patterns where the mobility pattern as a whole is moving in the same way as its members. In contrast, flow patterns of population migration can be seen as a series of patterns which migrate through a set of fixed individuals.

Our research premise here is that candidate context variables that can be mapped to any of these classification criteria are considered as relevant context information for annotation while the rest should be disqualified.

Step 2: Perform multi-level annotation

Trajectory Annotation. *A trajectory T of an entity E is represented as:* $T_E = \{<x_1,y_1,t_1, \ldots x_n,y_n,t_n>\}$ *where n is the number of sample points recorded during the movement of the entity* E. A *trajectory* is built from a sample of points recorded by mobile tracking devices, such as GPS, GMS, or WI-FI sensors. At this level, we propose two further types of contextualisation. Every trajectory in the dataset can be annotated, especially when dealing with a small and sparse dataset. By doing so, statistically significant classification results can be inferred. An alternative is to only annotate the trajectories that are involved in the discovered mobility patterns, which results in shorter processing time for the annotation.

Mobility Pattern Annotation. Our discussion here focuses on moving flock patterns. *Given a set of n trajectories consisting of line segments (i.e. sub-trajectories) that can vary in number for different trajectories, an (m,k,r)-moving flock* F_M *in a time interval* $I = [t_i, t_j]$, *where* $j-i+1 \geq k$, *consists of at least m entities such that for every discrete time step* $t_l \in I$, *there is a disk of radius r that contains all the m entities and the spatial extent* $ext(F_M,I) \geq r$ *(removed for blind review).* Hence, moving flocks have context variables such as a spatial extent covered by each flock as a whole, the start time and the end time of flocking. We

categorise the context variables into three groups: (1) the parameters used by the data mining algorithm, (2) the descriptions generated by the data mining algorithm, (3) and the aggregated properties of the moving entities involved in the flock pattern (See Sect. 4 for examples). The first two categories can be directly obtained from the algorithm while the last group requires an aggregation step wherein individual properties of entities belonging to a flock are combined in order to obtain a corresponding context variable describing the flock as a whole. The corresponding set of context variables can be obtained by expanding the trajectory property of a member based on all its possible values.

Step 3: Building Semantic Views of Mobility Patterns

The last step of the proposed context building process is concerned with building semantic views from the context variables used for annotation at both trajectory and mobility pattern levels. In this step, we propose two data mining techniques that can be used for building the semantic views. These techniques are hierarchical clustering and classification with decision tree.

For hierarchical clustering, correlation scores are first computed among the context variables. These scores are then used to build the distance matrix needed to perform clustering. This analysis produces dendograms that give an overview of the relations among the considered context variables. The classification approach, on the other hand, produces decision trees that focus on certain relations connecting the membership of the entities to a specific mobility pattern instance with related context variables. Though the decision trees do not cover all considered context variables and/or mobility pattern instances, they provide more details compared to the dendograms as to how the discovered relations are connected. The discussion of applying these techniques is detailed in the subsequent paragraphs.

Hierarchical Clustering. It groups together entities in a progressive way such that the most highly correlated entities are first grouped together. This is applied recursively until the entire set of entities is grouped into one cluster. The clustering algorithm requires the distance matrix, which summarises the dissimilarity scores among the items, as a parameter. The distance matrix for the context variables of a trajectory and mobility pattern level can be easily computed from the computation of correlation scores. The output of the clustering algorithm can be useful when the analyst needs to focus on certain groups of mobility patterns that are of interest. This is remarkably useful when there are a large number of discovered mobility patterns. Furthermore, clustering may reveal interesting relationships, which may or may not be obvious to a user, and they can support the analyst in uncovering possible reasons for the occurrence of the mined patterns. However, the obtained results are still quite limited though in terms of aiding the user in understanding the nature and occurrences of the mobility patterns.

Classification. Though clustering is able to pinpoint which context variables are correlated, it is not able to pinpoint how or why these context variables are correlated. Therefore, we propose the use of a decision tree based classification

algorithm in order to study the relations among each mobility pattern and their respective trajectories since decision trees are simple and intuitive. We have specifically selected a cost-sensitive version of the J48 classification algorithm, whose implementation is accessible in WEKA, to generate the decision tree connecting context variables of the trajectories that contribute to the membership of an entity to a specific mobility pattern. The use of a cost-sensitive algorithm is appropriate when the dataset considered consists of entries biased to a specific value. Putting more weight to the less occurring value eases the bias present in the dataset. J48 is WEKA's Java implementation of the C4.5 algorithm, which in turn is a known standard classification algorithm. C4.5 was developed by Ross Quinlan as an extension of his earlier ID3 algorithm. It builds decision trees by considering the attribute that most effectively splits the training dataset into subsets having the most homogenous values for the target attribute. This is determined by computing each attribute's normalized information gain and splitting the dataset using the attribute with the highest normalized information gain. This step is applied recursively to the subsets obtained at each step.

4 Experiment

The semi-synthetic version of a dataset containing the trajectories of 372 visitors of a recreational park has been used in this experiment. This data was collected once a month during spring and summer of 2006 for a total of 7 days, including both weekends and weekdays (Fig. 1).

Step 1: Mapping between classification criteria and context Variables

The dataset is interesting despite of its small size since it contains context variables derived from visitors' responses to conducted surveys, which contain 23 questions from which 73 context variables were derived. Whether the visitor is on holiday, the frequency of visits, the number of accompanying children, adults and dogs, and the main attractions visited are examples of such context variables. Running the data mining algorithm on this dataset allowed the discovery of 11 moving flocks when the radius was set to 150 m. The algorithm used in this experiment is open source available on http://www-kdd.isti.cnr.it/moving-flock/ and its description can be found in Wachowicz et al. (2011). Figure 2 shows an example of the discovered moving flocks. The top three moving flock patterns ranked by extent for radius 150 m are found in Table 1.

The discovered moving flocks are a series of patterns that has a fixed location but the location of its members (i.e. visitors) is variable. They explicitly occupy a region on space that is computed by using its extension and radius parameters. A larger spatial extent means that the flock covered a longer distance in the trails (i.e., the visitors moved together for a longer distance).

The context variables were selected based on the classification criteria (i.e. membership and location) and a mapping between them with these criteria for the

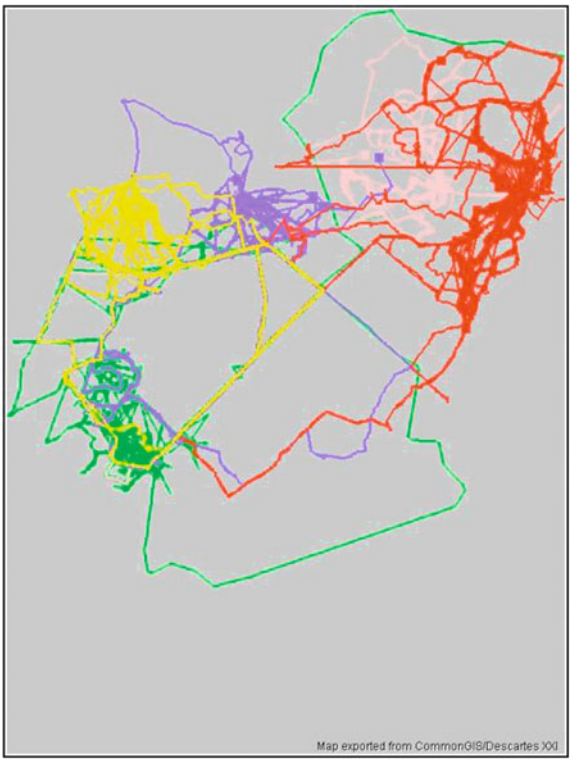


Fig. 1 Overview of four trajectories of the dataset used for the computation of moving flocks

dataset is shown in Table 2. Moreover, redundant context variables (e.g., properties that can be derived from other properties), unary context variables (i.e., properties consisting only of one value), or context variables that have no semantic meaning (e.g., ID) were also removed.

Step 2: Multi-Level Annotation

Once the flocks were discovered, both trajectory and mobility pattern annotation were applied to the dataset using the proposed taxonomy for collectives. At the

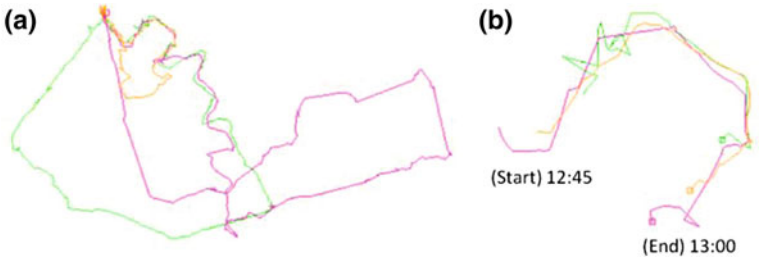


Fig. 2 Example of a discovered moving flock. **a** The moving entities included in the flock. **b** The trajectory segments wherein the flock pattern occurs

Table 1 Example of three moving flock patterns discovered from the dataset

Radius	Start time	End time	Flock extent	Flock members
150 m	12:15	12:30	991.9375	96; 288; 15
	9:40	9:55	870.5	228; 287; 104
	11:55	12:05	692.4375	118; 249; 346

Table 2 Some examples of the mapping between classification criteria and context variables

Criteria	Context variables
Membership	Is in the area for holiday?
	Frequency of visit
	Visited since when
	Total number of visited attractions
	Number of information sources used
	Is a local?
	Has visited an attraction?
	Is a browser?
	Is a repeater?
	Is a dogwalker?
	Is with children?
	Age category
Location	Has visited the following…? [picnic areas, mound, currant forests, information centre, woods, bird watching sites, prayer areas, juniper berries, fens, sheepfold areas, snack bar areas, sightseeing areas, radio telescope, david lakes, orienting, teahouses];
	Followed a route?
	Has stopped? type of stop. Stopped for the following…? [catering, beautiful, quiet, seat, lunch]; has followed the…? [white route, whiteLheederzand, redSpier, blue route, redLheederzand, yellowLheederzand, redLheederzandEast, redDiepveen, yellowLheebroekerzand, whiteSpier)

trajectory level, the trajectories belonging to a discovered flock were annotated with their corresponding visitor characteristics. An example of a visitor characteristic is visitor type, which specifies whether the visitor is an elderly person, an adult, an elderly couple, an adult couple, a family with children, a group of adults, or a family consisting of adults. Aside from visitor characteristics, each trajectory was also annotated with a set of membership context variables, indicating whether the visitor that made the trajectory belongs to a specific flock or not. Table 3 shows a sample annotation of 3 flock members in the dataset. In this table, r_id refers to the entity ID, on_holiday is a Boolean value describing whether the visitor is in the area for a holiday or not, freq_visit describes how often the visitor comes to the park, adult_num is the number of adults represented by the current entity, and children_num is the number of children included in the current entity. The context variables adult_num and children_num specifically imply that each entity in the dataset may consist of a group of adults and/or children. Meanwhile, picnic_areas, mound, bird_watching_site, and prayer_areas indicate if the entity visited these

Table 3 Example of context variables of three members (i.e. visitors) belonging to the same moving flock pattern

r_id character varying(10)	On_holiday integer	Freq_visit integer	Adult_num integer	Children_num integer	Picnic_areas integer	Mound integer	Bird_watching_site integer	Prayer_areas integer	Flock0 integer	Flock1 integer
R195	1	2	2	0	0	0	0	0	1	0
R647	0	1	2	0	0	0	0	0	1	0
R015	0	3	2	0	0	0	1	0	1	0

attractions. Finally, flock0 and flock1 are context variables generated by the flock discovery algorithm and they indicate if the visitors belong to a specific flock.

In addition to annotating individual trajectories of flock members, flocks themselves are also annotated. Recall that the set of flock properties is categorised into three groups. The first group (parameters used to discover flock patterns) include min points, radius, min time slices, and synchronisation rate. Meanwhile, the start and end time of flocking, the spatial extent covered by the flock, the duration of flocking, the number of flock members, the average speed of flock members are examples of the second group (generated flock descriptions). Some examples of the third group (collective properties based on the individual property) are visitor_type_1 (i.e., elderly alone), visitor_type_2 (i.e., adult alone), visitor_type_3 (i.e., elderly couple), etc. Since visitor_type can have the following values: 1, 2, 3, 4, 5, 6, 7, 99 (i.e., unknown), there is a corresponding collective property for each of these possible values (i.e., having eight corresponding collective properties in this case). The value of each collective property depends on the number of members satisfying the considered individual property value. For example, if a given flock pattern has 1 out of 3 members whose visitor_type property is equal to 1, then the flock property visitor_type-1 of the flock is set to 0.33 (i.e., 1/3).

Table 4 shows a subset of the context variables at flock level for the dataset. In this example, there are 10 discovered flocks, each one having the on_holiday and the freq_visit aggregated context variables. The context variable on_holiday at the trajectory level contains 3 possible values: 0, 1, null. Thus, at the flock level, there are 3 context variables associated with it. Likewise for the freq_visit, it has 6 possible values: neg,1, 2, 3, 4, 5 and hence, there are 6 corresponding context variables at the flock level. These context variables describe the percentage of flock members having the specified value for the considered trajectory level. For instance, on_holiday_0 for flock0 has a value of 0.666667 indicating that 66.7 % of flock0's members have a value of 0 for the on_holiday property. Meanwhile, on_holiday_1 has a value of 0.333333, which means that 33.3 % of flock0's members have a value of 1 for the on_holiday property (i.e., 33.3 % of flock0's members are in the area for a holiday while the remaining percentage are not).

A total of 108 context variables, which include survey-based properties and algorithm generated descriptions such as start time of flocking, were used for pattern level annotation. The parameters used to extract the moving flocks were considered as unnecessary since the flocks considered were obtained using exactly the same parameters. In other words, these parameters have unary values and thus, were disregarded as context variables for the pattern annotation level.

Moreover, pairs of complementary survey-based flock properties were redundant. For example, the complement of bird_watching_site_0 is bird_watching_site_1 and, vice versa since the value of a property can be easily computed from the other. Thus, one of these properties was removed. Recall that

Table 4 Example of context variables of three members (i.e. visitors) belonging to the same moving flock pattern

Flock_id integer	On_holiday_0 real	On_holiday_1 real	On_holiday_null real	Freq_visit_neg real	Freq_visit_1 real	Freq_visit_2 real	Freq_visit_3 real	Freq_visit_4 real	Freq_visit_5 real
0	0.666667	0.333333	0	0	0.333333	0.333333	0.333333	0	0
1	0.333333	0.666667	0	0.333333	0.333333	0	0	0	0.333333
2	0.333333	0.666667	0	0	0.666667	0.333333	0	0	0
3	0.666667	0.333333	0	0	0.333333	0	0.666667	0	0
4	0.333333	0.666667	0	0	0.333333	0.333333	0.333333	0	0
5	0.333333	0.666667	0	0	0.333333	0	0.333333	0.333333	0
6	1	0	0	0	0.666667	0	0.333333	0	0
7	0.666667	0	0.333333	0	0	0	1	0	0
8	0.333333	0.666667	0	0	0.333333	0	0.666667	0	0
9	0.333333	0.666667	0	0	0.333333	0	0.666667	0	0
10	0.5	0.5	0	0	0.25	0	0.75	0	0

bird_watching_site_0 is the percentage of flock members who did not visit the bird watching site, while bird_watching_site_1 is the percentage of those who did.

Step 3: Classification

We focused on the 11 moving flock patterns that were extracted from the semi-synthetic version of the dataset using the following parameters: min points = 3, radius = 150 m, min time slices = 3, synchronization rate = 300 s. Each flock has 3–4 members each, and the members remain spatially close for 3–7 time instances (i.e., 10–30 min).

This section describes the results obtained from performing classification analysis when considering only trajectories belonging to a moving flock. The J48 classification algorithm was run 11 times, one for each of the discovered flocks in order to obtain a decision tree for each flock. We will now describe how the decision trees correspond to semantic views representing the context.

Figure 3 presents two semantic views, which were obtained when the target class was set to Flock0 (left) and to Flock2 (right). The semantic view on the left implies that Flock0 may have taken place because its members were interested in visiting the radio telescope attraction. Additionally, each member consists of a couple who were either adults or elderly. On the other hand, the obtained semantic view on the right explains that Flock2 has occurred probably because their members have either visited the mounds or the bird watching sites.

Another interesting semantic view obtained using J48 is shown in Fig. 4. The target class in this case is Flock9 and the semantic view implies that members of Flock9 also belongs to Flock8. Those that do not belong to Flock8 should follow the White route while members of Flock8 should have visited the park before. Note that since_when = −1 means that there is no data on when the visitor last visited the park.

The J48 classification algorithm was also executed on the context variables at the pattern level. When the target class was set to main_activity_1, the semantic view shown in Fig. 5 was obtained. since_when_5 indicates the percentage of flock members who have visited the park more than 10 years before. The semantic view suggests that this pattern took place when visitors who have visited the park more than 10 years ago have the tendency to have walking as their main park activity.

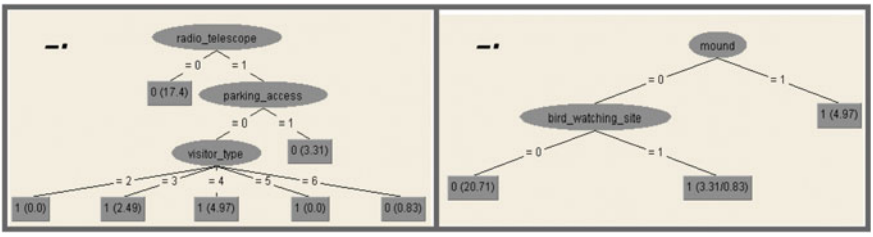


Fig. 3 Two semantic views obtained from the classification of context variables of trajectories belonging to moving flock patterns when the target class is flock0 and flock2, respectively

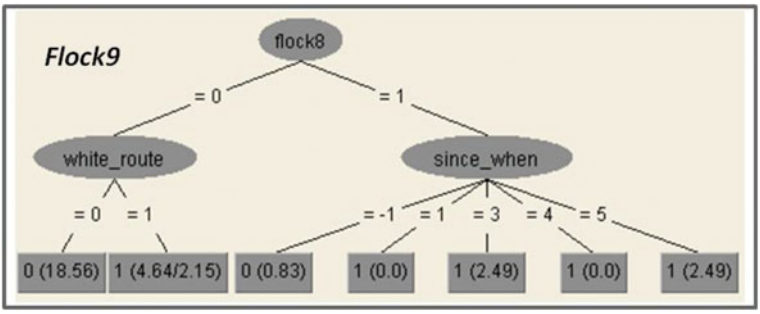


Fig. 4 Decision tree obtained based on individual properties of flock members when the target class is flock9

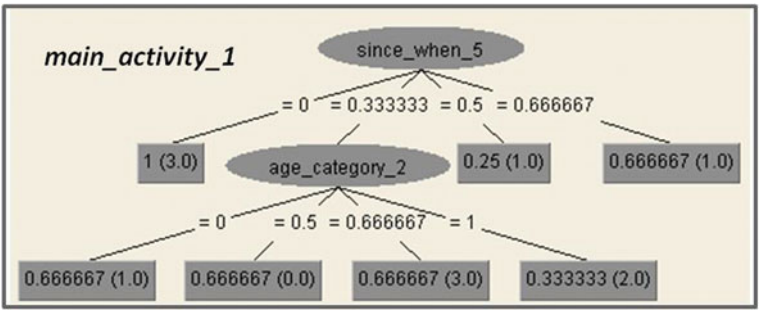


Fig. 5 The semantic view obtained based on flock properties when the target class is main activity 1

This type of relations between main activities and other context variables could be derived from the other classification results as well.

5 Conclusions

In this chapter, we demonstrated the importance of a context building process for the purpose of interpreting mobility patterns. The interpretation results obtained in this work could not have been acquired without such a process. First, we propose the use of Wood and Galton's taxonomy for selecting the appropriate context variables for our context building process. In the experiment, we have shown the importance of using the classification criteria proposed in this taxonomy. In particular, the membership and location criteria were used to illustrate the need to minimise the number of context variables in order to focus on those that are most

relevant for interpreting the mobility patterns. Second, we propose the combined use of two levels of annotation, namely trajectory and mobility pattern level. Both of which are important since context variables at the trajectory level influence context variables at the mobility pattern level. Aside from describing how trajectories and patterns can be contextualised, we have also shown how a decision tree classification algorithm can be used in generating contexts as dynamic semantic views. The main advantages of using this type of algorithm for analysis are the following: (1) the decision trees can be efficiently generated due to the simplicity of the algorithm, (2) the obtained trees are intuitive and easy to understand, in general, (3) and the algorithm allows the user to concentrate on the most important relations among the different context variables and the discovered patterns.

The obtained results were encouraging as we have found interesting relations among trajectory context variables despite of the fact that there were only 29 trajectories involved in flocking out of the 370 trajectories available in the dataset. The most interesting relations among trajectory properties are those that are between a flock membership property and a visitor characteristic property since this type of relations allows the analyst to build a context that possibly causes them to flock together. Meanwhile, among mobility pattern properties, the most interesting relations are between activity related properties and other visitor characteristic properties, allowing flock activities to be correlated with flock characteristics.

Currently, the taxonomy used is at a very high level in comparison with the actual context variables available in a dataset and hence, making the filtering power of the selection process limited (i.e., few context variables were discarded as irrelevant for further analysis). The selection of context variables needs to be enhanced further by using a set of criteria whose granularity is closer to those of the actual context variables. Another direction for future work is the application of other techniques aside from annotation in order to allow a higher level of automation and a deeper level of embedding relevant context information.

References

Alvares L, Bogorny V, Kuijpers B, de Macêdo J, Moelans B, Vaisman A (2007a) A model for enriching trajectories with semantic geographical information. In: Proceedings of the 15th annual ACM international symposium on advances in geographic information systems, pp 1–8

Alvares L, Bogorny V, Kuijpers B, Moelans B, de Macêdo J, Palma A (2007b) Towards semantic trajectory knowledge discovery (Tech. Rep.). Hasselt University, Belgium

Axhausen K, Schönfelder S, Wolf J, Oliveira M, Samaga U (2003) 80 weeks of GPS-traces: approaches to enriching the trip information. In Transp Res Rec 1870:46–54

Baglioni M, de Macêdo J, Renso C, Wachowicz M (2008) An ontology-based approach for the semantic modelling and reasoning on trajectories. In: Song I, Piattini M, Chen Y, Hartmann S, Grandi F, Trujillo J et al (eds) Advances in conceptual modeling challenges and opportunities, lecture notes in computer science, vol 5232. Springer, Berlin, pp 344–353

Baglioni M, de Macêdo J, Renso C, Trasarti R, Wachowicz M (2009) Towards semantic interpretation of movement behavior. In: Cartwright W, Gartner G, Meng L, Peterson M (eds) Advances in GIScience, lecture notes in geoinformation and cartography. Springer, Berlin, pp 271–288

Bolchini C, Orsi G, Quintarelli E, Schreiber FA, Tanca L (2011) Context modelling and context awareness: steps forward in the context-ADDICT project. IEEE Data Eng Bull 34(2):47–54

Brézillon P, Pomerol J-C (1999) Contextual knowledge sharing and cooperation in intelligent assistant systems. Le Travail Humain 62(3):223–246

Chen G, Kotz D (2000) A survey of context-aware mobile computing research, technical report TR2000-381, department of computer science, Dartmouth College. Retrieved from. http://www.cs.dartmouth.edu/reports/TR2000-381.pdf on November 07 2012

Crestani F, Ruthven I (2007) Special issue on contextual information retrieval systems. Inf Retrieval 10(2):111–113

Dey AK (2001) Understanding and using context. Pers Ubiquit Comput 5(1):4–7

Göker A, Myrhaug HI (2002) User context and personalisation. In: Workshop proceedings for the 6th European conference on case based reasoning. Aberdeen, Sept 2002. Retrieved from http://openaccess.city.ac.uk/624/2/User_Context_and_Personalisation.pdf on November 07 2012

Guc B, May M, Saygin Y, Körner C (2008) Semantic annotation of GPS trajectories. Paper presented at the 11th AGILE international conference on geographic information science. Girona

Hariri A, Tabary D, Lepreux S, Kolski C (2008) Context aware business adaptation toward user interface adaptation. Communications of SIWN, 3 June 2008, pp 46–52

Kofod-Petersen A, Mikalsen M (2005) Context: representation and reasoning. Representing and reasoning about context in a mobile environment. Revue d'Intell Artificielle 19(3):479–498

Marshall CC (1998) Toward an ecology of hypertext annotation. In: Grenbaek K, Mylonas E, Shipman FM, III (eds) Proceedings of the ninth ACM conference on hypertext and hypermedia, the association for computing machinery, 20–24 June 1998, Pittsburgh, New York, pp 40–49

Mei Q, Xin D, Cheng H, Han J, Zhai C (2006) Generating semantic annotations for frequent patterns with context analysis. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, pp 337–346

Mountain D, Raper J (2001) Modelling human spatio-temporal behaviour: a challenge for location based services. In: Proceedings of the 6th international conference on GeoComputation

Rasch K, Li F, Sehic S, Ayani R, Dustdar S (2011) Context-driven personalized service discovery in pervasive environments. Science + business media, LLC 2011. Retrieved from. http://www.infosys.tuwien.ac.at/Staff/sd/papers/Zeitschriftenartikel%20Fei%20Li%20world%20wide%20web.pdf on 7 Nov 2012

Ryan N, Pascoe J, Morse D (1998) Enhanced reality fieldwork: the context-aware archaeological assistant. In: Gaffney V, van Leusen M, Exxon S (eds) Computer applications and quantitative methods in archaeology. British Archaeological Reports. Tempus Reparatum, Oxford

Schilit BN, Theimer MM (1994) Disseminating active map information to mobile hosts. IEEE Netw 8(5):22–32

Schmidt A (2000) Implicit human computer interaction through context. Pers Ubiquit Comput 4(2/3):191–199

Schmidt A (2011) Interactive context-aware systems interacting with ambient intelligence. In: Riva G, Vatalaro F, Davide F, Alcañiz M (eds) Ambient intelligence, pp 159–178

Uren V, Cimiano P, Iria J, Handschuh S, Vargas-Vera M, Motta E, Ciravegna F (2006) Semantic annotation for knowledge management: requirements and a survey of the state of the art. J Web Seman 4(1):14–28

Wachowicz M, Ong R, Renso C, Nanni M (2011) Discovering moving flock patterns among pedestrians through spatio-temporal coherence. Int J Geogr Inf Sci 25(11):1849–1864

Wolf J (2000) Using GPS data loggers to replace travel diaries in the collection of travel data. Dissertation, Georgia Institute of Technology, Atlanta, GA

Wolf J, Guensler R, Bachman W (2001) Elimination of the travel diary: an experiment to derive trip purpose from GPS travel data. In Transp Res Rec 1768:125–134

Wood Z, Galton A (2009) A taxonomy of collective phenomena. Appl Ontol 4(3–4):267–292

Yan Z (2010) Traj-ARIMA: a spatial-time series model for network-constrained trajectory. In: Proceedings of the 2nd international workshop on computational transportation science, pp 11–16

Yan Z, Spaccapietra S (2009) Towards semantic trajectory data analysis: a conceptual and computational approach. Paper presented at 35th very large data base (VLDB) PhD workshop, Lyon, France

Yan Z, Chakraborty D, Parent C, Spaccapietra S, Aberer K (2011) SeMiTri: a framework for semantic annotation of heterogeneous trajectories. In: Proceedings of the 14th international conference on extending database technology, pp 259–270

Part VI
Decision Support Systems Related to Mobility

Facility Use-Choice Model with Travel Costs Incorporating Means of Transportation and Travel Direction

Toshihiro Osaragi and Sayaka Tsuda

Abstract Estimating the number of users and their spatial distribution is necessary in the planning process of new public facilities. In the present study, we construct a model based on the nested logit model that is composed of facilities' utility and users' travel costs to describe facility use-choice behavior with respect to the facilities. The travel costs are described in terms of network distance, means of transportation, direction of travel, and number of transfers. As a numerical analysis, validation of the proposed model is achieved by estimating the number of users and their spatial distribution for newly constructed public libraries.

1 Introduction

Two of the problems in planning the scale and location of public facilities are to predict the region whose people will actually use the facility (*catchment area*) and to estimate how many people will use it. The catchment area and number of users of a library or a hospital, for example, will fluctuate widely according to variety of factors, since users are free to choose which facility they use.

These issues are often discussed from the viewpoint of the distance decay theory, which is a geographical term describing the effect of distance on cultural or spatial interactions. The interaction between two locales declines as the distance between them increases. However, advances in transportation technology, such as automobile, railways and airplanes have decreased the effects of distance.

T. Osaragi (✉) · S. Tsuda
Department of Mechanical and Environmental Informatics, Graduate School of Information Science and Engineering, Tokyo Institute of Technology, 2-12-1-W8-10 O-okayam, Tokyo, Meguro-ku 152-8552, Japan
e-mail: osaragi@mei.titech.ac.jp

S. Tsuda
e-mail: tsuda@os.mei.titech.ac.jp

D. Vandenbroucke et al. (eds.), *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography,
DOI: 10.1007/978-3-319-00615-4_17,

Thus, the concept of distance has been recently replaced with that of accessibility, and there accordingly exists a large body of basic research in the area of accessibility.

The accessibility measures of recent research are based on Burns' (1979) utility-theoretic framework, which assesses accessibility in terms of the benefits accruing to individuals at particular activity locations. A central assumption of this framework is that individuals are both spatially and temporally constrained by a set of fixed activities that bind them to particular places at specific times of the day (Neutens et al. 2012).

Recent research has further extended this theme by using more complicated discrete choice random utility models. Hsu and Hsieh (2004) formulated an individual accessibility model to measure the accessibility benefits of daily activities, and extended it with the assumption that an individual chooses an activity/travel alternative with the maximum accessibility benefits. Furthermore, traditional accessibility measures are not well suited for analyzing accessibility in the context of task combination, since they assumed that individuals make single-stop, single-purpose trips from home. Arentze et al. (1994) and Ettema and Timmermans (2007) elaborated space–time accessibility measures to overcome this problem, building on Burns' (1979) theory, in the context of time geography and activity-based analysis. Also, Neutens et al. (2012) investigated the relationships between opening hours and accessibility, and proposed a method to optimize the temporal regime of public service delivery in terms of accessibility.

With the spread of digital maps and geographic information system (GIS) software, studies have begun to describe accessibility in terms of distances and travel times on the street networks (Green et al. 2009; Comber et al. 2008; Teixeira and Antunes 2008). Satoh et al. (2008) has described the placement of public facilities using indices of travel burden as they vary with topological conditions and physical fitness. Wang and Luo (2005) and Apparicio et al. (2008) investigated the planning of the location of hospitals from the viewpoint of travel time. Studies of accessibility are becoming more and more specific.

Nevertheless, there have not been many detailed analyses of the influence of travel route or the selected means of transportation on accessibility, due to lack of available actual data about such factors. Given this background, the present chapter examines "*travel cost*" as an accessibility measure, which include the overall sacrifice involved in getting to a facility, including monetary expense, labor, and psychological burden. Travel costs depend on a variety of factors besides travel distance and time, such as means of transportation, travel direction, presence of hills (resistance), and availability of public transportation. Of late, city residents have become quite mobile, and the factors determining travel costs are anything but simple. Nakamura and Kurihara (1998) conducted surveys of library users in catchment areas and showed that the routes a user takes in his or her daily life distorts the catchment area away from the nearest station. Also, Osaragi (2002) has written about the influence of the geographical characteristics (the presence of railways or highways, topography, etc.) around a facility on its catchment area.

In the present study, travel costs are modeled in close detail, and a model is constructed to describe the behavior by residential users of a facility. Actual data about users are analyzed. It is shown that this model is not only able to describe current travel by users but also to predict the appearance of new users, changes in the catchment area, and fluctuations in the number of users after the construction of a new facility.

2 Constructing a Use-Choice Model of Facilities

2.1 Formulation of Model

In order to predict the catchment area served by a facility and the number of users, we must establish whether a given resident will use a given facility (use behavior) and which facility that resident will actually use (choice behavior). Here, we construct a *use model* and a *choice model* using the logit model, and the combination of these two models is herein referred to as the *use-choice model*, which can describe hierarchical use-choice behavior of users.

We begin by representing the use behavior s ($s = 1$, "uses"; $s = 2$, "does not use") and the facility number m ($m = 1, 2, \ldots, n$). The utility corresponding to residents of location i engaging in activity s is represented by V_{is}, the utility of using facility m is represented by V_m (the attraction level of the facility), and the cost in terms of utility to travel to facility m is C_{im} (*travel costs*). The probability $P_i(m|s)$ that a resident of location i will choose facility m when he or she uses a facility can then be described using the *choice model* below. The travel costs C_{im} will be described in full detail in Sect. 2.2.

$$P_i(m|s) = \begin{cases} \frac{e^{V_m + C_{im}}}{\sum_{m'=1}^{n} e^{V_{m'} + C_{im'}}} & for\ s = 1 \\ 0 & for\ s = 2 \end{cases} \tag{1}$$

The probability $P_i(s)$ that a resident of location i will use any facility can be described by the *use model* below. The utility V_{is} of activity s will be described in full detail in Sect. 2.3.

$$P_i(s) = \frac{e^{V_{is} + \lambda \Lambda_{is}}}{\sum_{s'=1}^{2} e^{V_{is}' + \lambda \Lambda_{is}'}} \tag{2}$$

where λ is an unknown parameter and Λ_{is} is a function of both the attraction of the facility and the travel costs, given by the following:

$$\Lambda_{is} = \begin{cases} \ln \sum_{m=1}^{n} e^{V_m + C_{im}} & for\ s = 1 \\ 0 & for\ s = 2 \end{cases} \tag{3}$$

The probability $P_i(s,m)$ of a resident of location i using facility m is obtained by combining Eqs. (1) through (3) above into a *use-choice model*.

$$P_i(m,s) = P_i(m|s)P_i(s) \tag{4}$$

2.2 Formulation of Travel Costs

The travel costs are formulated for calculation on the following basis (see Fig. 1). (1) Foot travel costs may differ between ordinary daily walks in the vicinity of one's home (to the nearest train station or bus stop) and walks in the vicinity of a facility. Therefore, two variables were created, Z_{1im} and Z_{2im}, to distinguish between these. (2) Variable Z_{3im} was created to indicate the influence of the slope of a street (hill resistance, i.e., the altitude difference covered on foot). (3) Travel costs by train are low when the facility lies in the same direction as one's work or school destination and are high if in the reverse direction. There is much diversity in commuting directions, but a majority of people commute from the suburbs to the city center (central business district: CBD). The direction from residential locations toward the center of city is called "nobori" ("up-train") in the Japanese style on transportation schedules, and the direction away from the center of the city is called "kudari" ("down-train"); *nobori* and *kudari* are indicated with the variables Z_{4im} and Z_{5im}. It will be verified through analysis that the travel costs are lower if a facility lies in the *nobori* direction from a user's residence than if it lies in the *kudari* direction. (4) There is also a travel cost associated with changing the means of transportation between walking, cycling, buses, trains, etc. Variable Z_{6im}

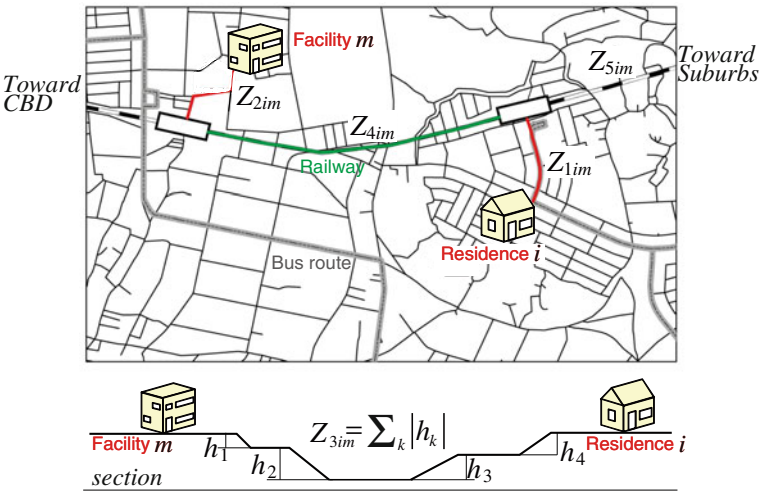


Fig. 1 Definition of explanatory variables for travel costs

represents the number of railway transfers. Variables Z_{7im} through Z_{Tim} represent the other means of transportation.

Thus, travel costs C_{im} are given by the following:

$$C_{im} = \sum_{t=1}^{T} \alpha_t Z_{tim} \tag{5}$$

where t ($t = 1,\ldots,T$) indicates the means of transportation and the travel cost parameter α_t is the weight applied to the travel costs by the explanatory variable Z_{tim}.

2.3 Determining the Attributes of Users

Since the potential for use of a facility depends greatly on the attributes of a given user, the utility function V_{is} in the *use model* must incorporate another function that accounts for user attribute effects. Here, we consider a simple linear model employing variables Z_{ti} expressing user attributes such as sex, age, and profession. The utility value was determined using the following form, which defines the utility as zero when the user does not use the facility ($s = 2$).

$$V_{is} = \begin{cases} \sum_{t=1}^{T} \beta_t Z_{ti} + V & for\ s = 1 \\ 0 & for\ s = 2 \end{cases} \tag{6}$$

3 Calibration of Model Using Data for Library Users

3.1 Pre-processing of Data

The address-based data for users of the Yokohama City Central Library (1994) were transformed into raster data to make the spatial distribution easier to understand. Specifically, a multiple regression model (multiple correlation coefficient = 0.975) describing the population densities in each land-use classification on the basis of national census data was employed to assign the users into cells 250 × 250 m in size. These cells are referred to as *minor districts* herein. The structure of the data used in the present study is shown in Table 1. Each number indicates the number of users who borrowed books at least once within a year. For calculating this value, the same person is counted only once. In case a user used multiple libraries, he/she was counted into the most-often-using library.

Topological information compiled in a digital map was added to the railway and roadway data. Bus routes were digitized using the bus maps published by the bus companies and the locations of bus stops were also added.

Table 1 Structure of dataset used in this research

Minor district	Population	Library					Total
		1	2	.	17	18	
1	2,134	36	46	.	0	1	758
2	3,721	35	108	.	1	0	1,076
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
1,712	2,457	0	.	.	38	75	122
Total	3,319,842	66,114	14,783	.	21,872	12,953	399,011

3.2 Elements of Travel Costs

Most of the libraries in Yokohama do not have any designated parking lots, because users are recommended not to come by car. Therefore, the means of transportation assumed for library visits were walking, buses, and railways. Table 2 provides a list of variables. One would also expect a large fraction of users to come by bicycle, but statistics differentiating pedestrians and bicycle riders were not available, and so bicycle riders were counted as pedestrians. Thus, we may be justified in assuming a fairly small value for the travel cost parameter for walking. Also, the train directions toward Yokohama Station were classified as *nobori* (up-train) and all other directions were classified as *kudari* (down-train). The directions for other railways not connecting directly to Yokohama Station were considered *nobori* if they were toward Tokyo.

3.3 Predicting Travel Routes

The parameter α_t for the *choice model* and V_m for the attraction of the library were then estimated. No data for the users' travel routes were available. It was assumed that the users would select the route with the lowest travel costs, and the routes from residential location i to library m ($m = 1,\ldots,16$) were predicted. Specifically,

Table 2 Explanatory variables for travel costs

Variables	Definition
Z_{1im}	Walking distance from residence i to facility m (the vicinity of one's residence) (m)
Z_{2im}	Walking distance from residence i to facility m (the vicinity of a facility) (m)
Z_{3im}	The influence of the slope of a street (the altitude difference covered on foot) (m)
Z_{4im}	Travel distance by railways from residence i to facility m (*nobori*: up-train) (m)
Z_{5im}	Travel distance by railways from residence i to facility m (*kudari*: down-train) (m)
Z_{6im}	The number of railway transfers from residence i to facility m
Z_{7im}	Travel distance by buses from residence i to facility m (m)
Z_{8im}	The number of bus transfers from residence i to facility m

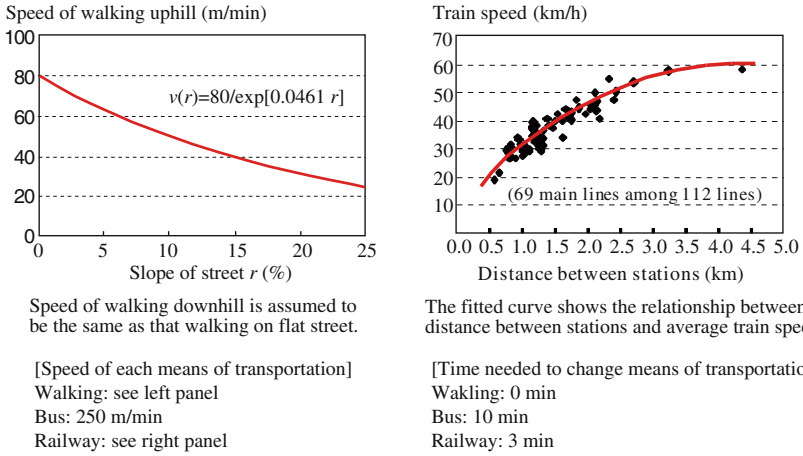


Fig. 2 Speed of each means of transportation for predicting the provisional route

(1) the speed of each means of transportation and the time needed to change from each means of transportation to another were set (Inoi and Nakaoka 2007; Geospatial Information Authority of Japan 1998; Osaragi 2009) (Fig. 2); and (2) the route offering the shortest travel time, combining walking, buses, or railways, was predicted. The predicted route was assigned to users and the travel time Z_{tim} was calculated. This time was used to calculate the parameters α_t and V_m for the *choice model*.

3.4 Convergent Calculation of Travel Cost Parameter

The parameter α_t described in the previous section was estimated from the provisional route based on an exogenously determined value. Once α_t had been estimated, this parameter was expected to provide results much closer to reality in the choice of route. (1) The estimated value of α_t was used for a second prediction of the route from residential location i to library m with the lowest travel costs, and (2) the value of Z_{tim} was found from this route, and α_t and V_m were re-estimated for the *choice model*. Processes (1) and (2) here were repeated until the parameter estimates converged.

3.5 Incorporating Local Population Characteristics

The utility function for the *use model* includes user attributes (see Sect. 2). Due to privacy considerations, however, the user attribute information was not included in the original data. A utility function V_{is} based on the attributes of the population of

Table 3 Explanatory variables for local population characteristics

Variables	Definition
Z_{1i}	Fraction of residential population over 65 years of age (%)
Z_{2i}	Fraction of workers who are technical experts/engineers (%)
Z_{3i}	Fraction of office workers (%)
Z_{4i}	Fraction of workers in the service industry (%)
Z_{5i}	Fraction of the residential population who are workers (%)
Z_{6i}	Fraction of the residential population who are school children/students (%)
Z_{7i}	Fraction of household who are workers (%)
Z_{8i}	Floor area of housing per person (m^2/person)

the minor district was created to replace the missing individual user information (Table 3). Namely, census data were used for describing personal attributes at the neighborhood-level. This is the limitation of the present research. Further study is, therefore, needed to confirm the reliability of the proposed models using a dataset including personal attributes of users.

After the preparations above, the values for parameters in the *use model* (β_t, V, and λ) were estimated.

3.6 Accuracy of Predictions by Use-Choice Models

The calibrated *choice model* and *use model* were combined to create the *use-choice model* Eq. (4). Figure 3 presents the accuracies of these models. The *choice model* had an extremely high description accuracy for the minor districts (Fig. 3a); however, the descriptive power of the *use-choice model* was somewhat low (Fig. 3c). This was due to the low descriptive power of the *use model* (Fig. 3b).

One reason for the unreliability of this model is that the individual user attribute data were replaced by the aggregated data (census data) for the appropriate minor district population. Still, the estimated values for the overall number of users aggregated for each library clearly showed quite high reliability (Fig. 3d and f).

4 Interpretation of Estimated Parameters

Table 4 shows the estimated values of parameter α_t for the *choice model*. The travel costs largely depended not only on the travel time and distance, but also on the travel direction and on transfers between means of transportation.

Examination of the travel burden of walking (α_1, α_2, α_3) reveals that walking in the vicinity of the library had about 1.2 times the resistance of walking in the vicinity of home. In other words, this indicated that the distance to the nearest train station or bus stop in the vicinity of the facility had a great influence on the facility

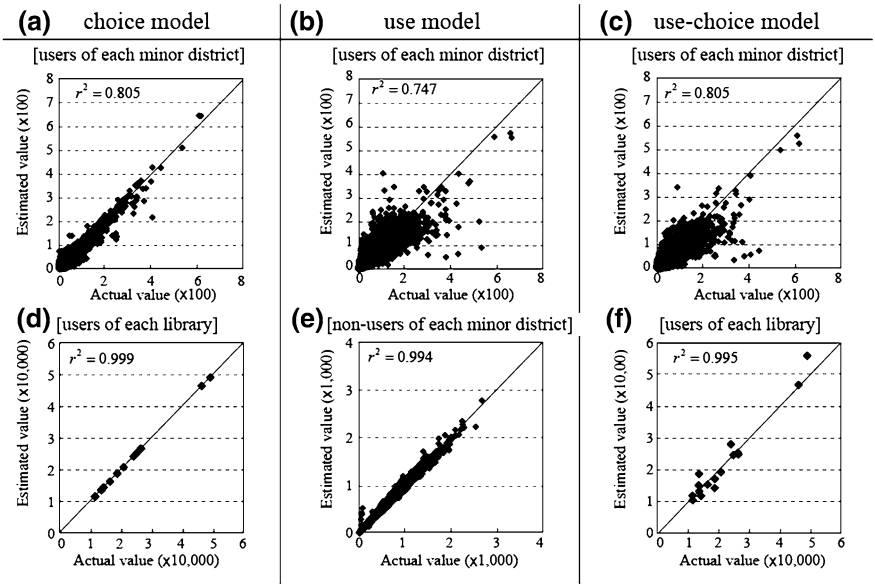


Fig. 3 Validation of choice model, use model and use-choice model

Table 4 Estimated parameters of travel costs

Parameters	Definition	Estimated values	*t*-values
α_1	Walking distance in the vicinity of the residence	-8.380×10^{-4}	−189.9
α_2	Walking distance in the vicinity of the library	-9.961×10^{-4}	−146.5
α_3	The influence of the slope of a street	-9.890×10^{-3}	−51.02
α_4	Travel distance by railways (*nobori*)	-0.797×10^{-4}	−55.74
α_5	Travel distance by railways (*kudari*)	-2.229×10^{-4}	−121.4
α_6	Number of railway transfers	−2.503	−203.3
α_7	Travel distance by buses	-6.928×10^{-4}	−228.7
α_8	Number of bus transfers	−0.272	−31.50

choice behavior of the user. Considering hill resistance, the travel costs of walking 1 m of vertical distance equaled the travel costs of walking 11 m horizontally; thus, the presence of hills was a factor that increased travel costs.

In the case of the travel burden of buses (α_7, α_8), the travel costs of the bus are about 0.83 times that of walking (near home), but those of a single bus transfer correspond to a travel distance (walking) of 300 m or more.

Examination of the travel burden of trains (α_4, α_5, α_6) reveals that the travel costs in the *nobori* direction were about 0.12 times those of buses, whereas those in the *kudari* direction were about 0.32. In other words, *kudari* travel had about 2.7 times the cost of *nobori* travel. The *nobori* direction was often the direction for going to work or school, making it convenient for riders to stop off on the way and

Table 5 Estimated parameters of local population characteristics

Parameters	Explanation	Estimated values	*t*-values
λ	Log-sum	0.440	180.60
β_1	Fraction of residential population over 65 years of age (%)	-2.49×10^{-1}	−63.56
β_2	Fraction of workers who are technical experts/ engineers (%)	6.81×10^{-3}	22.94
β_3	Fraction of office workers (%)	9.08×10^{-3}	25.66
β_4	Fraction of workers in the service industry (%)	-1.11×10^{-1}	−16.70
β_5	Fraction of the residential population who are workers (%)	-1.11×10^{-1}	−20.03
β_6	Fraction of the residential population who are school children/students (%)	-1.36×10^{-1}	22.71
β_7	Fraction of household who are workers (%)	-2.32×10^{-1}	27.27
β_8	Floor area of housing per person (m^2/person)	-1.90×10^{-3}	38.33
V	Attraction level of a facility	−2.49	−63.56

reducing the travel resistance for running errands. Also, there was a quite high cost for transfers between trains. Thus, for people using a facility, there was a high resistance associated with transfer between trains.

Table 5 shows the estimated values for the population attribute parameters. Facility use shows distinct variation with the characteristics of users. In minor districts that contained higher fractions of students or home-based populations, people were more likely to use facilities, while in those with large populations of residents over 65 years of age, people were less likely to use facilities.

5 Characteristics of Catchment Area

The catchment areas of some libraries predicted by the *use-choice model* were compared with the actual catchment areas, providing some observations about the influence of travel costs on catchment area.

1. Naka Library

Naka Library is the only library located distant from a train station. The contours of the equivalent travel costs to the library did not take the shape of concentric circles, as is predicted by the Euclidian distance approach for estimating travel costs; minor districts along bus routes to the library had low travel costs (Fig. 4c). The spatial distribution of the users clearly shows that the bus routes affected the catchment area; this is also well reproduced in the *use-choice model* (Fig. 4a and b).

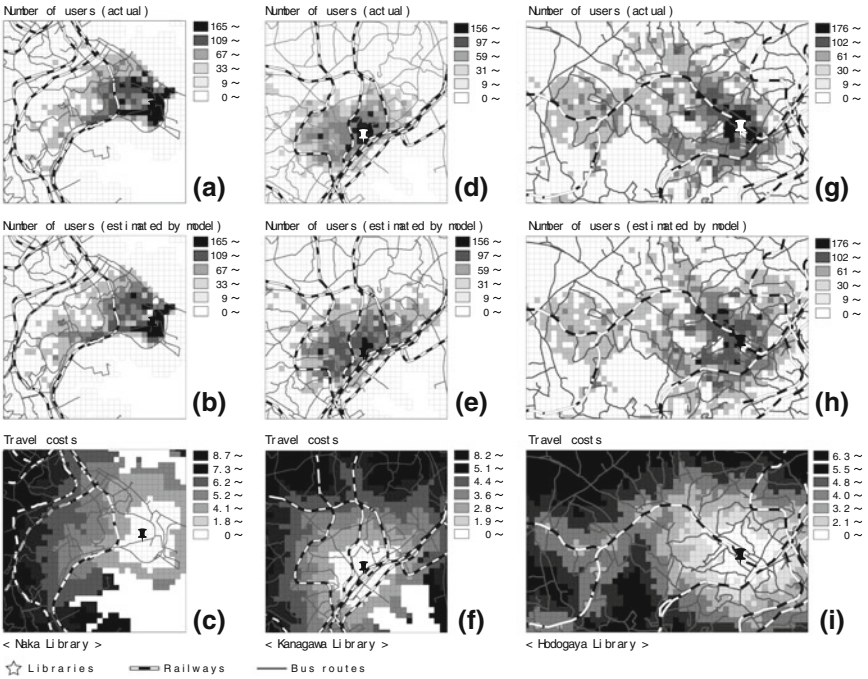


Fig. 4 Spatial distribution of catchment area and travel costs for each library

2. Kanagawa Library

The approach routes to the south-east corner of the Kanagawa Library were blocked by the railway and highways, and the approaches north of the library slope uphill. Therefore, the travel costs from the south-east and from the north are high (Fig. 4f), and the catchment area does not extend very far in the north or south directions (Fig. 4d). The *use-choice model* clearly shows these characteristics for the catchment area of the library (Fig. 4e).

3. Hodogaya Library

Hodagaya Library is situated next to a railway that runs east–west (the eastward direction is toward the CBD). The travel costs toward the CBD are lower than those in the opposite direction, and so are low for the minor districts along the railway west of the library (Fig. 4i). Accordingly, the catchment area is not a circle centered around the library, but rather, an irregular "comet" shape, expanding with distance from the CBD (Fig. 4g). This method provides a much more accurate description of the characteristics of the catchment area of this library, which would never be described by a conventional model based on the Euclidian distance, travel time, or other similar factors.

6 Predicting Number of Users and Catchment Area of New Libraries

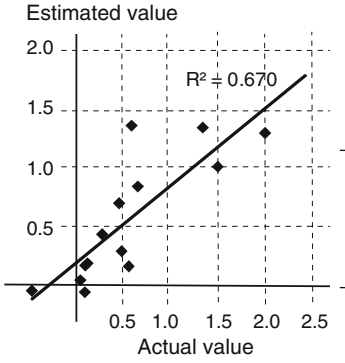
6.1 Construction of a New Library

Before constructing a new library, it is necessary to identify its catchment area and the number of expected users. One can also expect that the catchment areas and user numbers of neighboring libraries will be greatly affected. Below are some observations of the utility based on this *use-choice model* for predicting the conditions before and after the construction of a new library. Specifically, a *use-choice model* based on data from 1994 was employed to predict the number of users and the catchment areas for 1997 of two new libraries (Tsuzuki and Midori) constructed in 1995 and an older nearby library. The predictive power of the *use-choice model* for making such predictions was examined.

6.2 Values of the Attraction of New Libraries

The value of the attraction V_m must be established in order to predict the catchment area and the number of users of newly constructed libraries. The following linear model was created to describe V_m on the basis of the scale of the libraries and of their surroundings. Figure 5 provides details about the explanatory variables.

$$V_m = \gamma_1 x_{1m} + \gamma_2 x_{2m} + \gamma_3 x_{3m} + \gamma_4 \tag{7}$$



	Explanatory variables	Estimated values
γ_1	Number of employees of library	0.1483
γ_2	Total commercial floor area (m2)	7.800x10-6
γ_3	Total yearly restaurant sales (JPY)	9.864x10-7
γ_4	Constant	-1.596

Fig. 5 Validation of a model for estimating the attraction of libraries

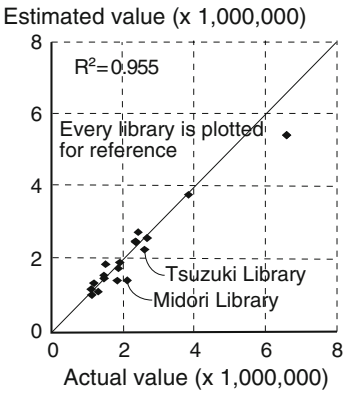


Fig. 6 Validation of numbers of users of new libraries

6.3 Predicting the Numbers of Users of New Libraries

It was attempted to predict the use and choice behavior as of 1997 with the *use-choice model*. Although the numbers of users in the minor districts were underestimated, the comprehensive estimates for users of each library were close to the target values (Fig. 6). The reason for the underestimates was that the attraction of the new libraries had been underestimated. An issue to be addressed in the future is to make more accurate estimates of the attraction of libraries.

6.4 Prediction of Catchment Area of New Libraries

Figure 7 shows the mean travel costs from each minor district (the sum of the products of the travel costs to each library multiplied by the probability of use) in 1994 and 1997. Comparing the two years, the reader can see that the construction of the two libraries greatly reduced the mean travel costs in the vicinity of the libraries and along the railways.

Figure 8a and b show the sizes of the actual catchment areas in 1994 and 1997, respectively. The reader can see that the minor districts where the mean travel costs fell by a large amount (the library neighborhoods and areas along the railways) also saw large increases in the numbers of users. Figure 8c presents the sizes of the catchment areas predicted by the *use-choice model* for 1997. Comparing this with the actual catchment areas in Fig. 8b, we see that the large changes from 1994 to 1997 were well predicted. In other words, the *use-choice model* provides dependable predictions of changes in user behavior that accompany the construction of new libraries. These results indicate that changes in catchment area and user number can be estimated with high accuracy.

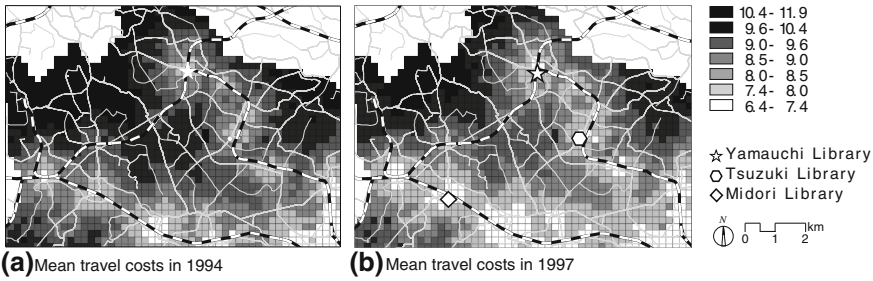


Fig. 7 The mean travel costs from each minor district to libraries

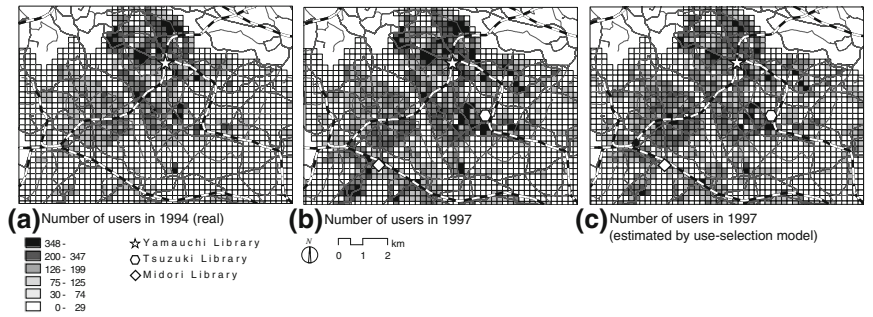


Fig. 8 Change of the catchment area after the construction of new libraries

7 Summary and Conclusions

This study produced a formulation for travel costs that vary according to the means of transportation, the direction of travel with respect to the CBD, the number of transfers between one travel means and another, and hill resistance. A *use-choice model* incorporating these travel costs was employed to estimate the behavior of library users, and the accuracy of this model was validated. The estimates for individual minor districts (250 × 250 m) were not very accurate, but the predictions for the catchment area and total number of users for each library were quite acceptable in accuracy. Also, the parameters used for these estimates provided a quantitative grasp of how several factors affect the shapes of catchment areas: the continuity of the street network, routes used in daily life and travel toward or away from the CBD, the potential for use of public transportation, and geographical characteristics such as hills. An analysis was also carried out on numbers of users and catchment areas before and after construction of new libraries; this analysis confirmed that the *use-choice model* is appropriate not only as a descriptive model but also as a predictive model. Specifically, it was shown that this *use-choice model* provides highly accurate predictions of the dynamic variations in the catchment area and user numbers of new and existing facilities, information which is vital during planning a new facility.

It is possible that the model described by this study can be applied by city officials, public health authorities, and other bodies to projects officially designated for certain uses. For example, this model could be used to examine how fair the location of a facility is on the basis of travel costs. Furthermore, accessibility measures can be of great value in geomarketing-analysis (Cliquet 2002; Peterson 2004; Schüssler 2006) as well as regional and environmental planning. High-detailed spatiotemporal distribution on human activities based on precise accessibility measures has great potential to extend the fields of gomarketing-analysis.

Acknowledgments The authors would like to acknowledge the valuable comments and useful suggestions from anonymous reviewers to improve the content and clarity of the chapter. A part of this paper was presented at Journal of Architectural Planning and Engineering, AIJ, "Travel costs considering means and direction of movement within a city incorporated with a model of facility choice behavior", Vol. 77, no. 676, pp. 1293–1300, 2012 (in Japanese).

References

Apparicio P, Abdelmajid M, Rival M, Shearmur R (2008) Comparing alternative approaches to measuring the geographical accessibility of urban health services: Distance types and aggregation-error issues. Int J Health Geograph 7(7):1–14

Arentze T, Borgers A, Timmermans H (1994) Geographical information systems and the measurement of accessibility in the context of multipurpose travel: a new approach. Geograph Syst 1:87–102

Burns LD (1979) Transportation, temporal, and spatial components of accessibility. Lexington Books, Lexington

Cliquet G (2002) Geomarketing—methods and strategies in spatial marketing, ISTE

Comber A, Brunsdon C, Green E (2008) Using a GIS-based network analysis to determine urban greenspace accessibility for different ethnic and religious group. Landscape Urban Plann 86(1):103–114

Ettema D, Timmermans H (2007) Space–time accessibility under conditions of uncertain travel times: theory and numerical simulations. Geograph Anal 39:217–240

Geospatial Information Authority of Japan (GSI) (1998) Instruction book of detailed digital information, Japan Map Center

Green C, Breetzke K, Mans G (2009) GIS based evaluation of public facility provision to achieve improved governance and equitable service delivery. Geo multimedia 2009: 14th international conference on urban planning and regional development in the information society, pp 1–94

Hsu C, Hsieh Y (2004) Travel and activity choices based on an individual accessibility model. Papers in Regional Science 83:387–406

Inoi H, Nakaoka R (2007) Research on the palliation of physical burden from slope by community bus. Infrastructure planning review, Japan society of civil engineers, vol 36. CD-ROM

Nakamura K, Kurihara K (1998) The way of planning the library system to regional area: Fundamental investigation for the planning of library system to community area (12). J Arch Plann Eng 512:123–130 in Japanese

Neutens T, Delafontaine M, Schwanen T, Van de Weghe N (2012) The relationship between opening hours and accessibility of public service delivery. J Transp Geogr 25:128–140

Osaragi T (2002) Accessibility evaluation: effects of free return system on users' behaviour of public libraries. Env PlannB: Plann Des 29(5):637–654

Osaragi T (2009) Estimating spatio-temporal distribution of railroad users and its application to disaster prevention planning (Lecture notes in geoinformation and cartography). In: Sester M et al. (eds) Advances in GI Science, Springer, Berlin, pp 233–250

Peterson K (2004) The power of place—Advanced customer and location analytics for market planning, Integras

Satoh E, Yoshikawa T, Yamada A (2008) Examination of continuity of local living based on the converted walking distance: Model for location planning of regional facilities considering topographical condition and aging society Part 2. J Arch Plann Eng 625:611–618

Schüssler F (2006) Gemoarketing—Anwendung Geographischer Informations-systeme im Einzelhandel, Tectum

Teixeira JC, Antunes AP (2008) A hierarchical location model for public facility planning. Eur J Oper Res 185(1):92–104

Wang F, Luo W (2005) Assessing spatial and nonspatial factors for healthcare access: Towards an integrated approach to defining health professional shortage areas. Health Place 2:131–146

Design Principles for Spatio-Temporally Enabled PIM Tools: A Qualitative Analysis of Trip Planning

Amin Abdalla, Paul Weiser and Andrew U. Frank

Abstract Current personal information management (PIM) tools do not sufficiently recognize the spatio-temporal, hierarchical, or conceptual relations of tasks that constitute our plans. Using behavioral observation methods we analyzed people planning a trip to attend a conference taking place in a region they had little or no prior familiarity with. The resulting open-ended records were coded into higher-level segments and categories. These served as a basis for a cognitive engineering approach, to propose better design principles for spatio-temporally enabled PIM-tools.

1 Introduction

Research on personal information management (PIM) is concerned with the […] *effort to establish, use and maintain a mapping between need and information* (Jones and Teevan 2007). Traditionally, research in this field was very much concerned with the organization of documents, pictures, bookmarks, etc. (Barreau and Nardi 1995; Jones and Tevaan 2007). But more recently voices called for a focus on *prospective memory* (Sellen and Whittaker 2010), i.e., the memory that stores the tasks and errands we are supposed to do in future. Tools that support our prospective memory are applications like calendars, todo-lists, etc. Such tools are, according to Norman (1991), cognitive artifacts, i.e., *artificial device(s) designed to maintain, display, or operate upon information in order to serve a representational function*.

A. Abdalla (✉) · P. Weiser · A. U. Frank
Department of Geodesy and Geoinformation, Research Group Geoinformation, Vienna University of Technology, Gusshausstr 27–29, 1040 Vienna, Austria
e-mail: abdalla@geoinfo.tuwien.ac.at

P. Weiser
e-mail: weiser@geoinfo.tuwien.ac.at

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

D. Vandenbroucke et al. (eds.), *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography,
DOI: 10.1007/978-3-319-00615-4_18,

Currently, these tools are relatively passive, because they provide little support in planning or monitoring our tasks. Their capabilities are mostly limited to the storage and display of information. Interactive features, such as alarms, do not take contextual changes into account. A dynamic notification in contrast, would automatically modify a set alarm if necessary. For example, in case of a traffic delay, it would tell the user to depart earlier. Research has shown that PIM-tools can greatly benefit from the integration of spatio-temporal and semantic information (Raubal et al. 2004; Raubal and Winter 2010; Janowicz 2010; Abdalla and Frank 2012). Calendars, for example, could be improved if they had a sense of space-time built into them. Currently, they allow creating two events that are geographically unreachable within the specified time the events are apart (e.g., Event A, 01.02.2013, 10 a.m. in New York and Event B, 01.02.2013, 3 p.m. in Berlin).

We argue that current PIM-tools require a change in representation, i.e., get a sense of space, to allow for a proactive support of our daily activities. The goal of this work was to gain valuable insights into the cognitive nature of the planning process and try to infer important principles that can serve as a basis for the design of spatial PIM-tools. This can be seen as a cognitive engineering approach, proposed by Norman (1986). The aim is to *devise systems that are pleasant to use-the goal is neither efficiency nor ease nor power, although these are all to be desired, but rather systems that are pleasant, even fun.*

In this work we analyzed people planning a trip to a scientific conference, applying behavioral observation methods. The results, open-ended records, were coded and provided a semi-formal segmentation and categorization of the planning process as the basis of our study.

The main questions we addressed are:

1. What are the (prominent) activities people carried out to plan the trip?
2. Do people share a common temporal ordering of their planning activities?
3. What is the nature of the information people used?
4. How can (computational) tools help to overcome difficulties during the planning of a trip and its execution?

The remainder of the work is structured as follows: In Sect. 2 we discuss relevant theories of planning from the cognitive sciences. Section 3 gives a detailed account of our study design and Sect. 4 explains the methods underlying our data analysis. Section 5 presents the result of the analysis and the last chapter lists the proposed design principles that stem from our findings.

2 Related Literature

From a cognitive standpoint, trip planning is the solving of a problem with an initial state, a set of transformations, and a goal state. Mayer (1990) defines problem solving as “cognitive processing directed at transforming a given situation into a goal situation when no obvious method of solution is available to the

problem solver". By doing so, we usually transfer information from past experiences or general knowledge onto the current problem to be solved. As a consequence, there is a considerable overlap between problem solving and transfer (Eysneck and Keane 2010). More knowledge in a given domain increases the effectiveness of problem-solving (expertise). This also plays a role in selecting the best option among several choices.

In general, problems can be classified into well-defined and ill-defined problems. Well-defined problems are those in which the initial state, all possible steps to achieve a solution are clearly laid out and the final goal state is specified (Eysneck and Keane 2010). The board game Backgammon is an example for a well-defined problem. There is only one start configuration, a finite set of legal moves and only one goal state (first to remove all their pieces from the board is to win). In contrast, trip planning is an ill-defined problem. Although, start and goal state are clearly laid out, the number of possible solutions to reach it is potentially infinite.

If one is to decide upon a possible solution to a plan, or a part thereof, different strategies may be applied. One of them is the domain principle (Gilhooly 1996) stating that "if option A is at least as good as option B in all respects and better than B in at least one aspect, then A should be preferred to B". However, this is not always the case. As Kahneman and Tversky (1979, 1984) have shown, people are potentially more sensitive to losses than to gains (loss aversion). This may influence decision making and produce results that stand in contrast to the domain principle. Similarly, the "sunk cost effect" lets people often continue an endeavor once an investment in money, effort, or time has been made (Arkes and Ayton 1999) even if it would be economically better to abandon the plan.

If there is more than one possible solution one might apply one of two strategies: Multi-attribute theory (Wright 1984) or elimination by aspects (Tversky 1972). The former lets people weight attributes relevant to a decision, finally selecting the solution that offers the highest summed weight. The latter lets decision makers eliminate options by considering one relevant attribute after another. For example, consider all the hotels of a city you are interested in, and then eliminate the ones that are not within walking distance from a particular place. Then further reduce the number of options by specifying a price limit. Continue until you have found one solution.

Also, some problems cannot be solved at all, because they are represented in a way that makes it impossible or very hard to retrieve a possible set of actions towards the goal. This mental block can only be broken if the representation is changed (Knoblich et al. 1999). Ohlsson (1992) in his representational change theory noted that such a change of representation can be achieved by (1) adding new problem information (2) re-encoding of information, or (3) constraint relaxation (what was not allowed before is now permitted). We argue that current PIM tools need a change of representation to allow for a better support in problem solving and plan execution.

3 Study Design

Our study investigated the process of planning a trip to attend a scientific conference. It was organized and hosted by the University of Tartu, Estonia and took place from the 22nd of August until the 25th of August 2012. The venue seemed particularly interesting, since only one of the participants has visited the country before and hence was familiar with its geography or transportation network. Detailed information about the conference was provided by the official conference website.[1] This included textual information on how to get there, a range of hotel options, the time table, etc. The participants were asked to plan the trip minimizing costs and time. The study was conducted around two weeks ahead of the conference. Participants were asked to comment and justify the actions they were carrying out, so their intentions became clear to the authors. Also, they were provided with pencil and paper. The computer screen was filmed and notes were taken by the authors throughout the entire process. To simulate the booking of flights or hotels the participants indicated to the observing person whenever they had made a final decision.

3.1 Study Participants

The subjects of the study (N = 10) were recruited from staff members of our institute and from the authors' family and friends. Half of the participants had a background in geospatial technologies. The age ranged from around 20–50 years and the sex ratio was 50:50. Experience in travelling and conference attendance among the participants varied from close to zero (never planned a trip abroad themselves) to very experienced (regular traveler to foreign places). Only one of the subjects has actually ever been to the country the conference took place. All other participants had no prior familiarity with the region of interest.

4 Data Pre-Processing

The observations resulted in a set of videos, transcripts and notes by both the participants and the observers. Because such kind of qualitative data is rather unstructured, the first task was to decide upon a rigorous analytical framework that allowed breaking the content into units that could be quantified. We followed the *coding* approach described by Montello and Sutton (2006). The method develops incrementally beginning with a *segmentation* of the raw data (videos) into "minimum meaningful entities". These can vary in nature, e.g., they can be words,

[1] http://www.ut.ee/mobiletartu/2012 (last accessed on 9th of November 2012).

phrases, sentences, situations, or actions (See Sect. 4.1 for a list of the segments). Once the minimum meaningful entities are defined they are grouped together into conceptually related categories. This resulted in analyzable quantifiable data.

4.1 Segmentation

For the segmentation we defined units that constitute the *meaningful entities*. Considering the questions posed in the first section and intensive preliminary analysis of the videos and transcripts led to the identification of a number of cognitive activities that constitute our segments. The following paragraphs describe them in detail:

- **Information-extraction**: Whenever a participant acquires information relevant for the planning process. It is indicated by looking at a text or browsing through a list or table. Additional hints for an ongoing information extraction are comments made by the subjects, e.g., […] *now I'm trying to find out what time the flight leaves*. Information-extraction increases the person's knowledge based on existing information. *Information-extraction* does not represent a single segment, but is an umbrella term for a group of segments. Each of these segments is determined by the kind of information extracted. Because we cannot explicitly list all segments subsumed by the term, they are represented by the lattice structure shown in Fig. 1. For example, information can be extracted *about* an event, a connection, or a place. Each of these can further be instantiated by a more specific description, e.g., event-conference, connection-flight, place-hotel. On the most detailed level, each instance of information can be of a specific *type, e.g.*, temporal, spatial, price, or place name. For example, the segment "information extraction-connection-flight-price" refers to extracting

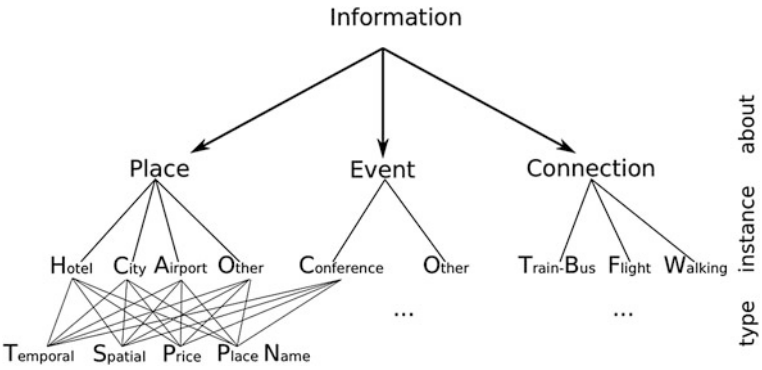


Fig. 1 The lattice structure illustrates all the different segments grouped under the term Information-Extraction

price information about a specific flight connection. This way we obtain a more nuanced view on extraction activities.

- **Information-input:** Whenever a participant feeds information into a data processing system (e.g., website) based on personal preferences and the task description. For example, selecting the departure date for a connecting flight.
- **Comparison & Ranking:** In this segment, the subject considers different options and attempts to rank one over the other and may apply some of the strategies mentioned in Sect. 2. This activity is indicated by verbal expression and/or by a constant back and forth switching between websites with comparable content.
- **Spatial-/Temporal-/Spatio-Temporal Inference:** An inference is mainly noticeable through verbal comments made when a participant concludes that something is relevant to the planning process based on observed facts. For example, […] *because the flight arrives very late and there are no connecting buses, I need to book a hotel for the night*. Inferences increase the subject's knowledge by inferring facts from given information.
- **Postponement:** This is the case when certain tasks are decided to be conducted in situ (e.g., buy tickets on the bus) or shortly before departure (e.g., order an airport shuttle) instead of planning them in advance.
- **Booking:** This is the activity of a final decision. Participants in the study were asked to clearly state when and what they were about to "book", e.g., a train ticket.
- **Consideration:** The act of determining a possible solution to a part of the plan. This is indicated by taking notes, possibly used for later comparison, and/or verbal comments.

4.2 Categorization

In the next stage, we grouped conceptually related segments together. We opted for a two level categorization. On the lowest level, 9 mutually exclusive categories cover the range of all segments. On the second level, three categories entail the lower category level (see Table 1). Higher order goals (category level 1 and 2) are achieved by lower level activities. As mentioned in the section before, we distinguish between different forms of *Information-Extraction*. In general, it has the goal to form an understanding or a description of a situation. We defined four different forms of understanding and one description:

- Event Understanding
- Network Understanding
- Place Understanding
- Spatial Understanding
- Task Description

Table 1 Activity Categorization

Segments (IsAbout, InstanceOf, Type)	Category level 1	Category level 2
Info extraction—event, conference, […]	Task description	Knowledge aquisition
Info extraction—event, […] \conference	Event understanding	
Info extraction—connection, […] \spatial	Network understanding	
Info extraction—place, […] \spatial	Place understanding	
Info extraction—[…], spatial regional	Spatial understanding	
Info extraction—[…], spatial local		
Info input—[…]	Query	
Spatial inference	Inference	Reflection and evaluation
Temporal inference		
Spatio/temporal inference		
Comparison and ranking consideration	Filtering and storing	
Booking postponement	Decision	

For example, extracting information about the conference contributes to the task description. Extracting spatial information (i.e.,: looking at a map) contributes to Spatial Understanding. The three types of inferences (i.e.,: spatial, temporal and spatio-temporal) are grouped into the Inference category. All forms of inputs are categorized as Query, and the activities of Comparing and Ranking as well as Consideration fall into the Filtering/Storing category. Finally, the segments Booking and Postponement form the Decision category. Table 1 gives a structured overview of the grouping mechanism applied to the segments. Please note the special syntax for the segments:

SegmentName—IsAbout, OfInstance, Type

The symbol "[…]" indicates that all subsequent levels are included. For example:

> "Info Extraction—Place, Hotel, […]" refers to "Info Extraction—Place, Hotel, [Temporal, Spatial, Price, Place Name]".

The symbol "\" stands for without. For example:

> "Info Extraction—Place, Hotel, […] \ spatial" refers to "Info Extraction—Place Hotel, [Temporal, Price, Place Name]".

5 Analysis

5.1 Data Analysis

Generating discrete data sets allowed us to aggregate and analyze it. The first question we addressed was "Are there any prominent activities evident?". Figure 2 illustrates the absolute number of the activities (See Sect. 4.2) per subject.

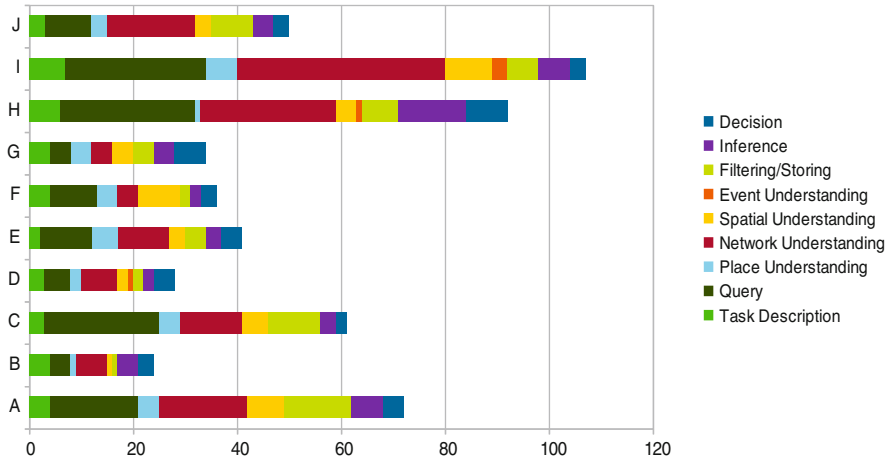


Fig. 2 Absolute number of segmented activities for each participant

It shows that the amount of effort spent for the task varied tremendously. It can be seen that experienced travelers (Subjects D, F, and G) tended to conduct less activities than the others. This can be explained by their strategy of postponing parts of the planning process to be carried out shortly before the trip started, e.g., how to get from the hotel to the conference. But there are outliers that do not support the statement, such as subject A, who can be considered to be an experienced traveler but nevertheless put much effort into the planning, as well as subject "B" who does not belong to the group of experienced travelers, but still was very brief.

In Fig. 3 each individual activity is shown relative to the sum of all activities. At that point it is important to state that the activities do not say anything about the temporal extent of the process.

The prevalence of certain activities varies from subject to subject, although some seem to be consistently prominent. Those are *network understanding* and *querying*. Thus, a lot of effort is put into understanding the transportation network that serves as a basis for the trip. The second important activity is querying databases to seek information that provides the foundation for the various *understandings* as defined in level 1 of the categorization table (See Table 1). In contrast, there are activities that are sometimes not present at all, such as *event understanding*. We believe that this is due to the simulated character of the study, resulting in most of the participants to consider it not necessary to check any temporal overlaps with their real schedule.

Another dimension of interest is the temporal ordering (See Question 2 in Sect. 1). Many of the activities were conducted in an unordered manner and were reoccurring. Thus, we used a method that allowed to make assertions about the general ordering of the activities. For each activity a number was assigned, representing its position in the ordered sequence of activities. We then took the first

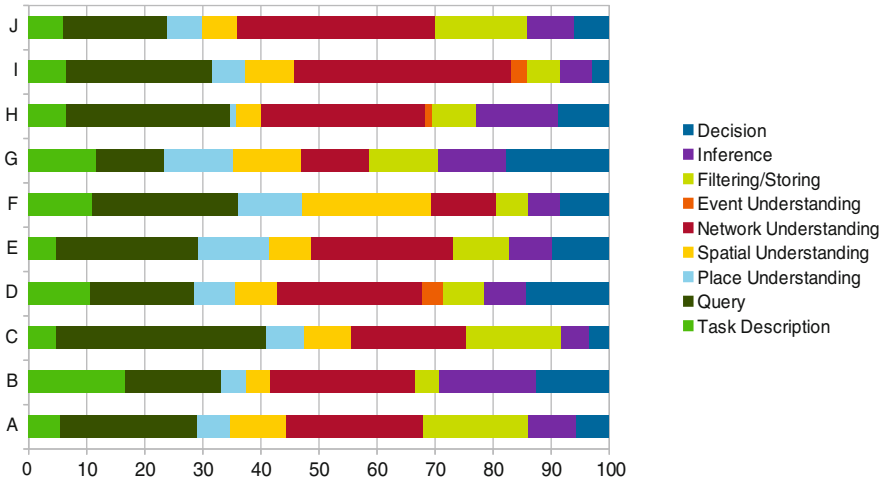


Fig. 3 Relative number of segmented activities for each participant

occurrence of each activity, i.e., the minimum value of all the positions that are occupied by it. That gave us the first occurrence of every activity in the complete list, from which we were then able to compute the relative order in comparison to the other activities. This was done for each subject. The result is visualized in Fig. 4.

Each line depicts an activity, while the axes stand for the subjects. The closer to the center a line is, the earlier the activity had its first occurrence in the overall process. Obviously, the activity that for all the participants has been the first thing to

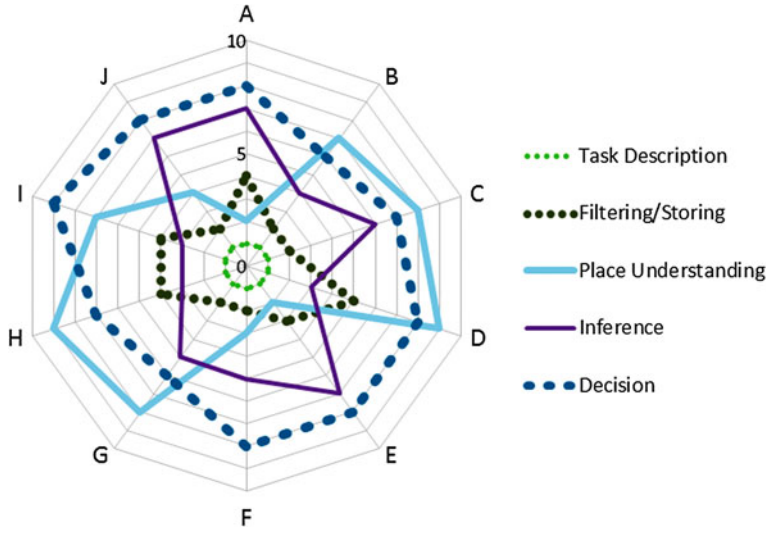


Fig. 4 Temporal order (*first occurrence*) of the various segments for each subject

do, was to define the goal (task description). From there on the picture appears to be more complicated and a clear delineation of the ordering is hard to find. Although the subjects vary in their sequential ordering, some of the activities took place before others. The first Query-activity, that is the second closest to the center, was by every subject conducted before the filtering and storing as well as the Decision-activities. Mostly decisions were made after each of the other activities was conducted at least once. We conclude that some of the activities imply functional dependencies. An activity that seems to vary strongly is Place Understanding.

A question we posed in Sect. 2 was concerned with the nature of information utilized to solve the problem. We recognized soon, in conformance with previous research (Timpf et al. 1992), that people plan on different levels of granularity. Therefore, we differentiated between two scales of spatial information, that is, local and regional. Figure 5 shows the absolute amount of times a map was looked at, distinguished by local (street level) and regional scale (country level).

Three of the subjects (B,G,J) did not look at a *regional scale* map at all. Hence, they did not build up a large scale spatial/topological representation of the region and places relevant for their search. It is interesting to note that none of them had a detailed familiarity with the region of interest.

Just like spatial information, temporal information was used on different granularities. Every subject started noting the temporal extension of the conference on a coarse *date*-level (i.e.,: 22nd–25th of September). However, at the point of looking for a flight, subjects refined the task description to an exact time on an hourly basis (i.e.,: 22nd at 11 am). Only one subject did book a flight based on dates, assuming one day before and after the conference allows for sufficient time to reach the conference and return to the airport on time. Hotels were only looked at based on date information, check-in times were hardly considered.

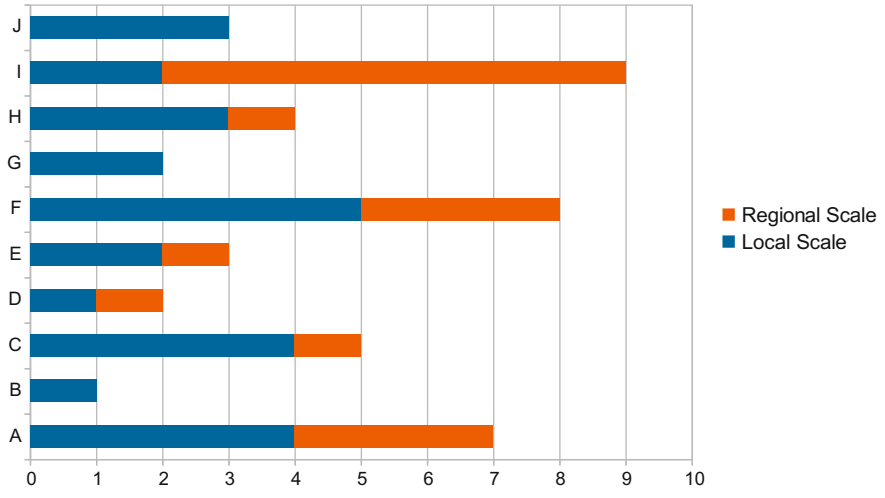


Fig. 5 Number of times subjects looked at a map, divided in regional and local scale

5.2 Observation Based Findings

The process of coding allowed us to quantify an open ended data-set, by cutting the continuous stream of information into discrete entities. Information as relations between the different activities or valuable insights became apparent throughout the observation. In the following we present the findings that are based on the observations made by the authors while conducting the study.

Information Transfer: The most striking cognitive activity was the extraction and feeding of information from different sources into various query interfaces. It was also one of the most error prone activities. 4 out of the 10 participants did err in the querying and booking process of flights, caused by wrongly put date information. This problem often accumulated throughout the process into a vast amount of different websites that contained relevant or irrelevant information. Some subjects, had at times, up to 20 tabs opened in the browser, among they attempted to extract or feed information.

Comparison and Inference: The next (related) evident problem was the comparison of different options that had to be done by constant back and forth switching between the tabs or websites. At some points subjects had opened two separate online maps, each of it containing a point of interest (such as a hotel- and conference venue). By back and forth switching between the maps they tried to see whether the points were close to each other. Inference was yet another activity, by examining a situation subjects inferred the need for something and defined subtasks. For example, by noticing that the flight departs early in the morning, it was inferred that there is a need for a hotel to stay overnight.

Place based versus Geographic Understanding: In Fig. 5 we showed that some subjects did not build up a geographic representation of the region. This lack of spatial knowledge resulted in a narrowed approach to the search for possible connections to Tartu. Almost all of those who had looked at a large scale map recognized that Riga is as close to Tartu as Tallin. Thus, they were able to look for solutions involving a flight to Riga and a connecting bus/train to Tartu. Those who did not bear such knowledge were not able to make that inference. This gap in geographic understanding of the situation even lead one subject to book a flight to Tallin, unaware of the fact that Tallin and Tartu are two different entities. The subject simply assumed that there is a taxi going from the airport to Tartu's city center. Interestingly, all of those who planned their trip solely based on city names, did not have an educational background in geospatial technologies.

Opportunities: Hayes-Roth and Hayes-Roth (1979) shaped the term *opportunistic planning*, describing the fact that people often recognize an opportunity in a plan to conduct a related or unrelated task. In our study it appeared that some subjects recognized such opportunities. In one case, it was noticed that there is a flight going via Brussels, what could have been an opportunity to meet friends who live

there. Consequently, such opportunities played a role in their weighing of a solution in comparison to others.

Assumptions: They build the basis for a lot of queries in the planning process. When people were unsure about things, they made assumptions. Facing a query interface of a flight search engine the following was stated: "[…] normally Tallin is cheaper, the capital is always cheaper". Since the subject was not sure what city to select, an assumption was made.

Postponement: When explaining the participants what to do, we always stated that they should plan the trip until they *feel* that it is "done". It occurred that for things which are still part of the trip and might have needed some prior preparation, people tended not to take care of it until very shortly before departure. Most of the subjects were satisfied by having a flight and hotel booked. They were not concerned on how to exactly get from the hotel to the conference venue, or how to go to the airport in their hometown.

Cognitive artifacts: In the observational part, we did not explicitly ask the subjects to use any sort of tools to support their planning. In fact, only very few used a calendar or similar tool. On the other hand, almost all of the subjects took notes on the provided sheet of paper. The purpose of the notes seemed to be twofold: (1) writing down the details of the conference (i.e., date, venue, etc.) and (2) comparing possible solutions.

6 Conclusion and Future Work

The aim of this study was to acquire a deeper understanding of a planning process that involves a clear set of spatio-temporally constraining factors. In this section we give a set of implications for the design of spatio-temporally enabled PIM-tools. We found 4 points that need to be addressed:

Goal based planning: Planning is a goal directed activity. It is therefore crucial that the goal is represented and defined well (task description), in order to come up with a successful plan. An application supporting the user needs to be able to store an (sufficiently) accurate representation of the goals. Both the temporal and spatial dimension plays an important role finding a good solution to the problem. As a result, an ideal application would let users set a task freely, but constrain them to choose only amongst solutions that "make sense". This helps to avoid errors, as those mentioned above, e.g., when users confused dates or places.

Planning is an evolving process: In general, several (evolving) plans are created during the problem-solving process. As an attempt to properly sequence or compare different legs of the trip people used cognitive artifacts (e.g., a piece of paper) to help them store spatial, temporal, or cost-related information. It is therefore

crucial that an application is capable of storing a partial (incomplete) plan. With this capability, other (reasonable) parts of a plan can be added, removed, replaced, or simply compared to each other. Also, it is important that the system is able to determine gaps or missing links in the plan (e.g., due to postponement by the user). Thus, a formal description for *partial plans* might be helpful, similar to the model proposed by Weiser et al. (2012).

Recognizing opportunities: An application should be able to recognize opportunities for other tasks, or simply suggest things based on other data sources. The example we presented in Sect. 5 showed that spatial information about a friend (her/his home address) was matched with a possible travel route and hence perceived to be an opportunity. Since such opportunities play an important role in the weighing of a solution it needs to be incorporated in the representation.

Planning over multiple granularities: Throughout the planning process subjects reasoned over multiple temporal and spatial granularities. As a result, a supporting system should be able to reason over coarse granularities (date or city level) but also be able to look into more detailed parts of the plan. This affords a way of grouping lower level into higher-order tasks that can be reasoned over. This is something well known in Artifical intelligence (AI) (Sacerdoti 1977; Tate 1975).

We expect that implementing the above mentioned features will give PIM-tools a better representation of spatio-temporal tasks and plans. This will help us with prior planning but also support us along the way of complementing and executing a plan, i.e., support our prospective memory. To achieve this goal, several open questions need to be tackled. Research in GIScience must increasingly address the issue of formal models of processes (plans) that can account for multiple levels of granularities (spatial and temporal). One of the crucial findings of the study was that **information transfer,** as described in Sect. 5.2, is a main issue that can be supported by computational tools. For the future we would like to take a closer look at the amount of time people spend for planning and examine the participants' notes. Building on the findings of this and future studies, we aim to implement a prototype PIM-tool.

Acknowledgments We would like to thank Dan Montello from the University of California, Santa Barbara for his thought-provoking comments and helpful advice during his visit at our institute.

References

Abdalla A, Frank A (2012) Combining trip and task planning: on how to get from A to passport. 7th international conference, GIScience 2012, Columbus, Ohio

Arkes HR, Ayton P (1999) The sunk cost and concorde effects: are humans less rational than lower animals? Psychol Bull 125(5):591

Barreau D, Nardi BA (1995) Finding and reminding: file organization from the desktop. ACM SIGCHI Bulletin 27(3):39–43

Eysneck MW, Keane MT (2010) Cognitive psychology, 6th edn, Psychology Press
Gilhooly KJ (1996) Thinking: directed, undirected and creative (3rd edn), Academic Press, London
Hayes-Roth B, Hayes-Roth F (1979) A cognitive model of planning. Cognitive Sci 3(4):275–310
Janowicz K (2010) The role of space and time for knowledge organization on the semantic web. Semant Web 1(1):25–32
Jones W, Teevan J (2007) Personal information management PIM, 14, University of Washington Press, Washington
Kahneman D, Tversky A (1979) Prospect theory: an analysis of decision under risk. Econometrica 47(2):263–292
Kahneman D, Tversky A (1984) Choices values and frames. Am Psychol 39:341–350
Knoblich G, Ohlsson S, Haider H, Rhenius D (1999) Constraint relaxation and chunk decomposition in insight problem solving. J Exp Psychol: Learn, Memory, Cognit, 25(6):1534–1555
Mayer RE (1990) Problem solving. In: Eysneck MW (ed) The Blackwell dictionary of cognitive psychology. Blackwell, Oxford
Montello DR, Sutton PC (2006) An introduction to scientific research methods in geography. Sage Publications, California
Norman DA (1986) Cognitive engineering. In: Norman DA, Draper S (eds) user centred system design. Lawrence Erlbaum Association, Hillsdale, pp 31–61
Norman DA (1991) Cognitive artifacts, In: John M, Carroll (eds) Designing interaction. Cambridge University Press, Cambridge
Ohlsson S (1992) Information-processing explanations of insight and related phenomena, Advances in the psychology of thinking. Harvester Wheatsheaf, New York
Raubal M, Winter S (2010) A spatio-temporal model towards Ad-Hoc collaborative decision-making. In: Geospatial thinking. lecture notes in geoinformation and cartography 0:279–297
Raubal M, Miller H, Bridwell S (2004) User-centred time geography for location-based services. Geografiska Annaler: Series B, Human Geo 86(4):245–265
Sacerdoti ED (1977) A structure for plans and behavior, Elsevier, North-Holland
Sellen A, Whittaker S (2010) Beyond total capture: a constructive critique of life logging. Commun ACM 53(5):70–77
Tate A (1975) Using goal structure to direct search in a problem solver. PhD Thesis, University of Edinburgh, Edinburgh
Timpf S, Volta GS, Pollock DW, Egenhofer MJ (1992) A conceptual model of wayfinding using multiple levels of abstraction. In: Frank AU, Campari I, Formentini U (eds) Theories and methods of spatio-temporal reasoning in geographic space. Lecture notes in computer science 639:348–367
Tversky A (1972) Elimination by aspects: A theory of choice. Psychol Rev, 79(4):281–299
Weiser P, Frank AU, Abdalla A (2012) Process composition and process reasoning over multiple levels of detail. 7th international conference, GIScience 2012, Columbus, Ohio
Wright G (1984) Behavioral decision theory: An introduction. Sage Publications, California

Publish/Subscribe System Based on Event Calculus to Support Real-Time Multi-Agent Evacuation Simulation

Mohamed Bakillah, Alexander Zipf and Steve H. L. Liang

Abstract Large scale disasters often create the need for evacuating affected regions to save lives. Disaster management authorities need evacuation simulation tools to assess the efficiency of various evacuation scenarios and the impact of a variety of environmental and social factors on the evacuation process. Therefore, sound simulation models should include the relevant factors influencing the evacuation process. More specifically, to be reliable in critical situations, evacuation simulations must integrate information on time-varying phenomena that can affect the evacuation process, such as the impact of meteorological conditions, road incidents, or other relevant events. Sensor networks constitute an efficient solution for gathering data on such events and feeding the evacuation simulation. However, the coupling of sensors with multi-agent simulation tools is not straightforward. In this chapter, we present a publish/subscribe system based on Event Calculus to support real-time multi-agent evacuation simulations. The system identifies relevant events from sensor data gathered through a Sensor Event Service that implements the OGC SWE standards and the Event Pattern Markup Language (EML). Then, the publish/subscribe system acts as a middleware between the Sensor data publishers and the multi-agent evacuation simulation through a Sensor Processing Service based on Event Calculus that infers the impact of events on the characteristics of the road network. The result is an evacuation simulation that can be deployed to assess various evacuation scenarios in real-time, during the crisis response.

M. Bakillah (✉) · A. Zipf
Institute for GI-Science, Rupprecht-Karls-Universität, Heidelberg, Germany
e-mail: mohamed.bakillah@geog.uni-heidelberg.de

M. Bakillah · S. H. L. Liang
Department of Geomatics Engineering, University of Calgary, Alberta, Canada

D. Vandenbroucke et al. (eds.), *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography,
DOI: 10.1007/978-3-319-00615-4_19,

1 Introduction

The management of natural or human disasters requires urgent response and the coordination of human and material resources. Disaster management includes the actions taken in prevention of disasters, those taken in response to an occurring disaster, and the recovery effort (Bakillah et al. 2007). The efficiency of the disaster management process, especially planning the response to disaster, can be improved by using evacuation simulations (Pel et al. 2012). In particular, the multi-agent transport simulation toolkit MATSim has been used to simulate regional evacuations (Lämmel et al. 2010). In existing evacuation simulations models, less focus has been given to modeling the agents' behavior and their reaction to changes in the environment (Fu 2004). For example, disasters may trigger contingent events such as road flooding or traffic bottleneck that affect the evacuation process (Kwan and Ransberger 2010). A simulation can only be useful if it realistically represents the relevant social and environmental factors that impact the evacuation time into a single simulation model. In this chapter, we focus on detecting and dealing with events that can affect the evacuation process. Sensor devices are a promising technology for monitoring such events.

To support the development of such evacuation simulation model, in this chapter, we propose a publish/subscribe system incorporating an Event Processing Service based on Event Calculus to support real-time multi-agent evacuation simulations. With the large amount of sensor data, the usual request/response communication used in traditional DBMS model becomes inefficient because it is based on point-to-point pulling interactions between users and data providers, which are not suitable for continuous data streams. To improve the efficiency of identifying events from sensor data, the publish/subscribe communication model and a continuous query approach are used to handle continuous data streams. The publish/subscribe model utilizes an intermediary middleware to compare predefined, continuous queries relevant for the evacuation simulation with the sensor data pushed to the middleware. More specifically, the proposed system identifies relevant events from sensor data gathered through a Sensor Event Service that implement OGC SWE standards and the Event Pattern Markup Language (EML). The publish/subscribe system acts as a middleware between the sensor data publishers and the multi-agent evacuation simulation through a Sensor Processing Service based on Event Calculus that infers the impact of events on the road network. The simulation tool is based on MatSIM, a Java agent-based toolkit for transportation simulation. The result is a simulation that can be deployed to assess various evacuation scenario in real-time, during the crisis response phase.

This scenario is a preliminary work towards a more comprehensive and generic framework for coupling sensor data with agent-based simulation, as several factors remain to be addressed, including: real-time streaming of sensor data into the simulation, sensor data quality issues that affect the accuracy of the simulation, and dealing with higher levels of heterogeneity among sensor data.

The chapter is organized as follows: the next section presents a background on evacuation simulations, while Sect. 3 presents the Sensor Web Enablement standards. The publish/subscribe system based on Event Calculus in support of multi-agent evacuation simulation is presented in Sect. 4. In Sect. 5, we present an application example. Conclusions are provided in Sect. 6.

2 Evacuation Simulation

In recent years, multi-agent simulations have been improved and they now support transportation simulation dealing with several million agents evolving in large regions (Balmer et al. 2004). As a result, multi-agent systems have increasingly been used in evacuation simulations (Chen and Zhan 2004; Pan et al. 2007; Lämmel et al. 2010). Evacuation simulation models based, for example, on cellular automata (Klüpfel et al. 2003) or on flow dynamics (Jafari et al. 2003; Kuligowski 2004), are limited in their capacity to represent characteristics of individuals and the complexity of their behavior. In contrast, multi-agent simulation allows incorporating more details and behavior of agents to accurately represent the parameters that influence the time it takes for a population to evacuate a given region. For example, some approaches (Lindell 2008; Xie et al. 2010) model the travel demand [i.e., the number of people who will evacuate and when they will depart (Pel et al. 2012)] with a response curve that establishes the percentage of departures in each time interval, regardless of individuals' specific circumstances. In fact, several parameters influence an individual's decision to evacuate or not, for example, the perceived risk, socio-demographic characteristics, the distance to the disaster, whether an issuance order has been given or not, etc. (Pel et al. 2012). Numerous approaches of agent-based evacuation simulation have been proposed, among which only some can be mentioned here. Some approaches, such as the DYNASMART model (Murray-Tuite 2007), take into account the fact that people that choose to evacuate may make decisions to adapt their route during their evacuation. As another example, Zhang et al. (2009) define three different types of agents: agents who stick to the route they had initially chosen, agents who choose to modify their route every time they face congestion, and agents who change route only half the time they face congestion. Another kind of agent-based approach incorporates the impact of interaction between agents. For example, Murakami et al. (2002) and Shanahan (1999) include in their simulation model the behavior of "leader" agents. Shendarkar et al. (2006) employed the belief-desire-intention (BDI) framework to model the behavior of agents during evacuation. The fact that agents can have complex behavior also means that they can react to changes in their environment. This has motivated some researchers to incorporate a dynamic environment in their evacuation simulation. For example, Lämmel et al. (2010) have used the notion of events to model the changes that affect the transportation network during the evacuation. More specifically, the attributes of roads, such as the flow capacity, can be modified at

arbitrary points in time. As another example, (Liu et al. 2006) present an approach of evacuation caused by a flood where the depth of the water, which varies over time, affects the speed of the evacuating people. In these evacuation simulation examples, the events that affect the evacuation process are simulated. In this research, our aim is to enable a real-time evacuation simulation where such events would be captured from sensors monitoring the real environment.

3 Sensor Web Enablement

The Open Geospatial Consortium (OGC)'s Sensor Web Enablement (SWE) initiative has developed a set of standards to achieve the so-called vision of the Sensor Web. The objectives of the latter is to support the discovery and access of sensor data over the Web while making sensor data interoperable and hiding from the users and applications the heterogeneities of sensor protocols (Bröring et al. 2011). In the context of our research, SWE standard information models and interfaces can be helpful to enable uniform and automatic access to the sensor data that capture the events we want to incorporate into the evacuation simulation. Before demonstrating how this coupling can be achieved, we briefly present the relevant standards.

The SWE suite of standards includes two main information models: the Sensor Model Language (Sensor ML), and the Observations and Measurements (O&M) model. Sensor ML is the SWE standard to describe metadata of sensors (Botts et al. 2008). O&M describes a conceptual model for the representation of measurements data (Cox 2007a, b). The SWE suite of standards also includes interface service models (Bröring et al. 2011). The Sensor Observation Service (SOS) is a standard interface for accessing descriptions of sensors and their observations (Botts et al. 2008). While the SOS was initially designed to encapsulate raw data streams from sensors and sensor networks, it has been increasingly employed to give access to more complex information, including processed sensor observations such as aggregated data (De Longueville et al. 2010). The Sensor Planning Service (SPS) is a service for remotely tasking sensors as well as setting their parameters for observation (Simonis 2007). The Sensor Alert Service (SAS) is another SWE interface service model that allows users to subscribe to events of interest captured by sensors through constraints; the SAS monitors the observations produced by registered sensors and notifies users in the same way as a publish-subscribe system would do (Open Geospatial Consortium (OGC) 2008). However, SAS has some drawbacks. Notably, it is not fully aligned with other standards such as O&M, which is an obstacle to the interoperability of SAS with other SWE-compliant services (Bröring et al. 2011). The Sensor Event Service (SES), although not yet a SWE standard, is considered as the successor of SAS. SES overcomes some of the limitations of SAS. It notably uses O&M for the encoding of sensor measurements (Bröring et al. 2011). It also takes as input subscriptions in the Event Pattern Markup Language (EML), a language which enables to define filters for events of

interest based on constraints on the values of properties observed by sensors. Because of its enhanced interoperability with other SWE services, we choose SES over SAS to implement our approach.

4 Publish/Subscribe System with Event Calculus for Real-Time Evacuation Simulation

The proposed publish/subscribe event-based system was designed to be coupled with the MATSim multi-agent evacuation simulation, which is first presented in the following section.

4.1 MATSim Multi-Agent Evacuation Simulation

The simulation framework uses the Java agent-based MATSim toolkit for transportation simulation. MATSim was designed for simulating traffic of vehicles in large cities. In MATSim, every agent represents a vehicle. Vehicles are moving on a road network composed of edges and nodes. Every vehicle has an initial trajectory corresponding to the shortest path between a person's departure place and targeted destination. The shortest path is computed with Dijkstra's algorithm. The population is distributed in the city according to available data on people's occupations and interpolation from census data. The queue model is used by MATSim to model the traffic flow. In the traffic flow simulation, each edge is assigned a set of parameters. These parameters include: the way's minimum and maximum width, as well as its length; the flow capacity, which determines the maximal outflow from the way, and which depends on the edge's width and empirical parameters; the maximal speed (free speed); and the maximal number of agents (vehicles) that can be on the edge at the same time (maximal density). The values of these parameters are used to determine the traffic flow. When an edge's maximal capacity is reached, it results in congestion on previous edges. The transportation network used for the simulation is extracted from Open Street Map (OSM). OSM is a volunteered geographic information (VGI) collaborative application where individuals can produce geographic information that emanates from their own knowledge of a geographic reality, or edit information provided by other contributors. Notably, contributors can describe map features—such as roads, water bodies, and points of interest—using "tags," providing information at a level of detail that often goes beyond the level of detail that can be provided by traditional geospatial data producers (Goetz et al. 2012). OSM has been used in numerous works (Goetz and Zipf 2011). The network model corresponds to the OSM data model and is formalized with XML. The simulation is also fed with initial trajectories of agents in this transportation network (XML format).

In our approach, events detected by sensor data are streamed into the simulation and modify the edges' parameters values, which forces agents to modify their trajectory. This process is enabled by a system based on the publish/subscribe paradigm.

4.2 Publish/Subscribe Systems

Publish/subscribe paradigm is a communication model frequently used in event-based computing (Eugster et al. 2003). It implements an efficient information-driven middleware for data sharing between clients and producers in large-scale and distributed systems (Muhl et al. 2006). Through the publish/subscribe paradigm, clients interact by exchanging event messages. Events are occurrences of interest for clients. The role of publishers is to publish events, while the role of subscribers is to consume events; however, a client can act as a producer and a consumer at the same time. Publishers monitor their environment, for example with sensor networks, and publish events that might be of interest to subscribers. Meanwhile, subscribers register their interests, called subscriptions, by defining the type of event messages they would like to receive. They can specify the sources of events they are interested in, or they can advertise their interests regardless of the source of these events. The main component of the publish/subscribe system is the middleware. The middleware is responsible for conveying event messages between publishers and subscribers asynchronously and in real-time. It handles the publications from the publishers, matches them with previously registered interests of subscribers, and pushes the published events to the matched subscribers. In this work, our focus is to use the publish/subscribe communication model to deal with events that are likely to modify an evacuation process. Therefore, the publish/subscribe-based system connects various sources of events (sensors) to a multi-agent evacuation simulation. Locations of car accidents and meteorological sensor observations are some examples of events that can affect the evacuation process. Real-time evacuation simulations, in particular, demand that occurring events be proactively disseminated to the simulation component in a timely manner. However, associating multi-agent evacuation simulation with events in publish/subscribe systems would raise new challenges to the underlying communication and event processing protocols. Thus, the proposed approach attempts to provide a suitable framework for integrating events into the evacuation simulation.

4.3 Architecture of the Publish/Subscribe System for Event Processing

The goal of the system is to deliver notifications of events that match predefined road events of interest to an Event Processing Service that will adapt the road network to the captured events. The road network therefore becomes a dynamic

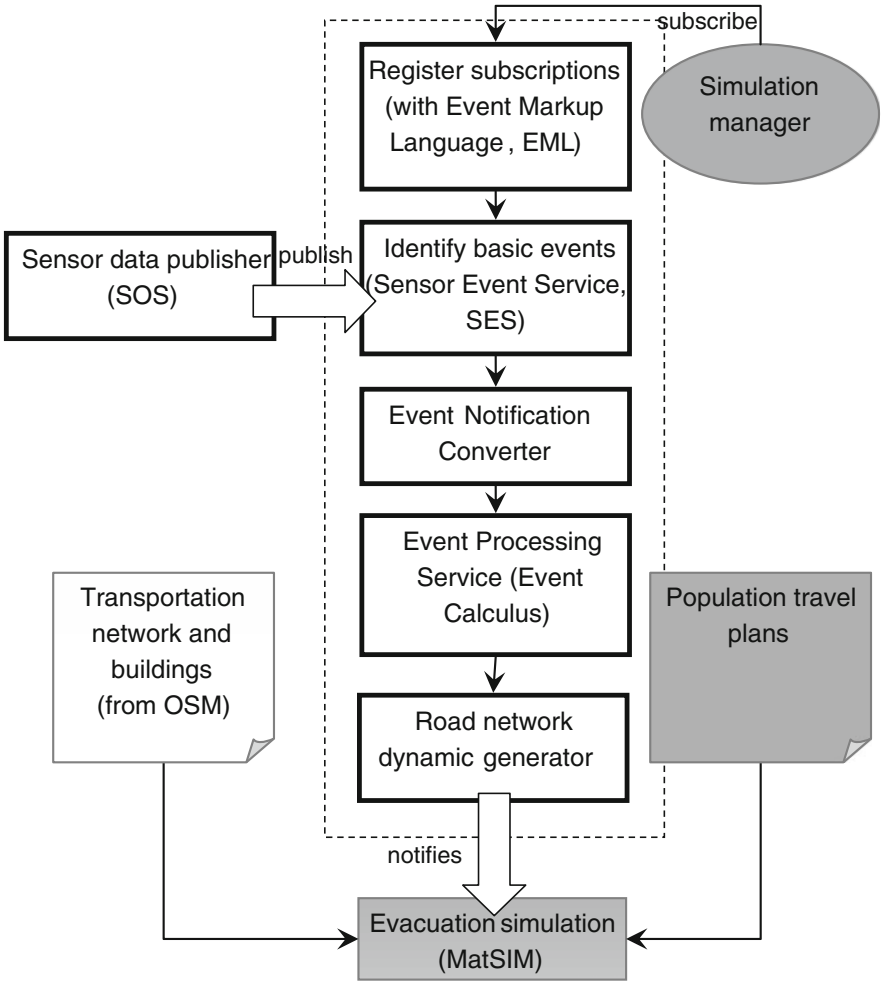


Fig. 1 Publish/subscribe system for event processing

road network, enabling to run evacuation simulations that take into account real-time or near real-time events that occurred in the reality. The system is illustrated in Fig. 1.

First, on the one hand, sensor data is captured by Sensor Observation Services (SOS) that act as "publishers". Sensor data is provided in O&M format. On the second hand, events subscriptions are provided by a simulation manager (expert) who indicates which events are likely to affect the evacuation process. The event subscriptions are formalized in EML format. A Sensor Event Service which implements the SES proposed standard filters sensor data streams to detect the events described in the subscriptions. The Sensor Event Service is built on the publish/subscribe principle, where continuous queries (instead of request/response

model) enable to deal with continuous data streams. The continuous queries are executed according to a query execution plan, composed of operators (which check whether constraints on observed properties' values are verified) and queues for incoming data that traverse the query plan. When events are detected, the Sensor Event Service notifies the Event Processing Service. The Event Processing Service is a deductive system based on Event Calculus. It deduces further events from the raw events to finally deduce the impact of the events on the parameters of the roads (maximal speed, flow capacity, etc.). Of note is that before being processed by the Event Processing Service, the event notification issued by the Sensor Event Service is previously converted into an Event Calculus statement. At last, the new road parameters are inserted into the road network file which serves as input for the evacuation simulation.

In the following, we explain the Event Processing Service which acts as a bridge between sensor data providers and the evacuation simulation.

4.3.1 Event Processing Service Based on Event Calculus

The Event Calculus is a logic programming formalism for representing events and their effects (Shanahan 1999). The Event Calculus is a deductive first-order predicate calculus that notably infers the consequences of events and actions. For example, if we know that a car accident creates an obstacle to traffic, which in turn reduces traffic flow within a given amount of time, then, given a car accident at time t1, the Event Calculus can deduce that the traffic flow will be reduced of an amount Q1 at time t2. The Event Calculus formalizes predicates for representing actions, events, as well as states and axioms constraining the relationships between these entities.

The fundamental elements of the Event Calculus are actions, event types, fluents, and time points. A fluent is defined as any property whose value may change over time or as events occurs, for example, the location or the speed of a vehicle, or the maximal flow associated with a road. Table 1 lists the predicates of Event Calculus and their interpretation.

To deduce fluent and events from input statements, the Event Calculus also provides a set of fundamental axioms, which are as follows (Shanahan 1999):

(A1) $\text{HoldsAt}(f, t) \leftarrow \text{Initially}_P(f) \wedge \neg\, \text{Clipped}(0, f, t)$

(A2) $\text{HoldsAt}(f, t2) \leftarrow \text{Happens}(e, t1) \wedge \text{Initiates}(e, f, t1) \wedge t1 < t2 \wedge \neg\, \text{Clipped}(t1, f, t2)$

(A3) $\text{Clipped}(t1, f, t2) \leftrightarrow \exists\, e, t\ [\text{Happens}(e, t) \wedge t1 < t < t2 \wedge \text{Terminates}(e, f, t)]$

(A4) $\text{HoldsAt}(f, t3) \leftarrow \text{Happens}(e, t1, t2) \wedge \text{Initiates}(e, f, t1) \wedge t2 < t3 \wedge \neg\, \text{Clipped}(t1, f, t3)$

(A5) $\text{Clipped}(t1, f, t4) \leftrightarrow \exists\, e, t2, t3\ [\text{Happens}(e, t2, t3) \wedge t1 < t3 \wedge t2 < t4 \wedge [\text{Terminates}(e, f, t2) \vee \text{Releases}(e, f, t2)]]$

(A6) $\neg\, \text{HoldsAt}(f, t) \leftarrow \text{Initially}_N(f) \wedge \neg\, \text{Declipped}(0, f, t)$

Table 1 Event calculus predicates [from (Shanahan 1999)]

Predicate	Interpretation
Initiates(e, f, t)	Fluent f starts to hold after event e at time t
Terminates(e, f, t)	Fluent f ceases to hold after event e at time t
$\text{Initially}_P(f)$	Fluent f holds from time 0
$t1 < t2$	Time point t1 is before time point t2
Happens(e, t)	Event e occurs at time t
HoldsAt(f, t)	Fluent f holds at time t
Clipped(t1, f, t2)	Fluent f is terminated between times t1 and t2
Released(e, f, t)	Fluent f is not subject to inertia after event e at time t
$\text{Initially}_N(f)$	Fluent f does not hold from time 0
Happens(e, t1, t2)	Event e starts at time t1 and ends at time t2
Declipped(t1, f, t2)	Fluent f is initiated between times t1 and t2

(A7) $\neg$ HoldsAt(f, t3) $\leftarrow$ Happens(e, t1, t2) $\wedge$ Terminates(e, f, t1) $\wedge$ t2 $<$ t3 $\wedge$ $\neg$ Declipped(t1, f, t3)

(A8) Declipped(t1, f, t4) $\leftrightarrow \exists$ e, t2, t3 [Happens(e, t2, t3) $\wedge$ t1 $<$ t3 $\wedge$ t2 $<$ t4 $\wedge$ [Initiates(e, f, t2) $\vee$ Releases(e, f, t2)]]

(A9) Happens(e, t1, t2) $\rightarrow$ t1 $\leq$ t2

A road network is represented with a directed and geo-referenced multi-graph, i.e., a graph where more than one edge (representing roads) can be associated with the same pair of nodes. The multi-graph is geo-referenced because each node is associated with a location and each edge with a localized curve. Traffic flow is associated with each edge. Traffic flow is a time-varying attribute. In addition, each edge is associated with the flow capacity, which determines the maximal outflow from the road; the maximal speed; and the maximal number of vehicles that can be on the edge at the same time (i.e., the maximal density).

We have identified the following types of events:

- Raw events as detected by sensors, such as:
  - Occurrence of a vehicle accident at time t
  - Change to a road condition (slippery, snow level, ice, flooding, low visibility) at time t
- Events modifying the traffic flow:
  - Introduction of an obstacle on an edge
  - Removal of an obstacle from an edge
  - Closure of an edge
  - Reopening of an edge
  - Reduction of the flow capacity
  - Augmentation of the flow capacity
  - Reduction of the maximal speed on an edge

  - Augmentation of the maximal speed on an edge
  - Reduction of the maximal density
  - Augmentation of the maximal density
- Events modifying the road network:
  - Introduction of an edge
  - Removal of an edge
  - Creation of a node (junction or intersection)
  - Merging of two edges
  - Introduction of a node
  - Removal of a node

The incidence of these events is related to attributes of the roads, for example, a vehicle accident causes the introduction of an obstacle on an edge, which in turn can cause the closure of the edge (if the accident creates an obstacle that spans all lanes of the road) or the reduction of flow capacity and maximal density (if the accident creates an obstacle that blocks only one of the multiple lanes of the road). Such relations are formalized with Event Calculus predicates, for example Happens(IntroductionOfObstacle, t). Happens(Accident, t). The Event-Calculus-based Service performs conjunction of domain-specific predicates with fundamental axioms to deduce new road attributes resulting from events captured by sensors.

The conversion of events expressed with EML into Event Calculus statements is performed by the Event Notification Converter. Below is an example of EML statement for an event pattern that can be detected by a visibility sensor, which measure the visibility in meters. The name of the event generated (new event name) is "LowVisibility." A filter "PropertyIsLowerThan" is used to define the condition for the event to be detected, i.e., when the Observed Property "visibility" is lower that 300.0 m, an event "LowVisibility" is generated. There is also a restriction on the sensor used to detect the event, where the sensor is identified with its ID.

```
<eml: SimplePatterns>
<eml:  SimplePattern  inputName=''InputSensorService''
patternID=''VisibilityThreshold''>
<eml: SelectFunctions>
<eml:     SelectFunction     createCausality=''false''
newEventName=''LowVisibility''>
<eml: SelectEvent eventName='' VisibilityThresholdMet''/>
</eml: SelectFunction>
</eml: SelectFunctions>
<eml: View>
<eml: LengthView>
<eml: EventCount > 1</eml: EventCount>
</eml: LengthView>
</eml: View>
```

```
<eml: Guard>
<fes:Filter xmlns:fes = ''http://www.opengis.net/fes/2.0''>
<fes:PropertyIsLowerThan>
<fes:ValueReference > input/doubleValue </
fes:ValueReference>
<fes:Literal > 300.0 </fes:Literal>
</fes:PropertyIsGreaterThan>
</fes:Filter>
<eml:/Guard>
<eml: PropertyRestrictions>
<eml: PropertyRestriction>
<eml: name > observedProperty </eml: name>
<eml: value > visibility </eml: value>
</eml: PropertyRestriction>
<eml: PropertyRestriction>
<eml: name > sensorID </eml: name>
<eml: value > visibility_sensor026 </eml: value>
</eml: PropertyRestriction>
</eml: PropertyRestrictions>
</eml: SimplePattern>
```

Whenever the event described in this pattern is detected, the Event Notification Converter selects the value in the field "NewEventName" (in the above example, "LowVisibility") and generates an Event Calculus Statement in the form "Happens(NewEventName_ID)," where a unique ID is associated with the event name to be able to identify it. The Event Notification Converter also generates a spatial statement to localize the event: HoldsAt(On(NewEventName_ID, Edge_ID)), where the Edge_ID corresponds to the edge of the road network where the sensor is located. Of note is that in the current approach, the Event Notification Converter only generates "HoldOn" spatial statement. Further research is being conducted to support the generation of additional statements corresponding to various spatial relations, such as "Beside," or "CloseTo," to allow more expressiveness. Finally, the Event Notification Converter also generates a temporal statement that corresponds to the time the event was detected: OccurrenceTime(NewEventName_ID, DateTime).

5 Application Example

To illustrate the system presented in this research, we show how events captured by sensors and disseminated by the Sensor Event Service can be modeled and processed by the Event Processing Service to automatically feed the MATSim-based multi-agent evacuation simulation with real-time data on the dynamicity of the road network. We model the impact of a car accident on the traffic flow.

First, the following is an excerpt of the input network file for the evacuation simulation. It provides the parameters of the edge identified with link id 103145, including free speed (maximal speed), capacity (flow capacity), permlanes (number of lanes), transport mode, etc.:

<link id="**103145**" from="**422**" to="**1544**" length="**315.0**" freespeed="**40.0**" capacity="**60.0**" permlanes="**2**" oneway="**1**" origid="**1577**" type="**42**" modes="**car**" time="**t0**"/>

The attribute "capacity" corresponds to the flow capacity in vehicles/min. By default, the flow capacity is equally divided between the different lanes of this edge (i.e., permlanes = 2). The excerpt also shows that these values for the edge are valid at time t0 of the simulation.

Consider that the following are raw events captured by the Sensor Event Service and converted from EML format to Event Calculus statements:

(RawEvent1) Happens(Accident1, t1)
(RawEvent2) HoldsAt(On(Accident1, Edge1))
(RawEvent3) Happens(Obstruction1, t1)
(RawEvent4) HoldsAt(On(Obstruction1, Lane1), t1)

Events 1 and 2 were captured by a "human sensor," i.e., a person who uses mobile application to communicate the occurrence of a car accident on edge1 at time t1. This event reporting does not indicate the consequence of the car accident (i.e., whether the road is closed or not). However, a traffic camera near the scene captured the obstruction of lane 1 of the same edge (events 3 and 4). The mobile application and traffic camera are two sensors that are connected, as "publishers" to the Sensor Event Service.

In addition, data on the road network can be expressed with Event Calculus statements. The following is a fact on the road network itself:

(NetworkFact1) Initially_P(HasLanes(Edge1, Lane1, Lane2))

Finally, the following are domain rules expressed with Event Calculus predicates:s:

(DomainRule1) Happens(LaneClosure, T1, T2)∧
HoldsAt(On(LaneClosure, Lane), T1) ← Happens(Accident, T1)
∧ HoldsAt(On(Accident, Lane), T1)
(DomainRule2) HoldsAt(HasFlowCapacity(Edge, FlowCapacity(Lane2))) ←

Happens(LaneClosure, T1, T2)∧
HoldsAt(On(LaneClosure, Lane1), T1)∧
Initially_P(HasFlowCapacity(Edge,
FlowCapacity(Lane1)+FlowCapacity(Lane2)))∧
Initially_P(HasLanes(Edge, Lane1, Lane2))

(DomainRule3) Released(LaneClosure, HasFlowCapacity, T2)

DomainRule1 expresses that when an accident occurs on an edge at time t1 (i.e., a road or a lane), the edge in question is closed from t1 to t2. The evacuation simulation component can therefore simulate traffic based on an estimation of the edge closure.

For the evacuation simulation purpose, the duration of the edge closure (t2–t1) can be determined from averaging empirical data. DomainRule2 expresses that if an edge is affected by lane closure, then the flow capacity of the edge will be reduced by the flow capacity of the closed lane. DomainRule3 expresses that when LaneClosure no longer exists (at time t2), then the flow capacity is back to the value it had prior to the LaneClosure event.

Let RE be the conjunction of RawEvent1 to RawEvent4, let DR be the conjunction of DomainRule1 to DomainRule3, and let A be the conjunction of axioms A1 to A9. We have:

$A \wedge RE \wedge NetworkFact1 \wedge DR \Rightarrow$

HoldsAt(HasFlowCapacity(Edge1, FlowCapacity(Lane2), t1) $\wedge$

HoldsAt(HasFlowCapacity(Edge1, FlowCapacity(Lane2) + FlowCapacity(Lane1)), t2)

As a result, the system outputs the following temporal snapshots of the road network, where flow capacity is reduced to 30 vehicles/min during the time of the lane closure:

<link id="**103145**" from="**422**" to="**1544**" length="**315.0**" freespeed="**40.0**" capacity="**30.0**" permlanes="**2**" oneway="**1**" origid="**1577**" type="**42**" modes="**car**" time="**t1**"/>

<link id="**103145**" from="**422**" to="**1544**" length="**315.0**" freespeed="**40.0**" capacity="**60.0**" permlanes="**2**" oneway="**1**" origid="**1577**" type="**42**" modes="**car**" time="**t2**"/>

The system can then forward the time-varying values of flow capacity for edge 103,145 to the evacuation simulation component.

We envision the system presented in this chapter as a first step towards a more comprehensive and generic framework for coupling sensor data with agent-based evacuation simulation to enable evacuation planning according to real-time conditions. However, several challenges must be addressed to fully achieve this goal. Firstly, while the Event Calculus is suitable for reasoning with events, in its current form, it does not support complex spatial reasoning. However, reasoning with spatial relations is required to enable the current system to deal with more complex situations. For example, it is very likely that the location of sensors does not exactly match the location of road segments or places of interests. Meanwhile, the system should be able to identify that events captured by these sensors may still be relevant (under given conditions) to deduce the characteristics of nearby road segments. Therefore, further research is necessary to achieve the coupling of the

Event Calculus with a spatial reasoning language. Secondly, multiple types of obstacles and events could be observed by various sensors and considered for the simulation: heavy rain, ice, construction, demonstrations, etc. Considering multiple events requires more research in terms of resolving semantic heterogeneities, as heterogeneous sensors would be used, as well as for fusion of sensor data. In this respect, useful work was presented for example in Bakillah et al. (2012) to support the retrieval and merging heterogeneous data that could be used to enhance our approach. Another issue raised during the experimentation was related to the temporal relationships between events. Obviously the scenario presented was simplified, but in the reality, for example, two events that are related to the same incident (i.e., the detection of the car accident and lane closure) would not be captured by sensors at the exact same time, while it was useful here to consider them as simultaneous. Therefore, we envision that the approach could be enhanced with fuzzy temporal reasoning.

6 Conclusion and Perspectives

This chapter is intended to be a preliminary work towards the coupling of sensor data with multi-agents evacuation simulations. We have introduced a publish/subscribe system based on Event Calculus to support real-time multi-agent evacuation simulations. This system demonstrates how events captured by sensor data can be dealt with in an automatic manner by combining Sensor Web Enablement standard information and service models with event processing capabilities. We have studied the specific case where sensors are employed to identify obstacles to traffic flow. However, the approach can also be extended to deal with events that affect, notably, the drivers' behavior. Future work will also aim at expanding the present research by finding appropriate techniques to deal with the fuzziness of temporal relations between events and higher levels of heterogeneity among sensor data.

References

Bakillah M, Mostafavi MA, Brodeur J (2007) Mapping between dynamic ontologies in support of geospatial data integration for disaster management. In: Proceedings of joint CIG/ISPRS conference on geomatics for disaster and risk management, Toronto, Ontario

Bakillah M, Mostafavi MA, Liang SHL (2012) Enriching SQWRL queries in support of geospatial data retrieval from multiple and complementary sources. In: Proceedings of 6th international workshop on semantic and conceptual issues in GIS (15–18 Oct 2012, Florence, Italy), SeCoGis 2012 (under press)

Balmer M, Nagel K, Raney B (2004) Large scale multi-agent simulations for transportation applications. J Intell Transport Syst 8:205–223

Botts M, Percivall G, Reed C, Davidson J (2008) OGC sensor web enablement: overview and high level architecture. In: Nittel S, Labrinidis A Stefanidis A (eds) Proceedings of geosensor networks, GSN 2006, LNCS 4540. Springer, Berlin, pp 175–190

Bröring A, Echterhoff J, Jirka S, Simonis I, Everding T, Stasch C, Liang S, Lemmens R (2011) New generation sensor web enablement. Sensors 11:2652–2699

Chen X, Zhan FB (2004) Agent-based modeling and simulation of urban evacuation: relative effectiveness of simultaneous and staged evacuation strategies. In: Proceedings of the 83rd annual meeting transportation research board, Washington, DC, USA

Cox S (2007a) OGC implementation specification 07–022r1: observations and measurements—part 1—observation schema. Open Geospatial Consortium, Wayland

Cox S (2007b) OGC Implementation Specification 07–022r3: Observations and Measurements—Part 2—Sampling Features. Open Geospatial Consortium, Wayland

De Longueville B, Annoni A, Schade S, Ostlaender N, Withmore C (2010) Digital Earth's nervous system for crisis events: real-time sensor web enablement of volunteered geographic information. Int J Digit Earth 3(3):242–259

Eugster P, Felber PA, Guerraoui R, Kermarrec AM (2003) The many faces of publish/subscribe. ACM Comput Surv 35(2):114–131

Fu H (2004) Development of dynamic travel demand models for hurricane evacuation. Dissertation, Louisiana State University, USA

Goetz M, Zipf A (2011) Extending open street map to indoor environments: bringing volunteered geographic information to the next level. In: Rumor M, Zlatanova S, LeDoux H (eds) Proceedings of the 2011 urban and regional data management. UDMS 2011, Delft, The Netherlands, pp 47–58

Goetz M, Lauer J, Auer M (2012) An algorithm based methodology for the creation of a regularly updated global online map derived from volunteered geographic information. In: Rückemann C-P, Resch B (eds) Proceedings of the 4th international conference on advanced geographic information systems, applications, and services (Valencia, Spain), pp 50–58

Jafari M, Bakhadyrov I, Maher A (2003) Technological advances in evacuation planning and emergency management: current state of the art. In: Proceedings of final research reports EVAC-RU4474, center for advanced infrastructure and transportation (CAIT), Rutgers University, NJ

Klüpfel H, Meyer-König T, Keßel A, Schreckenberg M (2003) Simulating evacuation processes and comparison to empirical results. In: Fukui M et al (eds) Traffic and granular flow'01. Springer, Berlin, pp 449–454

Kuligowski E (2004) Review of 28 egress models. Technical report National Institute of Standards and Technology (NIST), Gaithersburg

Kwan M-P, Ransberger DM (2010) LiDAR assisted emergency response: detection of transport network obstructions caused by major disasters. Comput Environ Urban Syst 34(3):179–188

Lämmel G, Grether D, Nagel K (2010) The representation and implementation of time-dependent inundation in large-scale microscopic evacuation simulations. Transp Res Part C 18:84–98

Lindell MK (2008) EMBLEM2: an empirically based large scale evacuation time estimate model. Transp. Res. A 42:14–154

Liu Y, Hatayama M, Okada N (2006) Development of an adaptive evacuation route algorithm under flood disaster, vol 49. In: Proceedings of annuals of disaster prevention research institute, Kyoto University, pp 189–195

Muhl G, Fiege L, Pietzuch P (2006) Distributed event-based systems. Springer, Germany

Murakami Y, Minami K, Kawasoe T, Ishida T (2002) Multi-agent simulation for crisis management. In: Proceedings of the IEEE on knowledge media networking workshop, pp 135–139

Murray-Tuite P (2007) Perspectives for network management in response to unplanned disruptions. J Urban Plan Dev 133(1):9–17

Open Geospatial Consortium (OGC) (2008) OGC sensor event service interface specification—version 0.3.0

Pan X, Han C, Dauber K, Law K (2007) A multi-agent based framework for the simulation of human and social behavior during emergency evacuations. AI Soc 22(2):113–132

Pel AJ, Bliemer MCJ, Hoogendoorn SP (2012) A review on travel behaviour modelling in dynamic traffic simulation models for evacuations. Transportation 39:97–123

Shanahan M (1999) The event calculus explained. In Woolridge MJ, Veloso M (eds) Artificial intelligence today, LNAI 1600. Springer, Berlin, pp 409–430

Shendarkar A, Vasudevan K, Lee S, Son Y (2006) Crowd simulation for emergency response using BDI agent based on virtual reality. In: Proceedings of the 2006 winter simulation conference

Simonis I (2007) OGC implementation specification 07–014r3: open GIS sensor planning service. Open Geospatial Consortium, Wayland

Xie C, Lin DY, Waller ST (2010) A dynamic evacuation network optimization problem with lane reversal and crossing elimination strategies. Transp Res E 46:295–316

Zhang B, Kin W, Ukkusuri SV (2009) Agent-based modeling for household level hurricane evacuation. In: Rossetti MD, Hill RR, Johansson B, Dunkin A Ingalls RG (eds) Proceedings of the 2009 winter simulation conference, Institute of Electrical and Electronics Engineers, Inc., Piscataway, pp 2778–2784

A Visual Analytics Approach for Assessing Pedestrian Friendliness of Urban Environments

Tobias Schreck, Itzhak Omer, Peter Bak and Yoav Lerman

Abstract The availability of efficient transportation facilities is vital to the function and development of modern cities. Promoting walking is crucial for supporting livable communities and cities. Assessing the quality of pedestrian facilities and constructing appropriate pedestrian walking facilities are important tasks in public city planning. Additionally, walking facilities in a community affect commercial activities including private investment decisions such as those of retailers. However, analyzing what we call pedestrian friendliness in an urban environment involves multiple data perspectives, such as street networks, land use, and other multivariate observation measurements, and consequently poses significant challenges. In this study, we investigate the effect of urban environment properties on pedestrian movement in different locations in the metropolitan region of Tel Aviv. The first urban area we investigated was the inner city of the Tel Aviv metropolitan region, one of the central regions in Tel Aviv, a city that serves many non-local residents. For simplicity, we refer to this area as Tel Aviv. We also investigated Bat Yam, a small city, whose residents use many of the services of Tel Aviv. We apply an improved tool for visual analysis of the correlation between multiple independent and one dependent variable in geographical context. We use the tool to investigate the effect of functional and topological properties on the volume of pedestrian movement. The results of our study indicate that these two

T. Schreck (✉)
University of Konstanz, Konstanz, Germany
e-mail: tobias.schreck@uni-konstanz.de

I. Omer · Y. Lerman
Tel Aviv University, Tel Aviv, Israel
e-mail: omery@post.tau.ac.il

Y. Lerman
e-mail: yoavlerm@post.tau.ac.il

P. Bak
IBM Research Lab, Haifa, Israel
e-mail: peter.bak@il.ibm.com

D. Vandenbroucke et al. (eds.), *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography,
DOI: 10.1007/978-3-319-00615-4_20,

urban areas differ greatly. The urban area of Tel Aviv has much more correspondence and interdependency among the functional and topological properties of the urban environment that might influence pedestrian movement. We also found that the pedestrian movements as well as the related urban environment properties in this region are distributed geographically in a more equal and organized form.

1 Introduction

Past attempts to explain pedestrian or walking movement patterns in urban environments mainly focused on two groups of factors that affect this movement: street connectivity and functional factors. Previous studies have applied multivariate and bivariate regression models, mostly by the Stepwise method, to obtain the highest coefficient of determination, by using the R2 measure for example [such as (Hillier et al. 1993) or (Ozbil et al. 2011)]. However, we know little about how the spatial structure of street networks and street connectivity, i.e. topological centrality, interacts with other important urban environment functional factors such as commercial land use and sidewalks in different urban environments, and what are the consequences of that interaction concerning pedestrian movement. Motivated by these questions, we investigated the correlation of urban environment properties on pedestrian movement, as measured by the number of pedestrians per time in different urban areas in the metropolitan region of Tel Aviv. Specifically, we investigated the inner city of the metropolitan Tel Aviv, a city that serves many non-local residents. We also investigated Bat Yam, a smaller city that serves mostly local residents. Our key research question asks for the spatial-functional configurations that enhance walkability (or pedestrian friendliness) in these urban environments. While many qualitative factors contribute to the walkability or pedestrian friendliness, in this work we assume the number of pedestrians is an indicator therefore, recognizing this is a simplification. We want to explore how these configurations create pedestrian friendly environments. The results of this study can potentially guide urban development policy in assigning priority to some of the identified properties.

The analytic problem in our case is a *correlation* problem, involving a set of four independent variables that describe the spatial and functional properties and one dependent variable—namely, the average number of pedestrians. Consequently, we need to also consider the geospatial map and the overall street network in the analysis. To address this challenging problem, we rely on approaches from visual data analysis to make sense of the acquired data. Specifically, we applied a method for visual cluster analysis that groups the sets of observations into a smaller number of categories, describing similar configurations of the independent variables. Using this, along with a linked map display, we then analyzed the dependency of the variables in terms of attributes and geospatial position in the

map context. We thereby provided both an analytical workflow to a general problem but also practical insights for the two specific studied cities, enabling us to give answers to questions of critical importance for urban planning.

The remainder of this chapter is structured as follows. In Sect. 2, we describe related work. In Sect. 3, we introduce the geospatial analysis problem at hand and describe our data acquisition process. In Sect. 4, we describe the setup of the visual analysis tool we implemented. In Sect. 5, we explain how we applied our tool on the collected data, discussing the main insights found and their implications. Finally, in Sect. 6, we summarize our approach along with interesting future work in the area.

2 Related Work

In this section, we review related work that analyzed data on urban environments with a focus on pedestrian movement patterns. We also briefly introduce related work on the visual analysis of data in geospatial and multivariate domains. Visual and analytical comparison between spatial distributions of objects and attributes within a GIS framework can be an essential tool for understanding and explaining geographic phenomena in urban areas.

Much evidence has been collected indicating that the geographic distribution of pedestrian movement along city streets is affected by two main characteristics of the urban environment—topological centralities of streets (or street segments) and the spatial distribution of retail and service facilities [e.g., (Golledge and Stimson 1997; Hillier et al. 1993; Jiang 2007; Ozbil et al. 2011)]. However, we still have no sufficient knowledge on these relationships, i.e., why, when, and where certain urban environment characteristics are more influential than others for predicting pedestrian movement in the city. This complexity may be related to the fact that the topological properties of individual streets are significantly correlated to the spatial distribution of retail and services [e.g., (Desyllas et al. 2003; Porta et al. 2006)]. Many studies were conducted to address this issue by combining various attributes in empiric investigations of urban pedestrian movement [e.g., (Orellana and Wachowicz 2011; Ozer and Kubat 2007; Raford and Ragland 2006; Torrens 2012)]. In practice, however, and as mentioned above, most investigations of the relationship among pedestrian movement, street network connectivity, and land use distribution have been conducted in statistical terms only, with no explicit and detailed consideration of geographic and multidimensional aspects.

Visual and analytical comparisons between spatial distributions of objects and attributes within a GIS framework can be an essential tool for understanding and explaining geographic phenomena in urban areas. They can do so, by offering an integrated view among multiple dimensions, including independent and dependent quantitative variables, in a geographic context. Accordingly, our approach to analyze pedestrian friendliness data is based on visual representations that compare and correlate the data from these perspectives. The recently evolving field of visual

analytics addresses the design, application and evaluation of tools that combine automatic data analysis methods with visual-interactive representations (Keim et al. 2011; Thomas and Cook 2005). Recently, visual-interactive approaches have also been extensively applied in the geographic data analysis community (Andrienko and Andrienko 2006). For understanding multivariate data, projection or clustering methods are often applied to group data by similarity and thereby ease their interpretation. Techniques for such grouping include principal components analysis as a prominent dimensionality reduction technique (Jolliffe 2002), and the self-organizing maps algorithm for projection and cluster analysis (Bak et al. 2010; Guo et al. 2005; Spielman and Thill 2008; Kohonen 2001).

3 Data Acquisition for Pedestrian Friendliness Analysis

Trends show that cities are becoming more populated, and the analysis and improvement of city infrastructure is an important goal. We acquired empirical data for a study to identify the influential factors that can determine the pedestrian friendliness of a city street network. To this end, we measure a set of independent variables that may explain the attractiveness of street properties to pedestrians, useful for city analysis and policy planning. In this section, we describe the setup of the data acquisition, which is the basis for the subsequent analysis.

3.1 Tel Aviv and Bat Yam: Background

We present research that deals with pedestrian movement in two urban areas—the cities of Tel Aviv and Bat Yam, focusing on the effect of the street network structure and land uses on the intensity of pedestrian movement. Both areas include orthogonal street patterns (similar to a grid). Yet, they also show hierarchical patterns as well as many internal loops, cul-de-sacs, and T junctions that characterize more modern planning approaches. The selected study area in Tel Aviv is divided into two sections that were designed and built at different times and have different characteristics. The western section was built during the 1930s according to a master plan made by the Scottish urban planner, Sir Patrick Geddes. The eastern region was built during the 1950s as part of the "East Tel Aviv" plan. These two areas differ in residential density, street grid, and land use mix. The city of Bat Yam has 130,000 residents. The study area in Bat Yam included the entire city, which has parts that were built based on plans made in the 1930s in the north-west area of the city, and parts that were built later in the 1950s and 1970s, in the south and to the east. The newer parts of the city have street networks that are less connected than the older parts.

3.2 Data Acquisition

Measurement points in the study areas in Tel-Aviv and Bat-Yam were selected to represent a range of different centrality measures and distribution of land uses. In both cities data was collected using the gate count method where each pedestrian who passed through the gate was counted. This method has been used by other studies (Desyllas and Duxbury 2000; Zhang et al. 2012) providing high-resolution count results efficiently for such as those that were conducted in this research. The selected research area in Tel Aviv covered 400 acres. Concerning its central location in the metropolitan region, we assume that non-local residents make up a significant portion of pedestrian movement. The count was done for 5 min every hour for 5 h at each survey point. The measurement took place on a sunny weekday between the hours 3 and 8 p.m. The survey took place in 51 different street segments at 95 measurement points. Bat-Yam is a suburb of Tel Aviv, therefore we assume that the pedestrians movements in this city are mainly due to local residents. The research area in Bat-Yam includes the entire city area, which spreads over 1,800 acres. Pedestrian volume sampling was done in 69 street segments (122 measurement points) throughout Bat Yam. At each survey point the count was done for 5 min every hour for 8 h. The Bat Yam pedestrian survey was done on a weekday during the hours 7 a.m. until Noon and from 3 to 8 p.m. In both cities, we collected data to understand the correlations between the built environments and the pedestrian movement. Specifically, we collected data that would enable investigating the correlations of the street network properties and the land use properties with regard to pedestrian movements, in different parts of each area and on different geographical scales. Although the sampling took place during slightly different hours in both cities, the overall average distribution of pedestrian movement during the measured period shows similar patterns in both areas.

In the two areas we investigated, the data for the functional independent variables were partially collected using a field survey and partially by using geographic information layers. Data on the land-use distribution was obtained as GISlayers from the Survey of Israel (MAPI) and the Mapa company.[1] The data include geo-referenced residential and public buildings and a detailed description of their land uses. The street connectivity independent variables—namely, connectivity and local integration—were measured at the level of individual street segments by using the space syntax methodology (Hillier 1996). When using this methodology, the built environments' spatial configuration is described by means of a topological analysis of its axial map. An axial map is defined as the smallest set of the longest lines of direct visibility and movement that pass through all of a citys' open spaces. For any particular axial line, connectivity denotes the number of directly linked axial lines. The integration measure indicates the closeness of an axial line to other axial lines by computing the shortest distance (or step depth) of

[1] The Survey of Israel, the Israeli official government agency for Mapping, Geodesy, Cadastre and Geoinformatics. Mapa is a private company. All data is updated to 2011.

the respective line from other axial lines in a given area. The local integration measure describes integration only up to a defined radius of topological distance (number of steps), which is restricted to three steps in the current research. To create the axial maps of the neighborhoods and the entire city as well as to compute the street connectivity independent variables, we used two software programs: Depthmap and AxialGen. Depthmap [(Turner 2004), version 8.15, UCL] was used to automatically create (and manually edit) axial maps based on the urban street network. AxialGen [(Jiang and Liu 2010), version 1.0] was used to calculate, analyze, and present the connectivity independent variables (space syntax attributes) within the ArcMap (ver. 9.3) GIS software. The data for the dependent variable were collected using a survey in selected street segments that represent geographical locations and street kinds in both of the studied areas.

3.3 Attribute Formation and Data Summary

In both Tel Aviv and Bat Yam, the street segments for the surveys were selected so that a range of topological and functional values would be represented in the sample. We compiled 122 measurement points for Bat Yam and 95 measurement points for Tel Aviv. As discussed above, at each measurement point the pedestrian movement was sampled during a few hours using the gate count method. The following list summarizes the four independent variables and the one dependent variable we used. This data was input to the visual analysis approach described in the next section.

- **CommFront**: Binary variable, indicating whether a retail commercial front is present at the measurement point (independent).
- **BusStation**: Binary variable, indicating whether a bus stop is present at the measurement point (independent).
- **Connectivity**: Axial connectivity value of the street network at the measurement point (independent).
- **LocalInt**: Axial local integration value of the street network at the measurement point (independent).
- **AveragePerHour**: Number of pedestrians counted at the measurement point, averaged per hour (dependent).

4 Visual Analysis Design

In this section, we describe the system design by its components. The design is derived from the main problems of analyzing the data at hand and is inspired by the workflow presented in (Bak et al. 2010):

1. We form groups of locations that are similar regarding their topological and functional configurations.
2. We visually compare the configuration groups against one another in terms of the distribution of topological and functional attribute values.
3. Then, we correlate the configuration groups against the target variable (the average number of pedestrians).
4. We also analyze the spatial distribution of the configurations and pedestrian counts over the map, giving the geographic context of the observations.

We address the first step in the workflow (1) by applying the self organizing map algorithm (SOM) (Kohonen 2001) to the four-dimensional configuration data samples. The method is a neural-network type data reduction and projection algorithm, which is often used in visual cluster analysis. We use it to obtain from the larger number of configurations, a small number of representative (cluster) configurations, on which the subsequent analysis is based. Training a small SOM of size 3×2, we obtained four main distinct clusters (configurations) contained in the four corner fields of the 3×2 map. Note that the size of the SOM grid influences the number of prototypes one obtains. The SOM algorithm does not by itself yield an indication of the number of clusters one can assume in the data, but this needs to be done by inspection of the analysis. In our case, we tried also larger resolutions but found empirically, that a 3×2 grid yields four distinct configurations which are suitable to our task at hand. We also note that the two central fields in the SOM result remained unpopulated interpolation fields—an indication that the found prototypes are rather discriminative against each other.

The Visual Analytics approach in our design is reflected by enabling users to iteratively define the input parameters of the SOM algorithm (size of the grid, and consequently the number of obtained prototypes). This interaction allows the user to decide when the resulting prototypes are significant enough; that is, when the number of contained data instances is sufficiently large for each luster, and the prototypes are discriminative enough. In addition, Visual Analytics as a core component in our approach is expressed through the interactive linkage between the visualization elements, resulting in a feedback loop between results of the algorithm and the users' search for insight. In this context, visualization is used as the interface to algorithmic refinement and an interactive display of algorithmic results.

We visualize each configuration group by a radial parallel coordinate (or radar) chart glyph, on which the four dimensions span a planar coordinate system. The mapping, clock-wise from the top is: CommFront, BusStation, Connectivity, LocalInt. Figure 1 (left and middle) shows an illustration. A black polyline represents the average values of the cluster (the cluster prototype) in terms of these independent variables. Yellow semitransparent bands indicate the distribution of actual measurements represented by the cluster. We also color-code the average number of pedestrians observed at the respective configurations as the background color. There, light shades represent low pedestrian counts and dark shades represent high counts [colors used according to (Brewer 2012)].

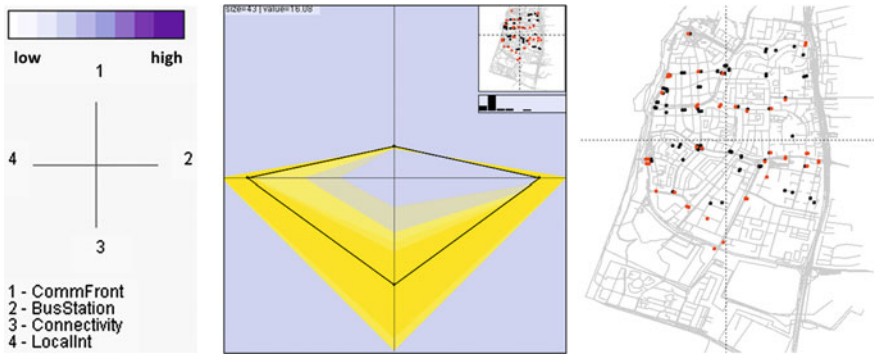


Fig. 1 The components of our visual analysis design. A four-dimensional glyph represents clusters of similar configurations of the input variables (*left*). The glyph represents the average number of pedestrians observed for a group of similar street configurations by color-coding the back-ground and by a histogram (*darker indicates larger numbers*). A *dot map* display shows the occurrence of measurements represented by this cluster on the map (*right*, highlighted by *bright/red dots*)

We extend the glyph to include a small dot map inset which shows the geographic position of the represented observations. Also, we show the distribution of the pedestrian counts represented in the groups by a histogram [see Fig. 1 (middle, top-right)]. Our approach also allows for zooming in the dot map for a more detailed inspection of the spatial distribution of configurations.

The visualization decisions for the glyph construction are taken consciously and in close cooperation with domain experts. The four-dimensional glyph was chosen to represent the input variables using a radar-chart. Literature suggests (Inselberg 2002) the effectiveness and expressiveness of this representation to detect patterns in datasets with multidimensional attributes. The choice of using the radar-chart version of this representation type was to save real-estate on the display. As our users were exclusively with geographic background, their request for corresponding spatial representation was accommodated by the map view. The radar-chart view and the map-view were used in a highly integrated manner, in order to generate hypotheses and reflect on the distinctiveness of the SOM clusters. We are convinced that the existing design is a good starting point. Future work may include a systematic usability test which could be the basis for further improvements on the analytical workflow.

5 Pedestrian Friendliness Analysis for Bat Yam and Tel Aviv

Our main question is how do the spatial structures of street networks and street connectivity interact with other important functional factors such as commercial land use in their effects on pedestrian frequency in different urban environments in

the same metropolitan region. That is, how do different spatial-functional configurations affect urban pedestrian movement in different geographical conditions? Based on visual data analysis, we address aspects of this question in the following.

5.1 Bat Yam Analysis

We first consider the Bat Yam case. We start our analysis by clustering the observation data according to the functional and topological measurements (see Sect. 4). The SOM analysis of the Bat Yam data reveals four distinct clusters of spatial-functional configurations (see Fig. 2, top row). In the radar chart, the four directions, up/east/south/west represent the independent variables CommFront, BusStation, Connectivity, and LocalInt (see also legend in Fig. 1 left). We can see at a glance that: (1) in general, the values of functional variables are strongly related to the values of the topological variables—the four-dimensional glyphs tend to expand rather equally in all directions, which means an interdependency among these variables is present. This is a side results of our analysis. And (2) a positive correlation exists between the values of the spatial-functional independent variables and the dependent data, as the frequency of pedestrian movements increases, represented by the background color in the image. The background color gets darker (more pedestrians) as the four-dimensional glyph shape becomes larger. This illustrates well the effect of the built environment's properties on pedestrian movement in the city.

Moreover, a detailed examination of the four configurations reveals that high spatial-functional values provide sufficient conditions for high volumes of pedestrians. The large cluster that represents walkable configurations in Fig. 2 (top row, left-most cluster) illustrates that well—showing higher values in all the independent values with small (standard) deviation. However, as the second cluster from the left in the row illustrates, high values of independent variables are not necessary conditions of pedestrian friendliness; lower values in one or in several variables still coincide with high volume of pedestrian movement. This is exemplified by the relatively lower values and higher deviations in the Connectivity and the BusStation variables, as apparent from the larger spread bands shown in yellow in the glyph. This means that high values of variables, mainly of local integration and commerce, are necessary for significant pedestrian movement. The other two configurations of variables that characterized low volume of pedestrian movement illustrate that the variable of commerce is low (with a small standard deviation) for both.

In light of the results described above, we found it necessary to examine the geographical patterns of the configurations that are identified in the SOM process and their locations in the urban environment. Such examinations are possible through maps that display the location of the members of each cluster. Figure 2 (middle row) displays the configuration locations represented by the respective clusters. We found that the street segments that are represented by the two most

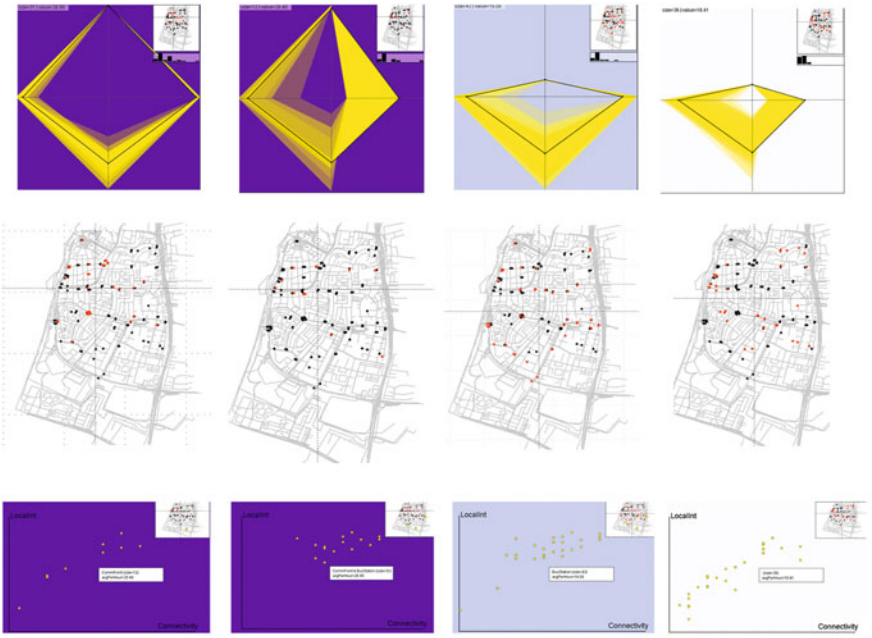


Fig. 2 Visual analysis of the Bat Yam Data. The *top row* shows the four different configurations of the functional-topological variables as found by cluster analysis for Bat Yam. The distribution of the measurements represented by the clusters on the map of Bat Yam in shown in the *middle row*. The *bottom row* shows scatter plots of the independent variables Connectivity (x-axis) versus LocalInt (y-axis) for the occurrences of CommFronts (*first chart*), CommFronts + BusStation (*second chart*), only BusStation (*third chart*), and neither presence of CommFronts or BusStations (*fourth chart*)

walkable configurations (with high values of pedestrians all the independent variables; left two clusters in Fig. 2, top row) are located mainly in the northeast part of the city. The other two configurations are distributed equally over the city areas. This unequal geographic distribution may be related to the central role of the northeast part of the city in its functioning and to a high level of connectivity and integration of its street network. We should notice in this respect that significance correlation was not found between the volume of pedestrian movement, and residential density or population size in different city areas. Therefore, we assume that part of the northeast area pedestrians come from other city areas due to its attractiveness. However, the focus of the investigation is on the distribution of pedestrians within a given area according to the selected functional and spatial variables.

To explore the exact relations among the values of selected variables in our data set, we also extended the visualization tool with scatter plot diagrams (see Fig. 2, bottom row). To illustrate the potential contribution of our tool, we chose to examine the relation between the two topological centrality variables—LocalIntegration and Connectivity, for the four possible combinations of

CommFront and BusStations being present or absent. Our question was: What is the relationship between the values of these variables in each of the identified configurations? Specifically, we are interested in how the topological centrality values at street segment level are associated with the distributions of functional land uses (in our case, CommFront and BusStations) and with pedestrian movement. The accompanied scatter plots help to explore the interrelation among the three components—functional properties, spatial properties (topological centrality), and pedestrian movement—in different geographical areas. In the case we present, we can see that the topological centrality values tend to correlate positively. Within this general tendency, the more walkable configurations (high volume of pedestrian movement, darker background color in the scatter plot charts) tend to be with higher values of local integration. On the contrary, the less walkable configurations (third and fourth chart in the row) tend to represent street segments with low local integration and connectivity values. These findings indicate that the topological centrality variables create the basic conditions for both, distributions of functional land uses and pedestrian movement.

Figure 3 enhances the geographic exploration in more detail. We used our tool to produce a diagram that presents the volume of pedestrian movement (via circle size) in each point/segment, together with an indication to which of the clusters they belong (via color mapping). This enables obtaining essential information at different geographic scales on how the principal functional-spatial configurations are distributed in the geographic space, how they relate geographically to one another, and how each of them is related to the geographical distribution of pedestrian movements. It can assist us, for example, to locate areas that consist of segments of different functional-spatial configurations and walkability levels. Namely, it enables simultaneously to identify geographical areas with differential walkability levels and to clarify some of the reasons for that differentiation through investigation of the obtained configurations of independent variables. E.g., as shown in Fig. 3, some places comprise points that have similar environmental configurations, but differ in their walkability levels. In such situation, one of the main tasks in pedestrian planning policy is to find barriers and spatial partitions that prevent pedestrian flow in a given urban environments [e.g., (Hillier 2002; Orellana and Wachowicz 2011; Torrens 2012; Zook et al. 2012)]. Thus, such empiric knowledge leads to better designs of pedestrian paths that widen the areas of pedestrian movement; for example, as part of renewal of commercial areas [e.g., see Zampieri et al. (2009)].

5.2 Tel Aviv Analysis

The case of Tel Aviv is similar to that of Bat Yam with regard to the general interdependency among the spatial and functional independent variables and their relationships, as configuration types, with the volumes of pedestrian movements. The four-dimensional glyphs (see Fig. 4, top row) tend to expand rather equally in

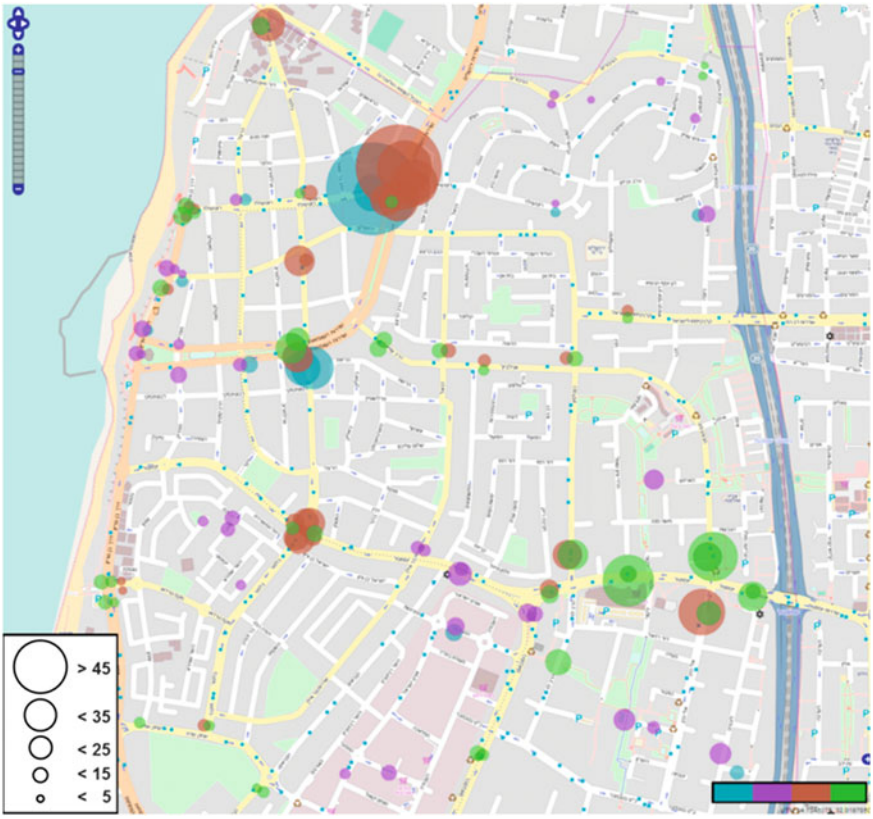


Fig. 3 The distribution of configuration types in selected street segments by volume of pedestrian movement for the city of Bat Yam. The colors (see legend in *bottom-right* part of image) denote the different configuration types from Fig. 2. Colors 1 and 3 correspond to configurations 1 and 2, and colors 2 and 4 to configurations 3 and 4

all directions, and along with that also varies the background color. This similarity between Tel Aviv and Bat Yam means interdependency among the urban environments' properties and positive correlation with the pedestrian movements. However, several significant differences exist between these two cases. First, while in Bat Yam, high values of independent variables are sufficient but not necessary conditions for creating significantly walkable locations, in Tel Aviv, high values of independent variables are sufficient as well as necessary for achieving such high walkability levels (see Fig. 4, the first cluster from left (note: Please zoom in for better readability). E.g., significant presence of pedestrians in Tel Aviv exists only where bus stops are located, and at the same time, bus stops are located in places where a significant presence of pedestrians exists. This indicates that there is stronger reciprocity between the urban environments' attributes and pedestrian movement. Second, a stronger interdependency and better fit also exists among the

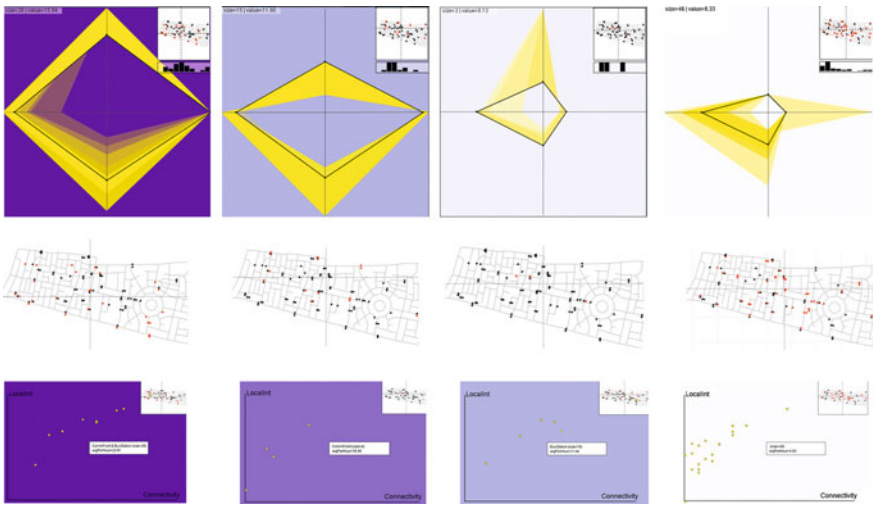


Fig. 4 Visual results for the Tel Aviv data set

independent variables themselves, especially between the two spatial variables (see the scatter plots of the independent variables in Fig. 4, bottom row). Thus, the results also indicate that more interdependency and even a structuration among all the variables involved in the phenomenon of urban pedestrian movement exists in Tel Aviv, which is more pronounced than in Bat Yam.

Third, the case of Tel Aviv differs greatly from the case of Bat Yam concerning the geographic distributions of the spatial-functional configurations. As Fig. 5 shows, in Tel Aviv the different configurations of variables are distributed equally and create a sort of hierarchical structure. We can see that the most walkable configuration type (denoted by the brown color), which tends to be with the higher volume (big circles), is distributed over all the area and creates a sort of skeleton or backbone, mainly in the western section, which was built earlier and is characterized by a street grid. At the same time, the other configuration types, with a relatively low volume of pedestrians, tend also to be located between and around the most walkable configurations, i.e., the locations with higher pedestrian movements. On the contrary, the different types of configurations in Bat Yam tend to concentrate in a few geographical areas with no clear spatial relation among them. We can also see that the locations with higher volume of pedestrian movement are located in two or three geographic areas that function only as "main spots" of pedestrian movements (see Fig. 5). That is, unlike Tel Aviv, Bat Yam is characterized by a geographical clustering of configuration types, and geographical inequality in the distribution of pedestrian movement. This can be seen as an expression of different spatial autocorrelations of the independent environmental variables. In Tel Aviv, the geographical scale of the land use mix is significantly more local, i.e., different land uses such as commerce and residence can be found in a relatively small geographical area. Accordingly, Tel Aviv also has much more

Fig. 5 Configuration types by volume of pedestrian movement for the city of Tel Aviv. The colors (see legend in *bottom-right* part of image) denote the different configuration types from Fig. 4. Colors 1 and 3 correspond to configurations 1 and 2, and colors 2 and 4 to configurations 3 and 4

equality in the geographic distribution of the volume of pedestrians at the level of individual locations and at the level of geographical areas. And again, much more correspondence exists between the configuration types and the volume of pedestrian movement.

6 Discussion and Conclusion

In this chapter, we explained and compared pedestrian movement patterns in two urban areas in different parts of the Tel Aviv metropolitan region. The first area, inner city Tel Aviv, referred to as Tel Aviv in this chapter, is the functional center of the metropolitan region that serves many non-local residents from the entire region. The second area is Bat Yam, a smaller city in the Tel Aviv metropolitan region, which serves mostly local residents.

Using adapted methods from visual data analysis, we found that these two urban areas differ greatly in several respects. Tel Aviv has much more correspondence and interdependency among the urban environments' variables. This is seen between the spatial and functional variables involved in the phenomenon of urban pedestrian movement. In addition, we found that this correspondence also has a geographical expression—pedestrian movements as well as configurations of independent environmental variables are distributed equally over the environment area and create a sort of hierarchical structure. On the contrary, Bat Yam has relatively less correspondence among the variables involved in the phenomenon of pedestrian movement and no significant spatial organization of pedestrian movements. These differences can be related mainly to the functional statuses of Bat Yam and Tel Aviv. While Bat Yam serves mostly local residents, Tel Aviv, as one

of the main central areas of the metropolitan region, serves a relatively high rate of non-local residents, who are highly dependent on the functional activity in the area. In our case, this is commerce and bus stations. The findings we obtained by using visual analysis methods can assist policy making with the aim of improving the conditions for higher levels of walkability in urban environments throughout the metropolitan region.

Our approach is based on a visual, multidimensional correlation analysis. As such, it presents patterns of co-located measurements, and concentrates mainly on geometric properties. The results are determined by the selection of variables and the interactive analysis process. The study could be extended by considering additional descriptive factors, such as affordances of places like presence of tourist attractions, work place density, aesthetic criteria of the given architecture, and many more. Such factors could help to substantiate our initial findings. We already extend beyond the basic space syntax paradigm by including the functional variables CommFront and BusStation, but certainly, and extension would be interesting.

We identified a number of interesting items for future research. First, the geovisual analytics methods we developed for pedestrian movement analysis in different urban environments should be adapted for use by urban planners during the design of walkable environments, or when implementing walkability-oriented spatial policy. Second, the study could be extended to other urban parts in the metropolitan region to explore how pedestrian movements are created in different functional and morphological environments. Third, while we used the number of pedestrians as the empirical expression of pedestrian friendliness, qualitative analysis, e.g., by interviews, could further specify the exact notion of pedestrian friendliness. In addition, with the wide-spread use of smartphone technology, including wireless Internet and GPS modalities, we envision much more detailed and high-resolution pedestrian data becoming available in the future. Thereby, more complex patterns of pedestrian movements could be taken into account. Then, further advanced approaches for geovisual analytics need to be developed to adequately represent such extended measurements in user-friendly interfaces for use by planners and designers of walkable urban environments.

Acknowledgments We thank Sebastian Bremm and Tatiana von Landesberger of TU Darmstadt for fruitful discussions on the topic as part of a working group meeting of the DFG SPP 1335 on Scalable Visual Analytics.

References

Andrienko N, Andrienko G (2006) Exploratory analysis of spatial and temporal data—a systematic approach. Springer, Berlin

Bak P, Omer I, Schreck T (2010) Visual analytics of urban environments using high-resolution geographic data. Geospatial Thinking 25–42

Brewer C (2012) Color brewer 2.0

Desyllas J, Duxbury E (2000) Planning for movement—measuring and modelling pedestrian flows in cities. In: Proceedings of the royal institute of chartered surveyors conference
Desyllas J, Duxbury E, Ward J, Hudson-Smith A (2003) Pedestrian demand modelling of large cities: an applied example from London. CASA working papers
Golledge R, Stimson R (1997) Spatial behavior: a geographic perspective. Guilford Press, New York
Guo D, Gahegan M, MacEachren A, Zhou B (2005) Multivariate analysis and geovisualization with an integrated geographic knowledge discovery approach. Cartography Geogr Inf Sci 32(2):113
Hillier B (1996) Space is the machine. Cambridge University Press, Cambridge
Hillier B (2002) A theory of the city as object—or, how spatial laws mediate the social construction of urban space. Urban Des Int 7(3):153–179
Hillier B, Penn A, Hansonand J, Grajewski T, Xu J (1993) Natural movement: or, configuration and attraction in urban pedestrian movement. Environ Plann B: Plann Des 20(1):29–66
Inselberg A (2002) A survey of parallel coordinates. Math Vis 167–179
Jiang B (2007) A topological pattern of urban street networks: Universality and peculiarity. Physica A 384(2):647–655
Jiang B, Liu X (2010) Automatic generation of the axial lines of urban environments to capture what we perceive. Int J Geogr Inf Sci 24(4):545–558
Jolliffe I (2002) Principal component analysis. Springer, New York
Keim D, Kohlhammer J, Ellis G, Mansmann F (2011) Mastering the information age: Solving problems with visual analytics. Eurographics Assoc
Kohonen T (2001) Self-organizing maps. Springer, Berlin
Orellana D, Wachowicz M (2011) Exploring patterns of movement suspension in pedestrian mobility. Geogr Anal 43(3):241–260
Ozbil A, Peponis J, Stone B (2011) Understanding the link between street connectivity, land use and pedestrian flows. Urban Des Int 125–141
Ozer O, Kubat A (2007) Walking initiatives: a quantitative movement analysis. 6th international space syntax symposium
Porta P, Crucitti P, Latora V (2006) The network analysis of urban streets: a primal approach. Environ Plann B: Plann Des 5(33):705–725
Raford N, Ragland D (2006) Pedestrian volume modeling for traffic safety and exposure analysis: case of Boston, Massachusetts. Transportation research board 85th annual meeting
Spielman S, Thill J (2008) Social area analysis, data mining, and gis. Comput Environ Urban Syst 32(2):110–122
Thomas J, Cook K (2005) Illuminating the path: the research and development agenda for visual analytics. National visualization and analytics Ctr
Torrens Paul M (2012) Moving agent pedestrians through space and time. Ann Assoc Am Geogr 102:35–66
Turner A (2004) Depthmap 4—a researcher's handbook. School of Graduate Studies, UCL, Bartlett
Zampieri F, Rigatti D, Ugalde C (2009) Evaluated model of pedestrian movement based on space syntax, performance measures and artificial neural nets. In: Proceedings of the 7th international space syntax symposium, pp 1–8
Zhang L, Zuhang Y, Dai X (2012) A configuration study of pedestrian flows in multi-level commercial space. In: Proceedings of the 8th international space syntax symposium
Zook J, Lu Y, Glanz K, Zimring C (2012) Design and pedestrianism in a smart growth development. Environ Behav 44(2):216–234

Modelling the Suitability of Urban Networks for Pedestrians: An Affordance-Based Framework

David Jonietz, Wolfgang Schuster and Sabine Timpf

Abstract In this chapter, a framework for modelling the suitability of urban networks for pedestrians is presented. Based on the psychological theory of affordances, a model of spatial suitability is developed that acknowledges the fact that suitability must always be analysed relative to the agent, the task and the environment. We extend existing affordance concepts by moving beyond simple true/false statements to express that there are various degrees to which an action can be afforded by an environmental object. In our model, environmental dispositions and agent capabilities are repeatedly selected, calculated and specified until atomic property pairs are identified. These can be combined to compute suitability values. We test and implement the model on a routing scenario for mobility-impaired persons. The results show that the framework produces suitable paths for different agents and thus shows promise for future work.

1 Motivation

The need to automatically evaluate the suitability of urban networks for pedestrians is a problem that sustainable transportation planning and the development of user-adaptive routing systems have in common. Making urban networks more walkable for pedestrians and taking into account their individual differences might be the key to many traffic problems in European city centres and enable the high mobility of an urban lifestyle for many different groups of people.

Currently, both in transportation planning and in adaptive routing, the suitability of a specific route is rated by experts using a list of environmental attributes that are related to walking, such as the trip distance and the slope of streets or their

D. Jonietz (✉) · W. Schuster · S. Timpf
Institute for Geography, University of Augsburg,
Germany Universitaetsstr. 10, D-86159 Augsburg, Germany
e-mail: david.jonietz@geo.uni-augsburg.de

D. Vandenbroucke et al. (eds.), *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography,
DOI: 10.1007/978-3-319-00615-4_21,

crossings, factors which are usually recorded in a geographic information system (GIS) (see e.g. Clark and Davies 2009; Czogalla 2011). The rating process is highly subjective and cannot take into account the specific requirements of the individual.

In this study, we place our focus on the above mentioned rating process, meaning the translation of selected environmental attributes into a scaled suitability value for individual mobility. For this, experts tend to base their estimation on results from empirical studies, rules of thumb or their personal opinion. Suitability, however, is a complex concept, which implies a strong focus on the human perspective. Thus, the suitability of an environmental object for a particular action should always be understood in relation to the respective actor. Whereas a path segment of a pedestrian network can be almost perfectly suitable for a regular walker, for instance, it can be inaccessible for a mobility-impaired person. A translation process as mentioned above, therefore, must incorporate individual capabilities as well as environmental attributes.

The insight of a mutual dependence of environment and actor with regards to action potentials represents a core concept of ecological psychology, a movement in perceptual psychology. Its foundations were set by Gibson (1977) with his theory of visual spatial perception, describing how the perception of possibilities for action, which he termed affordances, are determined by the combination of properties of both the environment and the perceiving organism. In the past, the notion of affordances has provided the basis for numerous studies from other disciplines, including geographical information science. There have also been suggestions to use the concept in the context of pedestrians and walkability (Alfonso 2005; Jonietz and Timpf 2012).

In Geographic Information Science, the question of how to create semantic meaning in geospatial data is still an issue of discussion. With our framework, we propose the use of affordances as a potential method to interpret measurable attributes of the environment with regards to individual users with specific tasks, in our case in order to evaluate the suitability. Present approaches to formalize affordances, however, are restricted to assuming binary true/false values, meaning that an affordance is either provided or not, with no intermediate stages. This notion is not compatible with the concept of suitability. Therefore, we extend existing work by introducing the notion of affordances being expressed as scaled values, i.e. degrees to which an action can be afforded by an environmental object in the presence of an individual with a specific task to carry out.

This chapter is structured as follows. After providing background information on affordances as a valuable concept, we present our model on how to derive a suitability value from individual capabilities and environmental dispositions. We then show the practical value of our model by applying it to a realistic routing scenario for mobility-impaired persons. The last section contains our conclusions and shows future work.

2 Background

In this section, relevant background information will be presented, starting with a short presentation of affordances as a key concept of Gibson's perceptual theory. This will be followed by a non-exhaustive review of related work, which builds on this theoretical basis.

2.1 The Affordance Concept

The notion of affordances was originally developed by the psychologist Gibson (1977) as part of his ecological theory of visual spatial perception. In the course of several of his works, in which he investigated perceptual processes involved in the creation of spatial meaning by human observers, he gradually came to reject the prevalent psychological theories of his time (Jones 2003). Whereas it was generally assumed that complex mental processing must be necessary to interpret otherwise meaningless perceptual data, he proposed a more straightforward theory of perception, arguing that meaning, in the specific form of action potentials, is in fact perceived directly. In reference to the verb *to afford*, he coined the new term affordances:

> The affordances of the environment are what it offers the animal, what it provides or furnishes, whether for good or ill. (Gibson 1979, p. 127).

Thus, to give an example, a stone may offer throw-ability to a human being. Its potential for being thrown, therefore, can be seen as a result of its properties such as its weight or size. Accordingly, in reference to Gestalt psychologist Koffka (1935), who Gibson claimed to have influenced his work on affordances, "*each thing says what it is*" (Koffka 1935, p. 7).

The functionality of an environmental object, however, is not just determined by its attributes, but rather in relation to the acting animal (Gibson 1979). Returning to the previous example, for instance, it is clear that a human being will perceive and evaluate a stone's throw-ability based on his or her own capabilities, such as grasp size or strength. The described principle of agent-environment mutuality represents one of the key innovations of Gibson's work and a core concept of ecological psychology (Varela and Thompson 1991).

Until today, the notion of affordances remains a very vague issue, which is due to the fact that until his death, Gibson himself continuously modified his theory and repeatedly described it as "subject to revision" (Gibson 1977, p. 67). Nevertheless, his works still enjoy an outstanding popularity and provide the theoretical basis for numerous studies from a variety of disciplines such as spatial cognition, artificial intelligence, robotics or GIS. In the following section, a selection of this work will be discussed.

2.2 Related Work on Affordances

Whereas the previous section focused on the background of affordances as a core concept in perceptual psychology, here, a review of selected contributions from a variety of disciplines will be provided. For this, the scope will be restricted to work relevant to answer the following question:

How can the affordance concept be integrated into a GIS-based model of spatial suitability?

For this specific problem to be solved, it is necessary to further investigate what exactly is meant by the term affordance. As a reaction to the conceptual vagueness mentioned in the previous section, there have been several attempts to further refine and formalize this concept. Turvey (1992) was the first to provide a formal description of affordances. Assuming a system Wpq composed of a person Z and an environmental object X, the author defines an affordance as a property p of the environment, which, however, is determined in its existence by a complementing property termed effectivity q of an involved agent. Apart from Turvey (1992), there are several other authors who adopt similar views (e.g. Heft 2001; Michaels 2000; Reed 1996; Stoffregen 2000).

Nonetheless, in later studies, there has been profound criticism. Stoffregen (2003), for example, claims that allocating affordances solely to the environment naturally implies that further mental processing by the observing animal must be required, a fact that would contradict some of Gibson's most basic assumptions. He, in contrast, defines affordances as "properties of the animal-environment system […] that do not inhere in either the environment or the animal" (Stoffregen 2003, p. 123). With reference to the example offered by Turvey (1992), he presents a revised formalization of affordances:

> Let Wpq (e.g., a person-climbing-stairs system) = (Xp, Zq) be composed of different things Z (e.g., person) and X (e.g., stairs). Let p be a property of X and q be a property of Z.
>
> The relation between p and q, p/q, defines a higher order property (i.e., a property of the animal–environment system), h.
>
> Then h is said to be an affordance of Wpq if and only if
>
> (i) Wpq = (Xp, Zq) possesses h
> (ii) Neither Z nor X possesses h

In direct reaction to Stoffregen (2003), Chemero (2003) argues that affordances should be defined as relations between an animal's abilities and environmental features rather than properties of the animal-environment system. Despite these differences, however, there is much agreement between these two authors. In the context of this study, therefore, an affordance can be roughly described as a potential for a specific action. The potential is in its existence determined by certain properties of the animal or the environment, which together form a system.

In his study on the climb-ability of stair steps, Warren (1984) describes the components that are involved in creating an affordance in greater detail and, at the same time, demonstrates its value as a basis for practical analyses. In his work, he further explores the relevancy of the three aspects environment, actor and the task.

In his experiment, he could show that concerning the specific task to climb a stair step, test persons tend to perceive the affordance climb-ability only as long as the ratio between the step's height and the person's leg length does not exceed a threshold value of 0.88 (Warren 1984).

Jordan et al. (1998) build on Gibson (1977) in their work on creating an affordance-based model of place in GIS. In reference to Warren (1984), they identify the three aspects of affordances that must be modelled: the agent, the environment and the task. To determine, for example, the suitability of a restaurant for a potential customer, the authors claim that it is necessary to note not only the agent's capabilities and preferences, but also the actual task, such as eating, socializing or reading. Only when analysed in combination can these data serve to realistically represent a place's meaning (Jordan et al. 1998). Apart from modelling place, affordances have in the past been applied to human wayfinding and navigation (Raubal 2001), agent-based modelling (Raubal and Moratz 2006), as primitives for semantic reference systems (Scheider et al. 2009), similarity assessment (Janowicz and Raubal 2007) or semantic transformations of human observations (Ortmann et al. 2012).

3 An Affordance-Based Model of Spatial Suitability

We propose an affordance-based model of spatial suitability. In reference to Stoffregen (2003), we understand an affordance h as a higher order property of the agent-environment system Wpq, which is in its existence determined by environment- and agent-related properties p and q that are interconnected in complex dependency relationships p/q. Apart from the agent Z and the environment X, the task, meaning the respective action, is also a critical aspect, since an affordance is always task-related (Jordan et al. 1998).

The three mentioned factors agent, environment and task, together with their respective properties, are of decisive importance for the emergence of an affordance. Accordingly, in order to determine whether a specific affordance exists in an actual agent-environment system, all three aspects must be modelled with their properties and set in relation to each other. Accordingly, to refer back to the example used by Warren (1984), the existence of the affordance climb-ability (h) of a stair step can be determined by identifying the relevant properties step height (p) and leg length (q), creating a ratio (p/q) and determining whether it does or does not exceed the critical value of 0.88.

In contrast to previous work, however, we postulate that affordances are not restricted to mere binary true/false statements. Instead, in close relation to the concept of suitability, we hypothesize that there are various degrees to which an action can be afforded by an agent-environment system. In most realistic scenarios, we argue, there will be several objects, which afford the same action to a specific agent, but still, one or the other may be preferred because of a higher suitability for the task at hand. For instance, one could imagine two paths, one paved and one

muddy, which both afford walking to a human agent. Still, the former would usually be preferred since it would be perceived as more suitable to the action walking. Similarly, one could expect that a person-climbing-stairs system with a ratio close to the critical threshold value of 0.88 would be perceived as less suitable compared to a system with a lower value. Consequently, we argue that by comparing received ratio values to known critical threshold values, it is possible to derive scaled values for affordances, which in turn convey information about the task-specific suitability of an environmental object with reference to an individual agent.

When dealing with realistic scenarios, for example for routing applications, it must be acknowledged that actions are complex constructs that are in most cases not related to just one affordance, but rather have to be modelled as a hierarchically structured system of several sub-affordances. For the action *movement* to be afforded successfully by a path segment in a network, for example, an agent must be able to physically walk on the surface, perform wayfinding, surmount potential barriers such as ramps or stairs and so on. These sub-affordances, in turn, can again be broken down into lower level affordances, just as wayfinding includes being able to visually perceive signs or landmarks, understand their semantic meaning, create a cognitive map of the area et cetera. In our opinion, in order to calculate an affordance with regards to a specific agent-environment-action system, it is necessary to break down the affordance into its constituents until arriving at the most elementary level, where every property of the environment is confronted by just one property of the agent, similar to the formalization proposed by Stoffregen (2003). These atomic dependency relationships can then be further analysed, for example in the form of ratio values, as done by Warren (1984).

For our model of spatial suitability, which is illustrated in Fig. 1, we introduce a discretionary system (Wpq) composed of an agent_i (Z) with task_i and an environmental object_j (X). Other than Jordan et al. (1998), who placed the task at the same hierarchical level with the agent and the environment, we allocate it as an attributive quality to the agent. Instead of scaled affordances, we speak of the $\text{suitability}_\text{ij}$ that represents the final output of the model and can be expressed as a scaled value.

As a first step, in a process that is determined by the properties of $agent_i$ and $task_i$, relevant properties of the environmental object can be selected or calculated from the total of attributes. These are stored in what we call an environmental disposition $disp_{ij}$, which can be understood as a situation-specific dynamic model of the environmental object, therefore including only properties that actually matter with respect to $agent_i$ and $task_i$. For instance, whereas the existence of kerbstones might be neglected when calculating a route for a walking person, it may be of critical importance for a person sitting in a wheelchair (agent). Similarly, seating accommodation on a path segment can be neglected when simply searching for the fastest route (task).

The environmental disposition $disp_{ij}$, in turn, has an influence on the creation of the capabilities $capi_j$ of $agent_i$, which represent a collection of the situation-specific properties of $agent_i$ that are selected or especially calculated in reference

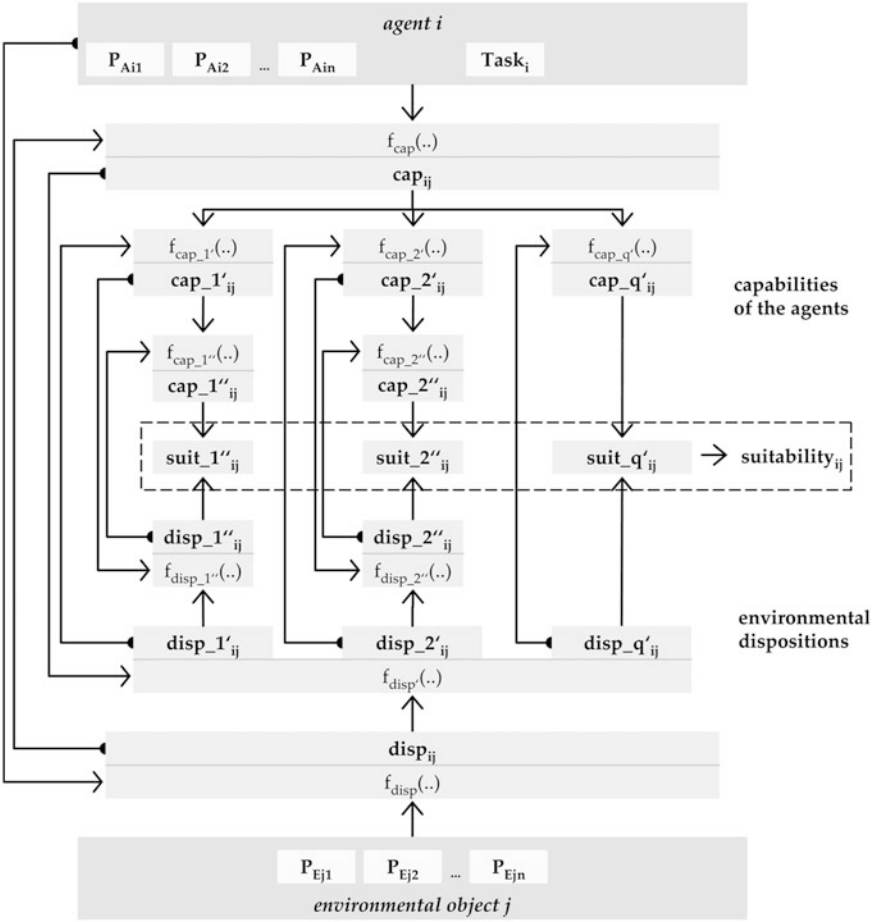


Fig. 1 Affordance-based model of spatial suitability

to $disp_{ij}$. The capability to climb stairs, for example, depends on other agent properties, such as leg length, than the capability to surmount a ramp.

After identifying the set of relevant properties of both agent and environment, and storing them in $disp_{ij}$ and cap_{ij}, similar selection processes must take place in order to gradually refine the sub-affordances. If, for example, $disp_{ij}$ of a path segment includes a barrier such as stairs, then movement is only afforded when stair climbing is also afforded. The stair climbing capability of $agent_i$, however, requires various physical efforts such as lifting the leg to the step height, pushing one's own weight up and so on, which can each be interpreted as lower level sub-affordances with related $cap\_n'_{ij}$ and $disp\_n'_{ij}$. As stated above, a repeated specification of capabilities and related environmental dispositions eventually leads to the identification of atomic property pairs on the most elementary level. The elementary level is reached when no further refinement is possible. In most cases,

this will be due to the restricted level of detail of the used data. The relation between the atomic property pairs can then be standardized and analysed to receive suitability values $suit\_n'_{ij}$, which can finally be combined to a single scaled value for the overall $suitability_{ij}$.

4 Testing the Model Using a Case Study

In order to assess the practical value of our model, we chose a routing scenario for mobility-impaired persons as an exemplary implementation. The suitability values that we receive are assigned as cost factors to the edges of our network for the routing algorithm. Our main focus, however, lies on the modelling of agents, tasks and the environment for the purpose of calculating these values and not on the routing algorithm itself.

4.1 Environment

In a first step, taking part of the campus of the University of Augsburg as a study area, data was gathered on stairs and slopes in the path network, with a focus on the height and the number of steps for stairs and the gradient and the length for slopes. For both types of barriers, the presence of handrails was also a criterion. Despite the fact that there are more potential barriers such as inappropriate surfaces or gutters, we concentrated our analysis on stairs and slopes as the main factors that may restrict the accessibility of footpaths. We deliberately limited the scope to these barriers and properties in order to maintain a simple and comprehensible testing scenario. For our routing scenario, the barriers are assigned to a non-directional graph that represents the footpath network of the campus.

Table 1 shows the properties of two path-segments of the routing-network as an example of environmental objects. One slope barrier with a length of 68 m, a gradient of 5.7 % and no handrail is assigned to path 1. Path 2 is modelled with two barriers, one slope and one set of stairs, which both need to be analysed. The lengths of the paths (respectively less the lengths of the related barriers) are included in the computation when calculating the least exhaustive path (see Sect. 4.3).

4.2 Agents

The agents are defined through their auxiliary device (e.g. a wheelchair), their personal fitness level and their potential to surmount individual steps and gradients with or without a handrail. We assume simplified potentials here, without going into detail about the actual physical properties that constitute them. Modelling the

Table 1 Properties of path segments

Path	Length	Barriers			
1	78 m	Slope ->	Length: 68 m	Gradient: 5.7 %	No handrail
2	44 m	Slope ->	Length: 18 m	Gradient: 5.0 %	Handrail
		Stairs ->	Step-height: 12 cm	Step-number: 26	Handrail

detailed relations of physical attributes of persons and structural characteristics of barriers is a very complex task and beyond the scope of our study at the present stage. This simplification allows us to model the accessibility of barriers in a manner, which is appropriate for the purpose of our work but without the need to consult empirical studies at this point.

Five different agents are defined for the case study (see Table 2):

1. pedestrian in good physical condition (i.e. with a fitness level of 8),
2. pedestrian in bad physical condition,
3. pedestrian in average physical condition with a problem in mounting stairs,
4. a wheelchair driver in good physical condition, and
5. a wheelchair driver in bad physical condition.

4.3 Task

The agents face two tasks in the case study: (1) moving along the most accessible path and (2) moving along the least exhausting path. For the first task, only the structural attributes of the barriers matter, whereas their length is not taken into account. That means that the focus is solely on the question whether an agent is able to surmount the barriers. For the second task, the fitness level of the agent is important. We expect both the structure and the length of the barriers to have an effect for determining the least exhausting path.

Table 2 Properties of agents defined for the case study

Agent	Auxiliary device	Fitness level (1–10)	Potential to surmount			
			Step (cm)		Slope-gradient (%)	
			With handrail	Without handrail	With handrail	Without handrail
1	None/ pedestrian	8	50.0	40.0	50.0	30.0
2	None/ pedestrian	2	20.0	10.0	18.0	9.0
3	None/ pedestrian	6	12.5	5.0	10.0	5.5
4	Wheelchair	10	5.0	5.0	10.0	10.0
5	Wheelchair	2	3.0	3.0	5.5	5.5

4.4 Interrelations Between the Components

The interrelations between agents, tasks and the environment are modelled in accordance with the framework as described above. Starting from an agent, a specific task and an edge of the path network, the resulting environmental disposition for this particular edge contains its barriers (if there are any). If there are no barriers on the path segment, only its length will be included in the environmental disposition and used for further calculation. The agent's capabilities are determined in relation to the environmental disposition, the task and the agent's properties. Again, the edge's disposition is further broken down to selected properties of the barriers that are related to certain capabilities. In our example, the functions determining the dispositions and the capabilities simply select certain properties. It would also be possible to calculate new capability values from existing properties. If there is more than just one barrier on an network edge, they all contribute to its suitability and must be analysed simultaneously.

If, for example, a set of stairs is analysed, its disposition determines what capabilities are needed to surmount it in relation to its special structure. There are certain factors (e.g. if the agent is a wheelchair-driver) that ultimately decide whether the stairs are insurmountable or not. By repeating the process of selecting and specifying the dispositions and capabilities, all factors can be identified. Finally, we arrive at single values for environmental dispositions and capabilities, which each can provide a suitability value for one particular sub-affordance. Dealing with task 2 (least exhausting path) and a stair barrier, for example, as a first step, the quotient between the actual step-height and the agent's potential to surmount an individual step is computed. If this quotient is smaller than 1, the action is afforded. Dealing with task 2 (least exhausting route), in order to receive the suitability value of this barrier, this quotient is multiplied with the length of the barrier (e.g. the number of the stairs) and put into proportion with the fitness level of the agent.

Finally, the partial suitability values can be added up to receive one value for the network edge with reference to a particular agent and task. If, however, one sub-affordance is not given, the edge itself is labelled as insurmountable.

Table 3 shows the suitability values for two path segments (see Table 1) for the five agents listed in Table 2. For both wheelchair-drivers (agents 4 and 5) path 2 is not suitable (insurmountable), because of the fact that it contains stairs. The three pedestrian agents are principally able to surmount the two barriers on path 2, but especially for agent 2, the cost to do so is relatively high (0.922 and 0.677). This is mostly because of a low potential to surmount a step and a low fitness level in relation to the many steps (26) to surmount. Although the potential to surmount steps of agent 3 (12.5 cm) is much lower than the potential of agent 2 (20 cm), for the second task the distinctly higher fitness level of agent 3 keeps the cost lower than the costs of agent 2. Path 1 (one slope) can be surmounted by two pedestrians and one wheelchair driver. For agent 3 and agent 5 the slope is too steep in comparison to their potential to surmount slopes. Again agent 2 has the highest

Table 3 Resulting suitability values for the two path segments

Task		Agent				
		1	2	3	4	5
Most accessible route	Path 1	0.268	0.711	–	0.648	–
	Path 2	0.384	0.922	1.504	–	–
Least exhausting route	Path 1	0.028	0.265	–	0.049	–
	Path 2	0.087	0.677	0.336	–	–

cost in surmounting the barrier because of a low potential to surmount the slope and a low level of fitness.

4.5 Results

The calculated values were used as cost factors in a Dijkstra-algorithm to compute the optimal, or most suitable, routes for the given tasks and agents. This additional computation serves as proof of concept for the proposed model. Figure 2 shows the results for five different agents and the two tasks mentioned above and shown in Table 3.

The left map in Fig. 2 refers to locomotion on the most accessible paths. According to the task, the length of the barriers is not taken into account at all while the length of the paths has only a minor influence on the calculation. For two agents (pedestrians in good/bad physical condition), the most accessible route includes a set of stairs with a handrail. Although the third pedestrian (problems in mounting stairs) has a higher fitness level than the pedestrian with a bad physical condition, it is not possible for him or her to surmount these steps. Consequently, in this special case the most accessible route is a relatively long one and proceeds over a relatively gentle slope, which is surmountable for this agent. As the staircase is also insurmountable for both wheelchair agents, this route is also the most accessible one for them.

The right map in Fig. 2 shows the least exhausting routes for the agents, as the lengths of the barriers are taken into account together with their accessibility. In addition, the lengths of the paths are used in relation to the agents' fitness levels. In comparison to the former task, there are no differences in the resulting routes for both agents with a bad physical condition, because alternative routes including less exhausting barriers are longer (and thus more exhausting) or shorter routes lead over more exhausting or insurmountable barriers (e.g. the slope with a gradient of 5.7 % for the wheelchair driver in bad physical condition). For the wheelchair driver in a good physical condition this barrier is surmountable and the respective route is less exhausting than the long route suggested in the previous task. Similarly, for the pedestrian with a poor stair climbing capability, the least exhausting route is not the longer route of the first task, but leads over relatively flat stairs (12 cm); the nearby slope with a gradient of 5.7 % is insurmountable for this agent

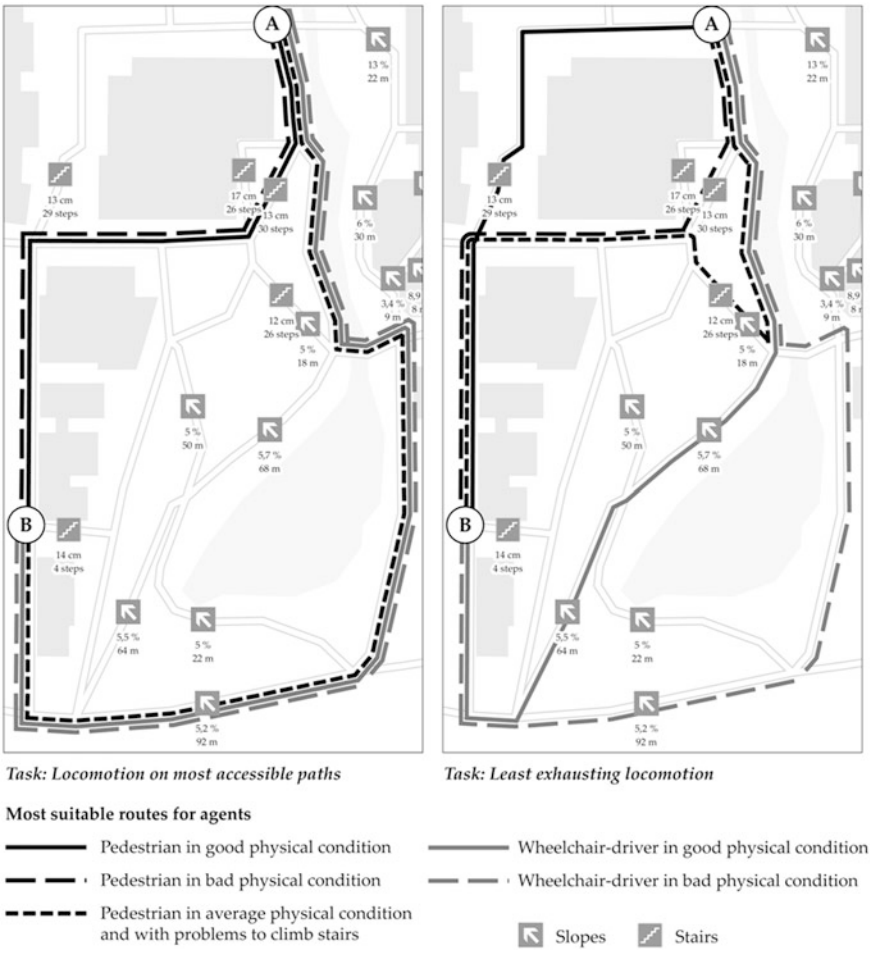


Fig. 2 Most suitable routes of five agents for two different tasks

(see Table 2). For the pedestrian in good physical condition the calculated difference between the route alternatives as suggested for both tasks is very small which is why both routes have an almost equal suitability regarding this agent and the second task.

Despite the quite restrictive assumptions in our case study, the calculated routes represent the actual environmental setting concerning route choices of different kinds of pedestrians in a realistic way. To practically improve the model, more types of barriers and more detailed structured barriers should be taken into account. Furthermore a directional graph (and also directional slopes) would improve the validity of the generic model.

By increasing the complexity of our data one main problem of our theoretical model emerges. Expressing and modelling the interrelations between agents, tasks and environment becomes very complex and partly fragmented when raising the complexity of the data. It is possible that applying ontologies to the model could help to reduce this emerging obscurity.

5 Summary and Outlook

Motivated by the insight that suitability can only be assessed in relation to a specific task and actor, we presented a framework to model the suitability of urban networks for several types of pedestrians based on the notion of affordances (Gibson 1977). We extended the affordance concept by moving beyond simple true/false statements to allow affordances to be expressed through scaled values. In our model, suitability values are determined by repeatedly selecting, calculating and specifying environmental dispositions and user capabilities until atomic property pairs are identified that can be analysed. Finally, we have implemented the model for a routing scenario for mobility-impaired persons with different profiles.

With our model, we provide a framework for the process of rating environmental properties in terms of their expected suitability values for a particular user and task. Our method can support experts in deriving their estimation but also be applied to other GIS contexts where the meaning of attributive data is dynamically interpreted for individual users or tasks. By breaking the affordances down to an elementary level, where every property of the environment is confronted by just one property of the agent, the evaluation process is made more transparent and comprehensible, while at the same time allowing for more detailed results.

Practical problems lie in the high level of detail, which is necessary for all required data, from the environment and the user model to the rules defining the interrelationships, sub-affordances and standardization procedures. Moreover, when applying the model to a realistic scenario, it may quickly become very complex.

For future work, it is planned to further refine the model and apply it as a basic framework to a web-based routing service for mobility-impaired persons and an agent-based simulation of pedestrian movement in an urban setting.

References

Alfonso MA (2005) To walk or not to walk? The hierarchy of walking needs. Environ Behav 37(6):808–836

Chemero A (2003) An outline of a theory of affordances. Ecol Psychol 15(2):181–195

Clark S, Davies A (2009) Identifying and prioritising walking investment through the PERS audit tool. In: Walk21 Proceedings, 10th international conference for walking. New York, USA, 7–9 Oct 2009

Czogalla O (2011) Parameters determining route choice in pedestrian networks. In: TRB 90th annual meeting compendium of papers DVD. Washington, D.C., 23–27 Jan 2011
Gibson JJ (1977) The theory of affordances. In: Shaw R, Bransford J (eds) Perceiving, acting, and knowing: toward an ecological psychology. Lawrence Erlbaum, Mahwah, pp 67–82
Gibson JJ (1979) The ecological approach to visual perception. Houghton Mifflin Company, Boston
Heft H (2001) Ecological psychology in context: James Gibson, Roger Barker, and the legacy of William James's radical empiricism. Lawrence Erlbaum Associates, Mahwah
Janowicz K, Raubal M (2007) A affordance-based similarity measurement for entity types. In: Winter S, Duckham M, Kulik L, Kuipers B (eds) COSIT'07 proceedings of the 8th international conference on Spatial information theory, 19–23 Sept 2007
Jones KS (2003) What is an affordance? Ecol Psychol 15(2):107–114
Jonietz D, Timpf S (2012) Towards an affordance-based model of walkability. In: Timpf S (ed) Proceedings of the short papers of the SDH2012. Bonn, Germany, 21–24 Aug 2012
Jordan T, Raubal M, Gartrell B, Egenhofer M (1998) An affordance-based model of place in GIS. In: eighth international symposium on spatial data handling, 11–15 July 1998
Koffka K (1935) Principles of Gestalt psychology. Harcourt Brace, New York
Michaels CF (2000) Information, perception, and action: what should ecological psychologists learn from Milner and Goodale (1995)? Ecol Psychol 12(3):241–258
Ortmann J, De Felice G, Wang D, Daniel D (2012) An egocentric reference system for affordances. Available via semantic web journal. http://www.semantic-web-journal.net/sites/default/files/swj243_0.pdf. Accessed 10 Nov 2012
Raubal M (2001) Ontology and epistemology for agent based wayfinding simulation. Int J Geogr Inf Sci 15(7):653–665
Raubal M, Moratz R (2006) A functional model for affordance-based agents. In: Dagstuhl seminar towards affordance-based robot control. Dagstuhl Castle, Germany, 5–9 June 2006
Reed ES (1996) Encountering the world. Oxford University Press, New York
Scheider S, Janowicz K, Kuhn W (2009) Grounding geographic categories in the meaningful environment. In: Hornsby S, Claramunt C, Denis M, Ligozat G (eds) Spatial information theory, 9th international conference, COSIT 2009, 21–25 Sept 2009
Stoffregen TA (2000) Affordances and events. Ecol Psychol 12(1):1–28
Stoffregen TA (2003) Affordances as properties of the animal environment system. Ecol Psychol 15(2):115–134
Turvey MT (1992) Affordances and prospective control: An outline of the ontology. Ecol Psychol 4(3):173–187
Varela FJ, Thompson E (1991) The embodied mind. MIT Press, Cambridge
Warren WH (1984) Perceiving affordances: visual guidance of stair climbing. J Exp Psychol 105(5):683–703

The Effects of Configurational and Functional Factors on the Spatial Distribution of Pedestrians

Yoav Lerman and Itzhak Omer

Abstract The research presented here deals with pedestrian movement in two adjacent areas located in the city of Tel-Aviv that were established in different periods and according to different city planning doctrines: pre-modern and modern urban planning. Consequently, these areas differ in the street network spatial-configurational attributes and in the functional built environment attributes. Statistical and geographical analysis showed that in spite of their physical proximity, the two areas examined in this study differed significantly in the volume and the geographical distribution of pedestrian movement as well as in the explaining attributes of this distribution. It was found that in pre-modern environment, pedestrian movement is more predictable and has higher correlation to the spatial-configurational attributes of street network than in modern environment. The findings of this research can contribute to a greater understanding of the factors that shape pedestrian movement in pre-modern and modern urban environments.

1 Introduction

The attempts to predict pedestrian movement in urban environment point to two groups of attributes that affect this movement: spatial-configurational and functional.

In studies that deal with the effects of street network attributes on pedestrian movement, the degree of connectivity and accessibility of a street (or a street segment) is a major factor. This approach to the street network and its various accessibility measures is referred to as the configurational approach.

Y. Lerman (✉) · I. Omer
Department of Geography, Tel-Aviv University, Tel-Aviv, Israel
e-mail: yoavlerm@post.tau.ac.il

I. Omer
e-mail: omery@post.tau.ac.il

D. Vandenbroucke et al. (eds.), *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography,
DOI: 10.1007/978-3-319-00615-4_22,

The configurational approach relies on the topological features of the street network assuming these features in themselves are capable of explaining the spatial behavior of people in the city. The methodology of space syntax (Hillier 1996) is especially known within the configurational approach. This methodology is based upon topological-visual analysis of the environment. It is also used as a basis for the claim that the urban street network structure affects not only traffic but also land use patterns and other urban phenomena (Hillier et al. 1993; Hillier and Iida 2005).

Space syntax based analysis is conducted by using an axial map, which expresses the degree of accessibility from the perspective of the moving person. This analysis produces a number of topological indices of the street network (Hillier 1996) such as connectivity, integration and choice. The connectivity index reflects the number of axial lines directly crossing the subject line. The integration index characterizes the average topological distance from a specific axial line to all the other lines in the network.

The choice index expresses the probability that a specific axial line will be used as a passage between two places. This index is calculated based on the location of a given axial line on the shortest topological route between all the other axial lines in the network. Therefore, the choice value is greater for axial lines that are used to connect with shortest routes to other axial lines. The integration and choice indices can be calculated both on a global scale (for the whole network) or on local scales with different distance radii. For example, local integration using radius value of 3 ($r = 3$) indicates an axial line topological proximity to all the other axial lines in its vicinity up to three topological steps away.

Hillier and Iida (2005) have checked the correlation of pedestrian and vehicular movement to the topological, geometrical and metric attributes of the urban street network. They have used the least angle paths which have smallest accumulated angular change to represent a geometrical distance, and fewest turns which have the least number of direction changes to represent topological distance. For the metric distance they have used the measure of least length. They have found that of in 11 out of 16 cases they have checked least angle (geometric) correlations were best. In the other five cases the fewest turns (topologic) correlations were best, but only marginally better than least angle correlations. In no case the metrically based least length correlations were best and most of the times the metric measure had markedly lower correlation that the topologic and geometric based measures. According to Hiller and Iida there are three implications for their findings:

1. Geometrical and topological architecture of the large scale urban grid is the most powerful shaper of urban movement patterns.
2. Axial graphs in most circumstances are a perfectly good approximation of the impact of spatial configuration on movement.
3. The architecture of the street network, in both geometrical and topological sense, can be expected, through its effect on movement shows, to influence the evolution of land use patterns and consequently the whole pattern of life in the city.

Another study (Turner 2007) used angular segment analysis to compare topologic, geometric and metric measures correlations to vehicular movement. The best correlations were between the geometric measures and the movement volume. Turner have used angular segment analysis and stated:

> It is shown that the new model, using the betweenness measure of centrality, is a better empirical model of vehicular movement than earlier axial models using closeness centrality, with a correlation of up to $R^2 = 0.82$ in an application dataset from the Barnsbury area in London. In addition, it is shown that the angular measures correlate better with movement than methods using metric shortest paths between nodes.

The configurational approach was applied to predict the movement of pedestrians in many places. Several studies were conducted in London and have demonstrated predictability of pedestrian movement in the range of 55–75 % (Hillier et al. 1993; Hillier and Iida 2005; Jiang 2009a; Penn et al. 1998) and in Amsterdam the predictability was found to be in the range of 60–70 % (Read 1999). Jiang (2009a) expressed the influence of the topological structure on pedestrian movement in the following words:

> As we can see, over 60 % of human movement can be predicted or explained purely from a topological point of view. In terms of statistics, the other 40 % is not predictable, and it may relate to land use, building height and road width, etc.

The effects of land use distribution on pedestrian movement are also taken into consideration in the research presented here. A study conducted in Hong Kong (Chu 2005) found that the location of different land uses such as retail, offices and public transportation stations has an influence on the amount of pedestrians nearby. For example, large pedestrian volume was recorded near public transportations stations during commute hours. The analysis used multi-variable regression and found that commercial fronts had the highest correlation to pedestrian movement.

Additional studies attempted to combine land use and configurational factors. One such integrated study compared two adjacent neighborhoods in the center of Istanbul (Ozer and Kubat 2007) and examined the hypothesis that one of the neighborhoods had a lesser presence of pedestrians due to its spatial structure, which reduced its accessibility to pedestrians. It was found that the sense of safety (from crime and traffic) had the highest impact on pedestrian movement volume, while street accessibility (based on space syntax integration) and degree of mixed land uses were also significant factors.

In Boston an attempt was made to create models predicting the volume of pedestrian movement in four different areas (Raford and Ragland 2006) based on spatial structure and accessibility to various land uses. The authors used a 'stepwise multiple regression' analysis to measure the influence of each variable on existing observed pedestrian volumes. In all four areas the highest correlation was found between the pedestrian movement observed and space syntax integration. Other significant variables that improved the model were proximity to public transit stations and proximity to tourist attractions. The degree of the correlation to the actual pedestrian movement ranged between 0.57 and 0.86. Generally

speaking, in statistical terms 50–70 % of human movement can be predicted or explained purely from a topological point of view, depending on the representation of the street network and on street pattern type.

Despite the progress made, we have no sufficient knowledge on how spatial-configurational attributes of urban street networks relate to functional attributes and how they affect together pedestrian movement in various urban environments. In this chapter we concentrate on the difference between pre-modern and modern environments.

1.1 Research Questions

In light of the above discussion, two questions are aired concerning the differences between pedestrian movements in pre-modern and modern environments:

1. Which spatial-configurational attributes are more appropriate for predicting movement? More specifically, in which conditions topological representations capture the actual pedestrian movement in each of the investigated environments?
2. What is the impact of functional features (i.e. land uses and transit stops) that can mediate or affect the relation between street pattern morphology and human movement rates?

The aim of this chapter is to examine the difference between areas that were formed by different urban planning doctrines taking into account relation between the urban environment attributes and pedestrian movement. For this end, we conducted the empirical investigation in two adjacent areas in the city of Tel-Aviv that represent different planning concepts. The Sect. 2 presents the research methodology, the Sect. 3 presents the findings and the Sect. 4 discusses the findings and contributions of this study.

2 Method

The investigation focused on the relationship between the spatial-configurational and functional attributes of the built environment and volume of pedestrian movement in two adjacent areas in the center of the city of Tel-Aviv. The research was based upon pedestrian movement volume data collected using a survey in selected street segments in the research area. The functional attributes data were partially collected using a field survey and partially by using a geographic information system (GIS). The spatial attributes of the built environment were calculated using a topological analysis of the street network.

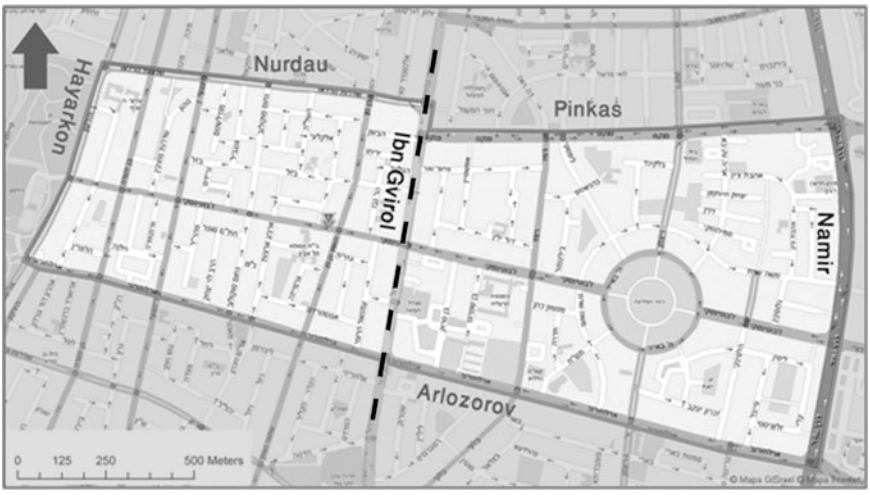


Fig. 1 Map of the research area with the shared boundary marked by a lineal line

Research Area

The selected study area covers 400 acres and is divided into two sub-areas which were designed and built during different periods, by different planning approaches, and have different characteristics. The study area is highlighted shown in Fig. 1 and its boundaries are as follows: North—Nurdau Boulevard and Pinkas Street; East—Namir Road; South—Arlozorov Street; West—Hayarkon Street.

Ibn Gvirol Street (indicated by a lineal line in Fig. 1) is the boundary that is shared by the western area and the eastern area. The western area was built during the 1930s according to a pre-modern master plan made by the Scottish urban planner, Sir Patrick Geddes, while the land was still under the jurisdiction of the British Mandate. The eastern area was built along more modern lines with more open spaces, less mix of land uses, wider major roads and less density during the 1950s as part of the "East Tel-Aviv" Plan (for further information on Tel-Aviv urban formation see: Marom 2009) after the establishment of the State of Israel. Although both areas share a major commercial street and have a few streets that traverse through both of them, the changes in the planning regime resulted in significant differences in residential density, street grid and land use mix.

Urban Environment Attributes

Calculation of the space syntax attributes was based on the axial map of the entire city of Tel-Aviv. The calculation was performed using the DepthMap[1] software.

[1] DepthMap is a software developed by UCL and is free for academic use: http://www.vr.ucl.ac.uk/depthmap/

For each of the street segments included in the survey data was collected about the following spatial-configurational and functional variables:

Spatial-configurational variables:

Street name based variables:

Street name connectivity—This index is based on the street network directly and not on the axial map used by space syntax analysis.

Axial based variables (topological space syntax):

Integration—both global and local (with r = 3)

Choice—both global and local (with r = 3)

Axial connectivity

Segment based variables (**geometric and metric space syntax**):

For the segment analysis we have used both geometric and metric analysis with different metric radii as follows: 250, 500, 750, 1000, 1250, 1500, 1750, 2000, 2500, 3000, 4000, 5000 m and no radius (entire system). The following variables were calculated for each of the metric radii:

Geometric mean depth
Geometric choice
Metric mean depth
Metric Choice

Functional variables:

Commercial fronts—Each street segment was given a value of 0, 1 or 2 depending on the amount of commercial fronts in it (retail on two sides, one side or none).

Proximity to public transit—Two variables were calculated to represent proximity to public transit:

No. of bus stations within a 100 m radius
No. of bus lines within a 100 m radius

Figure 2 illustrates the connectivity based on street names map (it can be seen that few of the streets continue outside of the research area) and Fig. 3 shows the commercial fronts distribution.

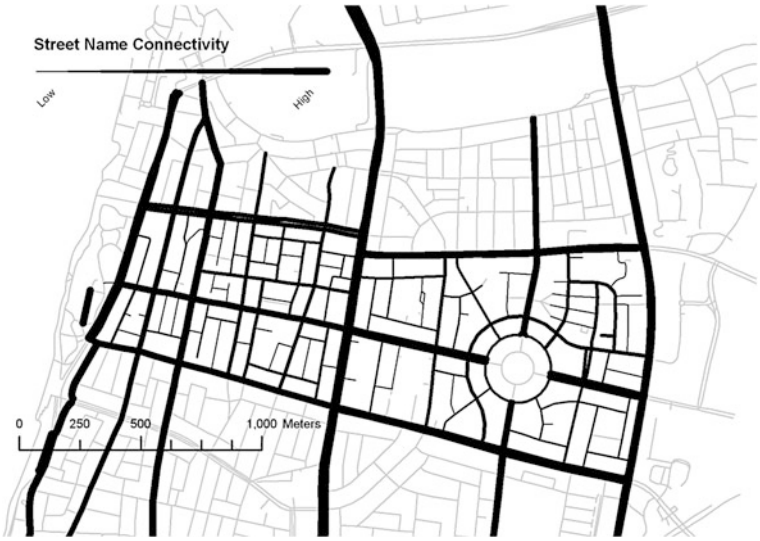


Fig. 2 Map of the street name connectivity in the research area

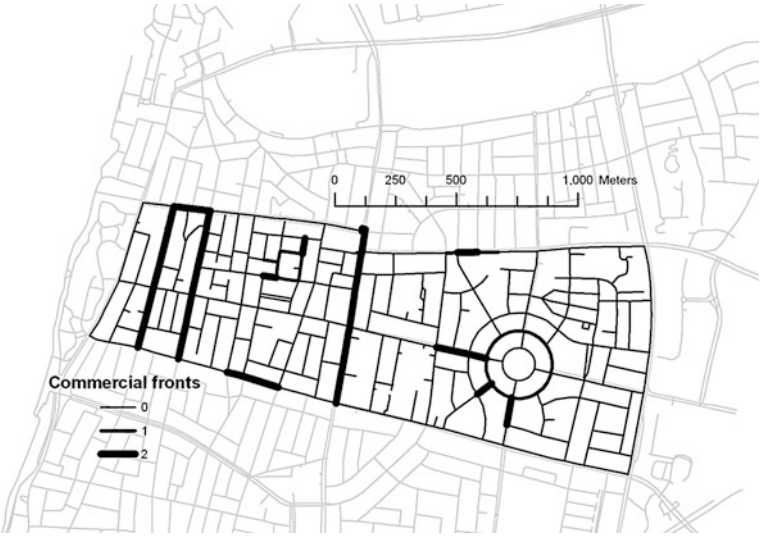


Fig. 3 Map of the commercial fronts distribution in the research area

3 Results

Pedestrian Movement Survey

Measurement points in the study area were selected to represent a range of different centrality measures and distribution of land uses. The measurement was carried out in 95 different points and in all locations sidewalks on both sides of the street were measured. The method of gate counts was used in each survey point and the amount of pedestrians who passed through the gate was counted. The count was done for 5 min every hour for 5 h in each point. The measurement took place on a sunny weekday between the hours 3 and 8 pm. Since most of the survey points were part of a pair of points on both sides of the street the amount of street segments where pedestrian movement was measured is different from the amount of the survey points.

The street segments for which pedestrian movement volume data was collected were divided as follows: in the entire study area pedestrian movement volume was collected for 51 different street segments. Of these 24 street segments were in the western area and another 24 street segments were in the eastern area. Three street segments were on Ibn-Gvirol Street—the boundary street that both western and eastern areas share. The calibrated average of the measured pedestrian volume per hour is displayed for each of the 51 street segments in Fig. 4.

The average for the 51 segments surveyed was 250.45 pedestrians per hour. It can be seen that on average the segments in the western area had more pedestrian movement. The average for the 24 segments in the western area was 280.7 pedestrians per hour while the average for the 24 segments of the eastern area was

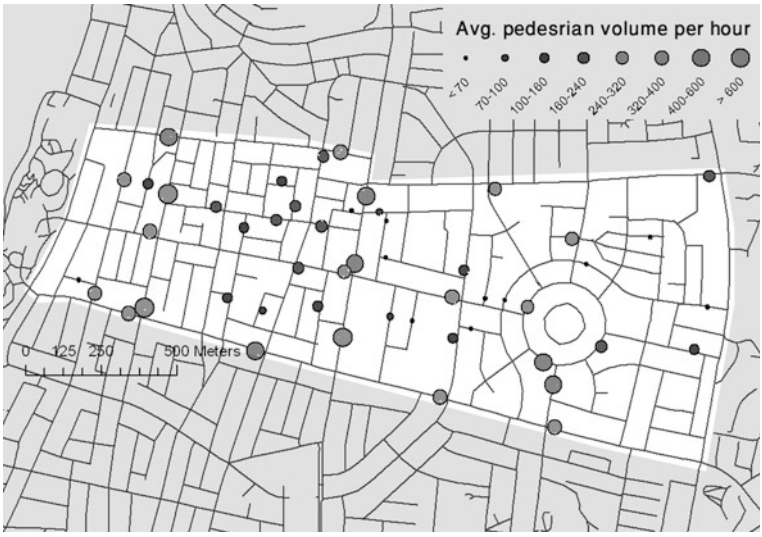


Fig. 4 Map of pedestrian movement survey results

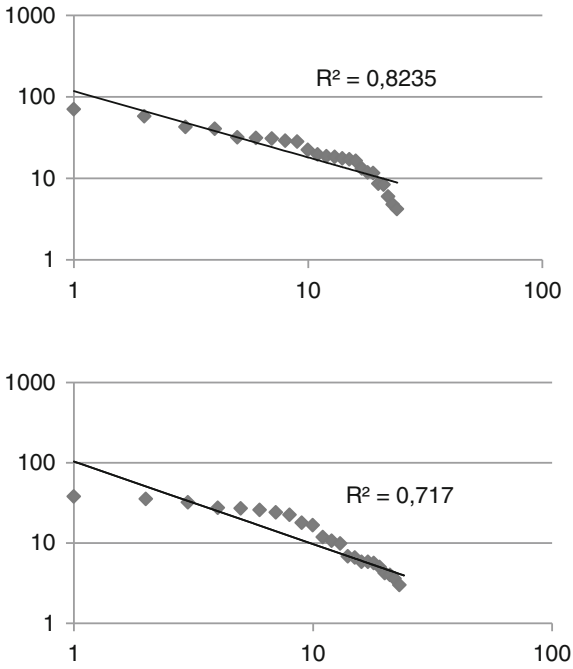


Fig. 5 Power law distribution (Log_{10}—Log_{10} scale plot) of distribution of pedestrian movement volume in the western area (*above*) and in the eastern area (*below*)

174.6 pedestrians per hour. The eastern area had on average 62 % of the pedestrian movement that the western area had.

As Fig. 4 shows, the distribution of the pedestrian flows creates a sort of hierarchical structure. It can be seen that there are high pedestrian flows in the western section that create a kind of spatial organization; the locations with high level of pedestrian movements are distributed more or less equally over all the western section and tend to relate to the weaker pedestrian movements.

Another significant difference between the examined areas is related to the distribution of pedestrians in each of the examined areas. As can be seen in Fig. 5 the distribution of pedestrians in the western area tends to be closer to power law distribution. This tendency indicates that the pedestrian movements in the western area have greater similarity to a self-organized process. The existence of power laws is one of the most striking signatures that the distribution of movement flows as well as street network centralities in cities are a product of self-organization that characterize unplanned urban environment (Jiang 2009b). The power-law behavior represents the level of heterogeneity of the pedestrian movement distribution among the different streets; while there are few streets with strong pedestrian flows, a significant majority of streets carry weak pedestrian flows.

Built Environment Differences

Table 1 summarizes the average values of the topological centrality attributes based on axial line and street names and the average values of the functional

Table 1 Spatial-configurational attributes averages of research area

	Both areas together	Western area	Eastern area
Connectivity by street name	6.12	6.73	5.61
Integration	1.16	1.17	1.15
Local integration	2.48	2.69	2.36
Choice (Log)	8.19	7.78	8.62
Local choice (Log)	4.13	4.30	4.11
Axial connectivity	5.63	6.51	5.51
Intelligibility	0.42	0.64	0.35

attributes for both areas, separately and together as a whole. It can be clearly seen that the western area has higher average levels of local integration, axial connectivity and street name connectivity. On the other hand, the eastern area has higher level of global choice. Another significant configurational attribute is the intelligibility second-order index which represents the correlation between global integration and connectivity axial-based measures and gives an indication on the legibility of the space (Hillier 2002). The intelligibility of the western area is much higher than that of the eastern area (0.64 compared to 0.35) and is also higher than both areas taken as one. The spatial-configurational and functional averages for both areas together and on their own are shown in Table 1.

Correlations with Built Environment Factors

We assumed that the different spatial configuration of the two areas as well as the different functional arrangement might be related to different movement patterns and different correlations between the pedestrian movement volume and the urban environment attributes. Bivariate correlations (R^2) between the log function of the pedestrian movement and the spatial-configurational attributes as well as the functional attributes are shown in Table 2 for both areas as one set (51 cases) and for each area on its own (24 cases). All choice and connectivity variables had a power law distribution and were normalized using the log function for the various correlations described below. All correlations with the spatial-configurational variables were statistically significant ($P < 0.01$). Correlations of pedestrian movement with the functional variables were less significant than those with spatial-configurational variables in both areas.

From Table 2 it's clear that in the western area there are higher correlations between pedestrian movement and spatial structure in all of the attributes. The dominant spatial-configurational attributes of connectivity and integration tended to be local rather than the global variables indicating that pedestrian movement is more locally oriented. Another aspect of the difference between both areas can be found in the fact that the best correlation in the western area is with street name based connectivity ($R^2 = 0.78$), while in the eastern area the best correlation is with the axial-based local choice ($R^2 = 0.55$).

Table 2 Correlations of pedestrian volume (R^2) with spatial-configurational and functional attributes (the highest correlations are marked in bold). All correlations are statistically significant ($P < 0.01$) unless mentioned otherwise

Log pedestrian volume	Both areas together	Western area	Eastern area
Connectivity by street name (Log)	0.59	**0.78**	0.43
Global integration	0.38	0.33	0.30
Local integration	0.56	0.48	0.50
Global choice (Log)	0.36	0.40	0.31
Local choice (Log)	**0.63**[a]	0.57[b]	**0.55**
Axial connectivity (Log)	0.57	0.53	0.48
Total commercial fronts	0.40	0.43	0.25[*]
Bus station count	0.23	0.17[*]	0.23[*]
Bus line count	0.08[*]	0.06[*]	0.09[**]

[a] 50 cases out of 51
[b] 23 cases out of 24
[*] Significant with $P < 0.05$
[**] Not significant

In addition, the local choice variable had the highest correlation with pedestrian movement in both areas taken together and the second highest correlation in the western area. It can be also seen that correlations of pedestrian movement with functional attributes were much lower than those of the spatial-configurational factors in all areas. Scatter plots of the log–log function between pedestrian volume and the dominant spatial-configurational attributes of local choice and street name connectivity are shown in Fig. 6. These scatter plots demonstrate clear differences between these two variables. While street name connectivity had a rather big variation in its correlation to pedestrian movement in the different areas, ranging from $R^2 = 0.43$ in the eastern area up to $R^2 = 0.78$ in the western area, local choice had similar correlation levels to pedestrian movement for all areas ranging from $R^2 = 0.55$ for the eastern area to $R^2 = 0.63$ for both areas taken as one.

Bivariate correlations between pedestrian movement and the segment based variables were generally weaker than those of the axial line based variables. However, even so, the western area had higher correlation between pedestrian movement and segment based variables than the eastern area. On average, higher correlation was achieved with geometric segment variables than with metric segment variables. Bivariate correlations between pedestrian movement and all the segment based variables are shown in Appendix 1.

The next step was to conduct a multivariate regression analysis by using the log function of the pedestrian movement. The multivariate regression was done using stepwise regression with the entire set of the variables described above. The best correlations and variables for each area as well as for both areas together are shown in Table 3. It can be seen that the best correlation ($R^2 = 0.88$) is achieved in the western area, compared to the eastern area and both areas taken together ($R^2 = 0.82$ for both). Adding the functional variable representing commercial fronts has added to the model correlation in all cases. All correlations presented Table 3 are significant ($P < 0.01$).

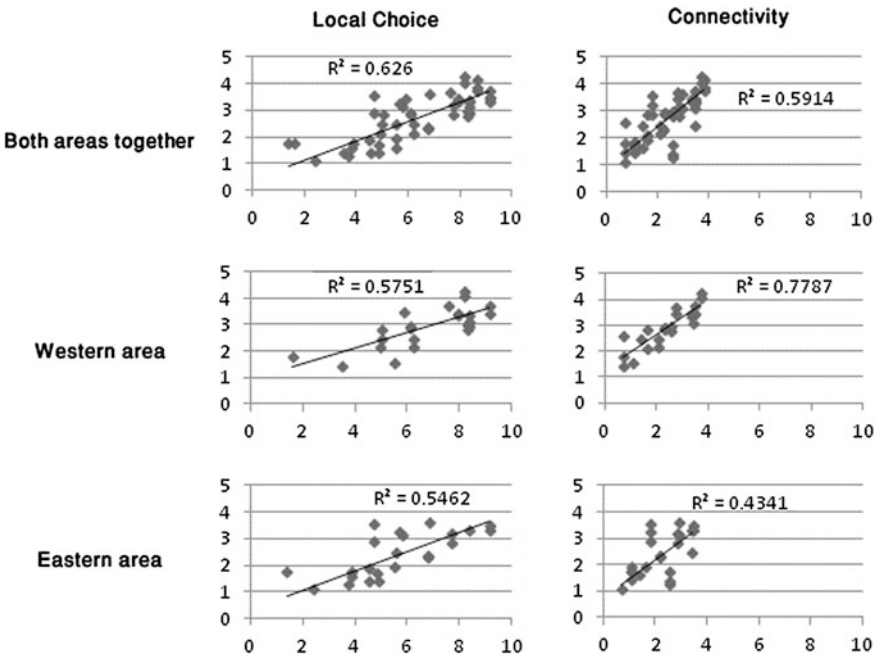


Fig. 6 Scatterplots of the log–log correlation between pedestrian movement (y-axis) and the spatial-configurational attributes (x-axis): axial based local choice (*left*) and street named based connectivity (*right*). Scatter plots for both areas taken together are in the *top row*, for the western area in the *middle row* and for the eastern area in the *bottom row*. For clarity sake points with negative log value were removed from the displayed graphs (at most one point per graph)

Table 3 Multivariate correlations of pedestrian volume (R^2) with all of the attributes. All correlations are statistically significant ($P < 0.01$)

Log pedestrian movement	Cumulative variables	Cumulative R square
Both areas together*	Local choice	0.63
*50 cases out of 51	Total commercial fronts	0.74
	Global choice	0.80
	Street name connectivity	0.82
Western area	Street name connectivity	0.78
	Total commercial fronts	0.85
	Mean depth R = 5,000 m	0.88
Eastern area	Local choice	0.55
	Global choice	0.75
	Total commercial fronts	0.82

Thus, the correlation between urban environment characteristics—spatial structure and land use—is significantly higher in the western area. Such difference in the correlation coefficient of the two areas can be related to the intelligibility

level in these two areas. That is, high level of intelligibility enables better prediction of pedestrian movement (see for example: Zhang et al. 2012).

The correlation of the commercial fronts to spatial-configurational variables was checked and exposed another meaningful difference between both areas. In the western area the highest correlation for the commercial fronts variable was achieved with regards to street name connectivity (which also had the strongest correlation with pedestrian movement in the western area) at $R^2 = 0.22$ with $P < 0.05$. In the eastern area no significant correlation was found between the commercial fronts and any of the spatial-configurational variables, meaning that the location of the commercial fronts in the eastern area had less to do with the street network. This aspect may also explain the lower movement rates of pedestrians in the eastern area and the lower correlations between the spatial-configurational attributes and the pedestrian movement in that area.

Bus stations did not come out as a significant factor in the multivariate regression analysis with pedestrian movement and had a slightly stronger correlation to spatial-configurational attributes in the eastern area (correlation of $R^2 = 0.38$, $P < 0.01$, with axial-based connectivity) compared to the western area (correlation of $R^2 = 0.28$, $P < 0.01$, with street name-based connectivity).

4 Discussion

The research presented here is unique in dealing with two adjacent areas that were established in different periods and according to different city planning doctrines: pre-modern and modern urban planning. This physical proximity enables focusing on the spatial and functional built environment differences between the areas while minimizing socio-economic differences as well as functional aspects related to the position of the areas in the city.

A statistical analysis of the relation between the built environment attributes—spatial-configurational and land uses attributes—showed that these two areas, in spite of their physical proximity, differed significantly both in the volume and in the spatial distribution of the pedestrian movement as well as in the effect of the built environment attributes on these aspects of pedestrian movement.

It was shown that there is a better correlation between pedestrian movement and spatial-configurational attributes of the built environment than between pedestrian movement and functional attributes. This result is similar to those found in previous studies (Jiang 2009a; Raford and Ragland 2006). However, it was also found that he older and denser western area proved to be more conducive to pedestrian movement and had over 50 % higher movement rate of pedestrians than the eastern area. The western area also had higher spatial-configurational averages, higher correlation between the spatial-configurational attributes and pedestrian movement and higher intelligibility score.

The greater fit of the spatial structure with the land use distribution, especially retail, in the western area can explain, at least partly, the relatively higher pedestrian presence in the western area as well as better correlations between the spatial structure and pedestrian traffic. While the commercial fronts in the western area can be found along the streets with higher centrality values, in the eastern area the commercial fronts are somewhat artificially concentrated around one big square. The eastern area exhibits a separation of the spatial-configurational and functional attributes, which hampers the flow of pedestrian movement in that area. We tend to relate this contrast between both areas to the difference between the planning approaches that stand behind the establishment of these two areas.

The data for this study was gathered from one area in the city of Tel-Aviv, which might have caused the results to be too particular. Further exploration of pedestrian movement patterns in different areas would seek to collect and examine data from more adjacent areas which differ in their urban planning concepts and have significant spatial and functional differences.

Other factors which were not included in this chapter are physical factors such as pavement width as well as demographic factors and residential and employment densities. These factors may also be influencing the way pedestrian volume is distributed and will be revisited in future research. In addition, following the findings related to named street based connectivity, future work should examine more named street—based variables such as Closeness and Betweenness centrality measures (Jiang and Claramunt 2004).

The practical contribution of this study might be the ability to intervene in the relevant environmental attributes in order to improve pedestrian movement in various urban environments (see also Ozbil et al. 2011).

5 Appendix 1 Bivariate Correlation of the Segment Based Variables

The R^2 values for the correlation between the pedestrian movement and segment based variables are shown in Table 4. Both geometric and metric variables are shown in various radii defined in meters. Variables with no radius in their name are global variables and are calculated with no limiting radius.

MDR = Geometric Mean Depth
MMDR = Metric Mean Depth
Tch = Geometric Choice
Mch = Metric Choice

Table 4 Correlations of pedestrian volume (R^2) with segment based variables. All correlations are statistically significant ($P < 0.01$) unless mentioned otherwise

Log Pedestrian Volume	Both areas together	Western area	Eastern area
MDR250	0.07**	0.38	0.03**
MDR500	0.21	0.39	0.08**
MDR750	0.36	0.43	0.18*
MDR1000	0.49	0.54	0.32
MDR1250	0.52	0.59	0.34
MDR1500	0.54	0.64	0.38
MDR1750	0.56	0.61	0.41
MDR2000	0.54	0.59	0.38
MDR2500	0.53	0.52	0.40
MDR3000	0.54	0.50	0.43
MDR4000	0.52	0.47	0.38
MDR5000	0.47	0.40	0.36
Geometric Mean Depth	0.46	0.44	0.33
MMDR250	0.00**	0.00**	0.02**
MMDR500	0.04**	0.13**	0.07**
MMDR750	0.00**	0.01**	0.00**
MMDR1000	0.00**	0.00**	0.00**
MMDR1250	0.01**	0.05**	0.00**
MMDR1500	0.03**	0.25*	0.00**
MMDR1750	0.04**	0.16**	0.00**
MMDR2000	0.04**	0.24*	0.03**
MMDR2500	0.05**	0.01**	0.02**
MMDR3000	0.07**	0.00**	0.00**
MMDR4000	0.09*	0.01**	0.00**
MMDR5000	0.11*	0.03**	0.01**
Metric Mean Depth	0.00**	0.06**	0.17*
Tch250	0.01**	0.00**	0.00**
Tch500	0.10*	0.18*	0.02**
Tch750	0.24	0.34	0.12**
Tch1000	0.33	0.43	0.21*
Tch1250	0.36	0.47	0.25*
Tch1500	0.39	0.49	0.29
Tch1750	0.40	0.51	0.29
Tch2000	0.38	0.48	0.29
Tch2500	0.39	0.50	0.29
Tch3000	0.38	0.49	0.28*
Tch4000	0.36	0.49	0.22*
Tch5000	0.33	0.50	0.17*
Geometric Choice	0.34	0.47	0.15**
Mch250	0.04**	0.00**	0.01**
Mch500	0.05**	0.03**	0.00**
Mch750	0.10*	0.09**	0.01**
Mch1000	0.17	0.19*	0.04**
Mch1250	0.23	0.25*	0.09**

(continued)

Table 4 (continued)

Log Pedestrian Volume	Both areas together	Western area	Eastern area
Mch1500	0.25	0.27*	0.11**
Mch1750	0.27	0.28	0.13**
Mch2000	0.28	0.30	0.14**
Mch2500	0.30	0.32	0.17*
Mch3000	0.29	0.32	0.17*
Mch4000	0.28	0.29	0.17*
Mch5000	0.24	0.23*	0.16**
Metric Choice	0.19	0.16**	0.15**

* Significant with $P < 0.05$
** Not significant

References

Chu SCH (2005) When and why do people walk in the city: the influence of urban elements on time-pattern of pedestrian movement. In: 6th international Walk 21 conference in Zurich

Hillier B (1996) Space is the Machine. Cambridge University Press, Cambridge

Hillier B (2002) A theory of the city as object: or, how spatial laws mediate the social construction of urban space. Urban Des Int 7:153–179

Hillier B, Penn A, Hanson J, Grajewski T, Xu J (1993) Natural movement: or, configuration and attraction in Urban pedestrian movement. Environ Plann B: Plann Des 20:29–66

Hillier B, Iida S (2005) Network effects and psychological effects: a theory of urban movement. In: 5th International space syntax symposium in Delft

Jiang B (2009a) Ranking spaces for predicting human movement in an urban environment. Int J Geograph Inf Sci 23(7):823–837

Jiang B (2009b) Street Hierarchies: a minority of streets account for a majority of traffic flow. Int J Geograph Inf Sci 23(8):1033–1048

Jiang B, Claramunt C (2004) A structural approach to the model generalization of an urban street network. Geo Informatica 8:157–171

Marom N (2009) City of concept: planning Tel-Aviv. Babel Press, Tel-Aviv in Hebrew

Ozbil A, Peponis J, Stone B (2011) Understanding the link between street connectivity, land use and pedestrian flows. Urban Des Int 16:125–141

Ozer O, Kubat AS (2007) Walking initiatives: a quantitative movement analysis. In: 6th International space syntax symposium in Istanbul

Penn A, Hillier B, Banister D, Xu J (1998) Configurational modeling of urban movement networks. Environ Plann B: Plann Des 25:59–84

Raford N, Ragland DR (2006) Pedestrian volume modeling for traffic safety and exposure analysis: case of Boston. Massachusetts. Transportation research board 85th annual meeting

Read S (1999) Space syntax and the Dutch city. Environ Plann B: Plann Des 26:251–264

Turner A (2007) From axial to road-centre lines: a new representation for space syntax and a new model of route choice for transport network analysis. Environ Plann B: Plann Des 34:539–555

Zhang L, Zhuang Y, Dai X (2012) A configurational study of pedestrian flows in multi-level commercial space. Case study Shanghai. In: 8th International space syntax symposium in Santiago de Chile

Examining the Influence of Political Factors on the Design of a New Road

Paulo Rui Anciaes

Abstract Transport policy is often influenced by political factors. The few existing studies on this topic analyse patterns arising from a series of past decisions. This chapter adds to this knowledge by exploring the implications of political bias in the design of an individual project. The objective is to compare socially and politically optimal decisions for different hypothesis about the level and nature of political bias. GIS methods are used to derive route alignments for a new road, taking into account the level and distribution of community severance effects. The results show that even when bounding the problem with several restrictions, attending to political interests leads to deviations from the social optimum, producing alignments not only with higher aggregate severance impacts, but also with higher land use costs and important distributional effects.

1 Introduction

The political organizations responsible for public policies are not neutral and have their own interests and motivations. The pursuit of these interests may clash with the social good. This means that a politically-biased policy may lead to an allocation of resources and distribution of benefits and costs that is not consistent with society's views on broad objectives such as economic efficiency, social equity or environmental sustainability. In general, the application of public policies is subject to the interaction of three different types of political structures: political parties, interest groups and social movement organizations (Burstein and Linton 2002). The decisions of policy-makers may then depend on their assessment of

P. R. Anciaes (✉)
London School of Economics and Political Science, London, UK
e-mail: P.R.Anciaes@gmail.com

D. Vandenbroucke et al. (eds.), *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography,
DOI: 10.1007/978-3-319-00615-4_23,

potential electoral gains or losses and may be permeable to lobby pressures or to protest by the local populations affected by the policy.

The influence of political factors is especially relevant in the case of transport planning, due to the asymmetry between the political power of well-organized lobbies, such as the car industry, and the power of other potentially interested parts, such as local business organizations, residents' associations, small environmentalist groups and non-car users. For example, Hillman (1997, pp. 72–77) argues that pedestrians and cyclists are systematically discriminated against in public policy, at the level of information gathering and decision-making. Empirical studies have also proved the influence of political factors on the regional allocation of investments in transport infrastructure (Congleton and Bennett 1995; Castells and Solé-Ollé 2005, Walden and Eryuruk 2012; Jussila Hammes 2012) and on decisions about investment and disinvestment in public transport infrastructure (Boschken 1998; Brent 1976). Delays and failures to introduce policies such as congestion charges and traffic restriction are also linked to the lack of public acceptability and to the socio-political characteristics of the population affected (Schade and Schlag 2003).

While there is scarcely any empirical evidence on the hypothesis of political bias regarding the distribution of the environmental effects of transport, some conclusions can be imported from studies of policies in other sectors. The question is often approached in reference to the principle of environmental equity. Camacho (1998, p. 18) argues that environmental inequalities arise because the governments depend on the resources of upper-class groups and large business. Several studies have shown that differences in the communities' political characteristics and especially in political power may explain options taken in environmental policy and in the location of pollution facilities leading to disadvantages for low-income groups and racial minorities (Hamilton 1993, 1995; Brooks and Sethi 1997; Earnhart 2004).

This chapter adds to the literature by focusing on the role of political aspects at the level of individual transport projects, focusing on the case of the construction of a new arterial road. Existing empirical work in both transport and non-transport sectors deals with broad options, where the influence of political factors operates at the level of discrete choices, either between the implementation or not implementation of a project and between the different locations of the project. There is however a lack of knowledge on the sensitivity of the design of specific projects to the influence of political factors, in the cases where the social worth of the project and the area of implementation have already been decided. While Taylor et al. (2010) show that the policy-maker is permeable to political influences at the stage of route alignment of new transport infrastructure, these influences have not been studied by reference to the political characteristics of the populations living in the neighbourhoods potentially affected by the project.

The production of further knowledge on this question is especially relevant in the case of urban transport projects, as the different alternatives for the alignment of the new project imply different distributions of local environmental effects among neighbourhoods, given the necessity for the project to cross areas with

generally high population densities. In addition, the influence of political factors may also be limited by geographic constraints and opportunity costs in terms of land use. The potential for introducing political bias in road projects then depends on the usually small set of feasible options available to the policy-maker.

The chapter also introduces two methodological novelties. The first one regards the definition of the political motivations of the policy-maker. Existing studies have worked with a priori measures of political pay-off, such as voters in specific parties or the probability of local populations of engaging in collective action. However, these measures may not be representative of the political interests of the policy-maker or may give only a partial view of those interests. Our approach is to assess the implications of different hypotheses about the nature of the motivations of the policy-maker. The focus is not on the behaviour of a specific political party but on an abstract policy-maker, whose political interests are multidimensional and depend on the political characteristics of the populations affected by the project. These characteristics are found adopting a data-driven approach.

The second methodological novelty is to evaluate route alignments considering their impact on community severance. This impact is one of the major sources of social and political protest at the local level following the construction of new transport infrastructure. However, severance effects are seldom analysed quantitatively, either in official evaluation studies comparing different alternatives for the infrastructure or in academic studies focusing on patterns arising from past policies.

The study is supported by a geographic information system (GIS) in all stages. An indicator of community severance is defined at the level of the census enumeration district, based on the connectivity of the district with a series of possible destinations for pedestrians. We then estimate the potential changes in community severance in all districts when the road crosses a given point. This information is used to derive the spatial distribution of the political costs of severance. These costs is defined in alternative ways, depending on the weights attributed to the variables that define the political characteristics of the population affected. An optimal path algorithm is then used to derive the route alignments that minimize political cost. These alignments are compared with the one proposed in municipal master plans and with the "socially optimal" alignment, defined as the one that minimizes aggregate severance effects. The comparison uses statistics based on the overlay of the alignments with other geographic information, such as land use and the socio-economic structure of the population.

The chapter is organised as follows. The next section describes the project and defines the optimal route alignment problem. We then analyse the political structure of the population potentially affected by the project, deriving vectors of variables used in subsequent analysis. The main section of the chapter presents the politically-optimal routes and compares them with the planned and socially-optimal routes. The final section concludes the chapter, discussing the results and suggesting questions for further work.

2 Problem Definition

The case study is a proposal for a new road in the Lisbon Metropolitan Area (Portugal), linking Lisbon with two municipalities to the west (Cascais and Oeiras). This road is a project planned at the municipal level and figures in the municipal master plans of both municipalities as "Via Longitudinal" (Longitudinal Road). The purpose is to create a new access corridor to Lisbon, improving the accessibility of an area that lacks a clear network of arterial roads and direct access to Lisbon. The new road will also ease congestion on the existing motorway crossing the middle part of both municipalities and contribute to a more rational distribution of road traffic at the entrance of Lisbon, when used in conjunction with a future arterial road in the western part of this city. However, as the main longitudinal arterial road in the region, the project will have multiple lanes and high levels of traffic and average speeds, with potential effects on local mobility, accessibility to local urban facilities and inter-community interaction. These effects apply not only to walking trips but also to trips by bicycle or local buses, as these trips will suffer delays and will be associated with increased accident risk and increased exposures to noise and air pollution at the intersections with the new road.

The analysis in this chapter captures concerns on these effects by using an indicator of community severance defined at the level of the census enumeration district. In the construction of this indicator, we assume that the population in each district has a specific set of potential pedestrian destinations. This set is obtained by a sampling process that ensures that each district has between 4 and 12 destinations at a maximum distance of 800 m. In each district, we assign to each destination an attractiveness score, equal to its population density corrected by a factor depending on distance. The indicator of community severance in the district is then the attractiveness score of the pedestrian destinations that cannot be reached on the street network unless crossing large transport infrastructure, given as a proportion of the attractiveness score of all the destinations of that district.

The use of community severance has methodological advantages over other indicators of the local environment costs of the project, as it depends only on the location of infrastructure (and not on traffic levels) and allows an unequivocal identification of the set of neighbourhoods affected when the road crosses a given point, as the road will not cross any walking route more than once. The effect on each neighbourhood can be estimated by selecting all the routes that link the neighbourhood with its pedestrian destinations and that cross that point. This information is then used to estimate the proportion of population-interaction potential that those routes represent in the set of all destinations of the neighbourhood.

Figure 1 illustrates the elements used to define the optimal route alignment problem. The routes link the two future junctions with the existing network, as defined in the municipal master plan, and are calculated over a cost surface modelled with GIS raster data with a resolution of 40 m. The set of feasible areas is restricted to a corridor, based on the rationale of the project. This corridor is defined by existing or planned motorways, except in the north part, where it

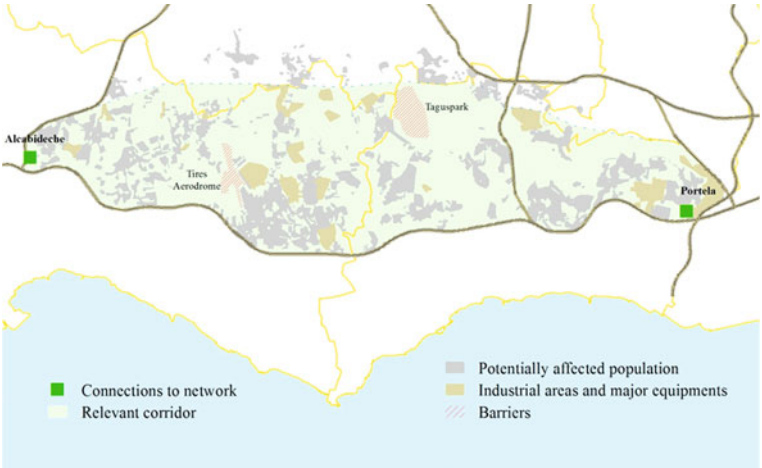


Fig. 1 Elements for the estimation of optimal routes for the *Via Longitudinal*

follows roughly the borders of the two municipalities. The districts potentially affected by the new road are the ones for which at least one destination is only reachable by crossing a point inside the relevant corridor. We also restrict the new road from crossing a set of areas inside the corridor, including an aerodrome, a large office park and locations with special cultural or environmental importance, given in the Portuguese National Gazetteer. Further restrictions are imposed on residential areas, industrial sites and major urban facilities, where the road can only use the space already accommodating arterial roads. This condition implies that in that space, the project corresponds to the upgrade of existing roads and not to the construction of a new road. We allow the same possibility in the space occupied by the three transversal motorways bordering and crossing the area, which are thus included in the feasible space for the new road.

The politically optimal route alignments minimize the political value of the aggregate community severance effects across the region (C_r). We assume that there is a single policy-maker, that is, the same party is holding the governments of the two municipalities involved. Political value is obtained by assuming that the policy-maker places a higher weight on the effects on the population with certain political characteristics, measured by a set of standardized factors described in the next section. The cost of the road crossing point r is then the sum of the severance effects in all the i affected enumeration districts ($\Delta CS_{r,i}$), weighted by population POP_i and by an exponential function of one of the political factors (P_x). The parameter ε measures the degree of priority assigned to the areas where the political factor is above the mean (that is, above 0). The value of this parameter is zero when the political factor is below the mean.

$$C_r = \sum_i POP_i * \Delta CS_{r,i} * Exp(\varepsilon * P_x)$$

The socially optimal route corresponds to the case $\varepsilon = 0$, that is, when no weights are placed on the political characteristics of the populations affected by severance. In this case, the route minimizes the sum of severance effects in all district weighted by their population.

3 Mapping the Political Landscape

The analysis in this section extracts from elections data a set of components measuring the political characteristics of the population in the areas affected by the project. These components are interpreted as independent elements in possible political strategies for the policy-maker. They may or may not correspond to the interests of a specific political party, as those interests depend on the political context at the time of decision and the strategy adopted by the party (for example, the choice between focusing on securing votes or on attracting new votes).

The hypothetical moment of decision for the projects is 2001. We consider that the political characteristics of the population in each area can be observed in the set of all elections hold in Portugal since 1991. The data come from the official election results database and include seven elections, for the central and local governments and for European Parliament. For each election, we define six variables. The first two variables are the vote shares of the two parties that dominate the political spectrum in Portugal, labelled in this study as "Orange" (right wing) and "Pink" (left wing). As possible indicators for social protest, we include the shares of blank and null votes and the shares of a set of left-wing parties with traditionally active militancy and labelled "Red". As issues of community severance have an environmental dimension, it is also relevant to include the vote share of the two Portuguese environmental parties, labelled as "Green". The final variable is the abstention rate, which can assess both the probability of electoral gains and the potential for social protest.

Possible electoral gains and losses for a party depend on the size and loyalty of its electoral base in each neighbourhood, while similar considerations apply to the probability of social protest, as measured by associated voting patterns. As such, we construct two sets of variables aggregating the data from the seven elections considered, containing the averages and the standard deviations of the variables defined above. The two sets of variables are then entered in the factor analysis.

The analysis is conducted at the level of the smallest administrative area (*freguesia*—civil parish) and as such, the resulting factor scores need to be disaggregated at the enumeration district level. This is achieved by estimating a regression model between the political factors and five socio-economic factors: Age, Qualifications, Urbanization, Slums and Migration.[1] The model is used to

[1] The Migration factor is related to variables such as length in current residence, place of birth, single-person households, shared dwellings and proportion of males in the adult population.

predict the political characteristics of the population in each district. The five factors were obtained by a previous factor analysis to census data at the enumeration district level and then averaged at the level of the administrative area.

Tables 1 and 2 show respectively the results of the factor analysis and the regressions between political and socio-economic factors. The factor models are satisfactory, extracting high proportions of the total variance and of the variance of each individual variable (as shown by the value of the communality), while the goodness of fit of the regression models is satisfactory for the three political factors.

The most distinctive factor (P1), explaining half of the variance in the dataset is the "Left" versus "Right" ideological opposition. This factor is characterized by high loadings on the historical averages of the shares of the parties representing the two ideological poles and by high loadings with opposite signs on the standard deviations of those shares. The regression models show that this factor is mainly explained by socio-economic status and to a smaller degree, by age and urbanization levels. The second factor (P2) groups variables suggesting the level to which communities participate in the political process. The variables with high negative loadings within this factor are the abstention rate and the shares of blank and null votes. The main explanatory variable for the factor is urbanization levels.

Table 1 Factor analysis of elections data (1991–1999)

		P1 Left	**P2** Political	**P3** Potential	*Communality*
% of variance		50.2 %	18.6 %	14.8 %	
Averages	Orange	−0.96	−0.08	−0.16	0.94
	Rose	0.92	−0.10	0.21	0.90
	Red	0.91	0.29	0.00	0.91
	Green	0.88	0.34	−0.02	0.89
	Blank/Null	0.50	−0.71	0.24	0.81
	Abstention	0.38	−0.68	−0.45	0.82
Std. Errors	Orange	0.71	−0.10	0.62	0.89
	Rose	−0.29	−0.68	0.63	0.94
	Red	−0.72	0.25	0.25	0.64
	Green	−0.92	−0.11	0.22	0.91
	Blank/Null	−0.54	−0.50	0.05	0.55
	Abstention	−0.18	0.51	0.75	0.85

Table 2 Regression between political and socio-economic factors

	P1	**P2**	**P3**
S1 (Age)	−0.24	−0.25	−0.11
S2 (Qualifications)	−0.84	0.08	0.24
S3 (Urbanization)	0.21	0.90	−0.52
S4 (Slums)	0.10	−0.33	−0.47
S5 (Migration)	0.19	0.10	−0.33
$\mathbf{R^2}$	**0.85**	**0.68**	**0.58**

The third factor (P3) is related to a low but variable abstention rate and to the variability in the shares of the Orange and Rose parties. This can be understood as a measure of the potential for both parties to attract new voters. This factor depends negatively on the urbanization level, location of slums and migration.

4 Politically-Optimal Route Alignments

4.1 Routes

This section presents the optimal routes obtained for costs based on the political factors obtained above. In the case of the first factor (Left *vs*. Right), the analysis considers both the factor scores (Left) and their symmetrical (Right). The estimation of the optimal routes uses Dijkstra's least-cost algorithm, implemented in ArcGIS 9.2 Spatial Analyst. Figure 2 shows for each political factor, the optimal route alignments obtained for different value of the parameter ε. The routes are mapped in Fig. 3, which also includes the route of the project as planned. The estimation produced five different politically-optimal routes. Route Ao1 is the socially-optimal route.

All the estimated routes differ considerably from the route as planned, especially in the middle section where they cross an area 1–1.5 km to the north. The consideration of community severance is decisive in this shift, as the routes cross through vacant land, reducing the separation of communities. In addition, in the

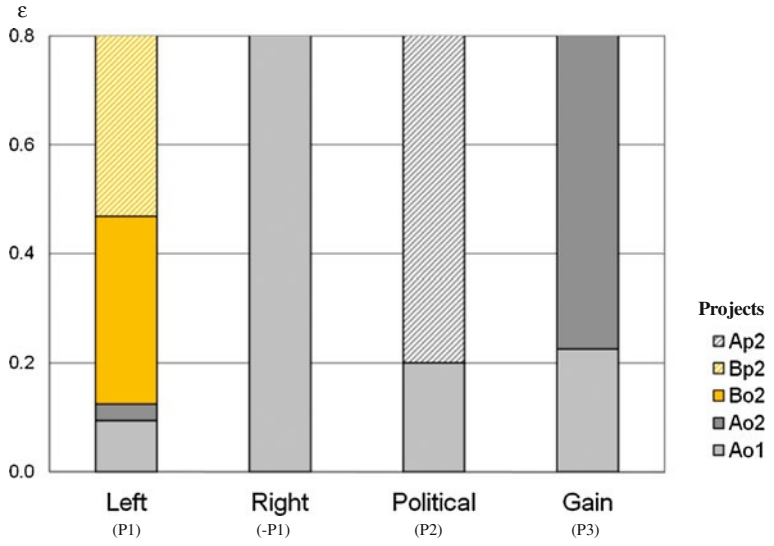


Fig. 2 Politically optimal routes: Project locus

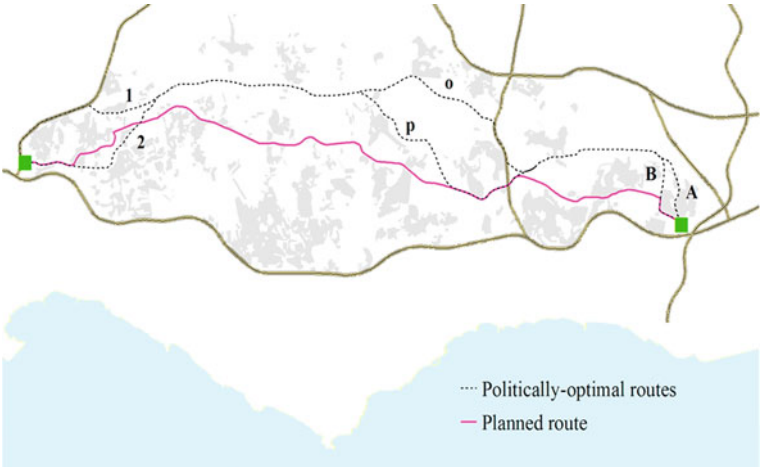


Fig. 3 Politically optimal routes and the planned route for the *Via Longitudinal*

eastern end of the road, the routes overpass one town, while the planned route is an upgrade of an existing arterial road crossing the town.

The figures shows that the consideration of political interest produces the same project as the socially optimum when the weights are placed on the right-wing factor [-P1], regardless of the value of the parameter ε. On the other hand, the project is sensitive to the size of the weights placed on the left-wing factor [P1]. As we consider higher values for the parameter ε, the route first changes in the western end (from *1* to *2*), then in the eastern end (from *A* to *B*) and finally in the middle section (from *o* to *p*). The first of these changes also occurs when we place weights on the factor measuring potential electoral gains [P3], but for higher values of the distributional parameter. The first and third shifts also occur when we place weights on the factor measuring the degree of attachment to politics [P2], but without a shift in the eastern end of the road.

The analysis that follows focuses on the effects of weights placed on the P1 and P2 factors. The effects of the P3 factor are understood as a particular case of the effects of the P1 factor, as the route obtained for P3 is also obtained in an interval of values for the parameters defining the weights placed on P1.

4.2 Efficiency

Table 3 gives statistics of the routes for several elements of cost. The comparison of the statistics of the politically-optimal routes and the socially-optimal route (Ao1) provides insights on the efficiency loss associated with the former. The length of the road is an approximation to the financial costs of the project. Length is also multiplied by a factor depending on slopes, to account for differences in

Table 3 Length, community severance and land use of politically-optimal routes

	Length (km)	Length x slope	Severance (000)	Land use (%)		
				Built-up	Agricultural	Ecological
Ao1	20.19	37.3	6.7	8.4	2.3	9.6
Ao2	20.17	37.9	7.4	14.6	3.7	16.3
Bo2	20.24	37.7	8.5	15.9	3.7	16.3
Ap2	20.40	38.6	7.7	17.4	14.2	15.2
Bp2	20.47	38.4	8.8	18.7	14.1	15.2
Plan	18.7	33.6	16.4	25.4	16.7	24.2

relief in the region. Aggregate severance costs are the sum of changes in the severance indicator across all districts weighted by their population. The percentage of the road length on built-up land (residential and non-residential) is an indicator of local costs such as exposure to pollution of local residents and workers and also an indicator of construction cost savings, given the constraint placed on the crossing of built-up areas, which implies the reformulation of existing roads in detriment of the more expensive option of building a new road. The last two columns in the table give the percentage of road length on land mapped in municipal master plans as areas with agricultural or environmental potential. These two indicators assess the economic and ecological opportunity costs of the new road.

All socially and politically optimal routes have very similar lengths. The socially-optimal route (Ao1) has the lower aggregate severance effects (by definition) but also the lowest proportion of road length crossing all three types of land use. Incremental departures from the socially-optimum towards political-biased distributions (following the patterns in Fig. 2) are associated with longer lengths, greater aggregate severance effects and greater land use costs. The cost in terms of agricultural land is especially noticeable in the projects that use the south middle section (Ap2 and Bp2). The results also suggest that the planned route considers financial cost as the main factor, as this route has the shortest length of all routes and a higher proportion of length in built-up areas. However, this route has the worst indices of all routes in terms of severance effects and negative effects associated with the land use on the areas crossed.

4.3 Distribution

This section presents the distributional impacts of the politically-optimal routes. In a first stage, we study the average severance impact of the routes across their whole length, based on the characteristics of the population affected. In a second stage we look into the profile of the impacts, taking into account the different levels of severance associated with the route. The analysis focuses on a set of 9 variables, including the three political factors and the five socio-economical factors described

Table 4 Averages of political factors, social factor and time gains, weighted by severance effects

	Political factors			Social factors					Time gains
	P1 (Left)	P2 (Political)	P3 (Potent.)	S1 (Age)	S2 (Qualif.)	S3 (Urban)	S4 (Slums)	S5 (Migr.)	
Ao1	0.45	0.03	−0.30	−0.25	−0.49	0.41	0.40	−0.30	0.13
Ao2	0.42	0.07	−0.32	−0.16	−0.45	0.43	0.35	−0.24	0.10
Bo2	0.31	0.24	−0.30	−0.14	−0.28	0.55	0.25	−0.19	0.13
Ap2	0.36	−0.05	−0.23	−0.12	−0.43	0.27	0.33	−0.25	0.04
Bp2	0.26	0.13	−0.22	−0.10	−0.26	0.41	0.23	−0.20	0.08
Plan	−0.01	0.39	−0.20	−0.03	0.10	0.63	0.12	−0.15	−0.13

in the previous section. The ninth variable is an indicator of the benefits that the new road will bring to the population in each district in terms of shorter times to work. This variable is obtained by modelling time to work in the pre- and post-policy scenario, considering data on commuting flows from the census and commuting surveys, which were then disaggregated to give peak and off-peak flows from each enumeration district to a large set of destinations in the Lisbon Metropolitan Area. The estimation of times to work makes use of GIS models of the private and public transport networks and includes the effect of congestion.

Table 4 and Fig. 4 show the distributional impacts of the routes, estimated as the average of several variables weighted by population and the severance effect. The charts illustrate the information in the table according to the sequence of projects obtained when increasing the parameter ε in the specification of the weights placed on the P1 and P2 political factors. The values of all variables are standardized in relation to their mean and standard deviation in the two municipalities involved in the project.

The socially-optimal project (Ao1) is associated with above-average values of the Left (P1), Urbanization (S3) and Slums (S4) factors and below-average values of the electoral potential (P3), Age (S1), Qualifications (S2) and Migration (F5). In other words, a project that minimizes aggregate severance effects has a higher effect on populations that are younger, less-qualified, live in more urbanized areas or in slums, live in the present area for a relatively long time and tend to vote in left-wing parties or to abstain from voting. The severance-weighted average of time gains is also above the regional average, which suggests that the socially-optimal route tends to distribute costs according to benefits.

As we consider higher weights on the P1 (Left) factor, the severance-weighted average of that factor decreases from 0.45 (project Ao1) to 0.26 (project Bp2). However, even in project Bp2, the value is still above the average of the two municipalities. The increase of the distributional weight on P1 does not have a linear relationship with the average of P2 (Political) and only a limited effect on the average of P3 (Potential). However, increasing the weights on P1 decreases the disadvantage of low-qualified individuals and populations living in slums. There is also a slight reduction of the disadvantage of younger individuals.

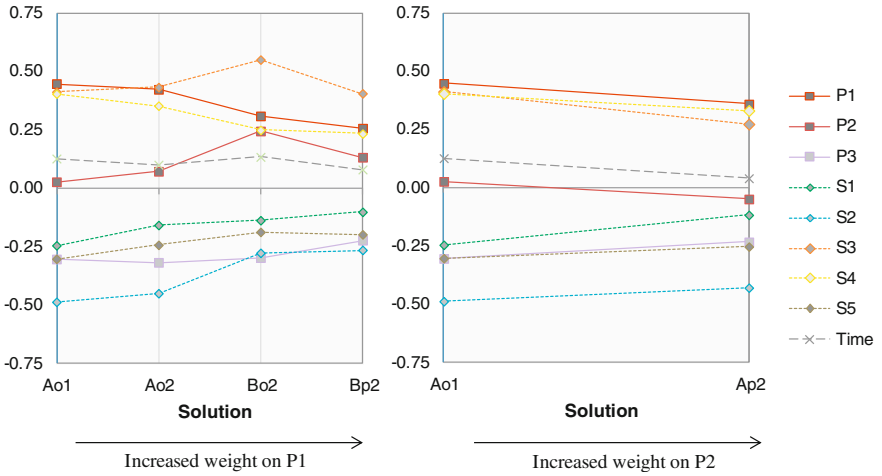


Fig. 4 Averages of political factors, social factors and time gains for increased weight values

Although the consideration of higher weights on the P2 (Political) factor is associated with a change in the optimal route (from Ao1 to Ap2), the effect of that change in the severance-weighted average of that factor is very small (from 0.03 to −0.05). However, the change is linked to patterns of redistribution among social groups that are similar to the previous case, reducing the disadvantages of regions with slum areas and low-qualified and young populations. The change in routes also brings the severance-weighted average of the time gains close to zero, which means that severance costs are distributed more equally among populations deriving different benefits from the road, comparing with the socially optimal route.

The alignment of the planned route is independent of the Left versus Right positioning of the population affected by the project, but affects to a higher degree populations with higher degrees of political mobilization and areas with lower potential for the policy-maker to attract new votes. The socio-economic distributional impacts of planned route are in general lower than the politically-optimal routes, with the exception of the impact falling on urbanized areas (S3).

Figure 5 illustrates with more detail the distribution of the severance effects of the different routes according to the socio-economic characteristics of the population, focusing on the Age and Qualifications factors. The charts plot the average value of the factors for each value of the severance effect. The left and right charts include respectively the projects associated with different values for the weights placed on the P1 and P2 factors.

In the case of the distribution of costs according to age groups, the results show that the effects of the projects associated with higher weights on either political factor is to move the distributional curve closer to the origin, especially in the case of the districts with the highest severance effects (above 0.35). The effects of project Ap2 (obtained placing weights on the P2 factor) are similar to those of project Ao2 (obtained placing weights on P1), showing that attending to different

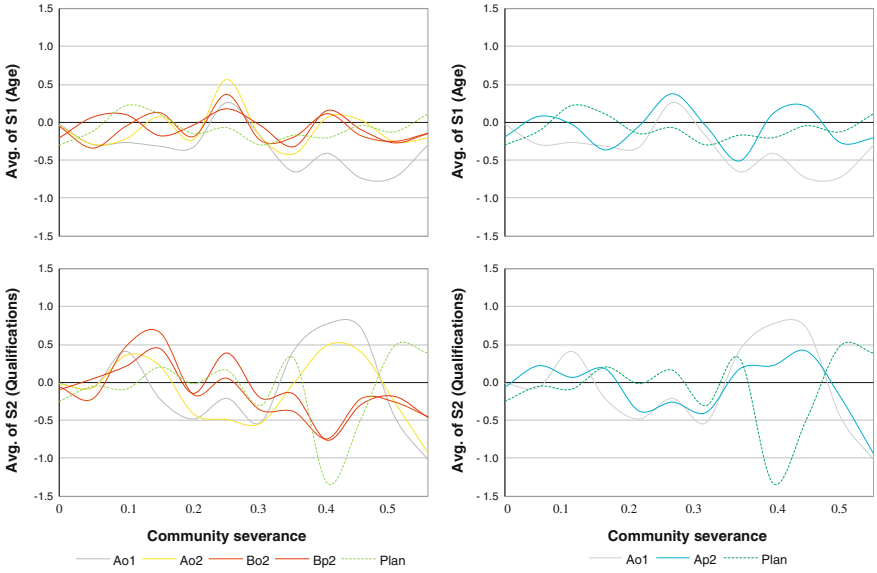


Fig. 5 Severance effects of politically optimal projects: Socio-economic profile

political interests produces the same distributional result. The charts also confirm that the planned route is distributionally neutral in terms of age groups.

In the case of the distribution of costs according to qualification groups, the figure shows that the average values presented in Table 4 mask some of the variation in the distribution of impacts of different degrees. The most important differences between projects regard the intervals with the highest impacts (above 0.35). This interval corresponds to the impacts on the neighbourhoods in the densely-populated eastern end of the route. Options *A* in this end redistribute the impacts among qualification groups differently from options *B* and the planned route. An increase in the weights of the P2 factor or a slight increase in the weights of the P1 factor has therefore a small effect on the qualifications profile of the route, as these increases do not imply a shift from "A" to "B" projects. However, a higher increase of the weights of the P1 factor lead to substantial changes, due to the shift to "B" projects.

The results of this section show that the introduction of political bias in the planning of a new road may contribute to a redistribution of the severance costs of the road among different groups in society, comparing with the socially-optimal route. This is explained in part because the political characteristics of the population are related to social factors, as shown in Table 2. However, the redistributive effect is observed equally for social factors with strong and weak relationships with political factors. For example, while the Left factor is related strongly with the Qualifications factor and only mildly with the Age, Slums and Migration factors, the redistribution impacts of the projects based on weights on the Left factor is observed for all the four social factors.

On the other hand, a given socio-economic distributional impact may be the result of different political strategies of the policy-makers. In this case, even though the factors measuring left wing voters, levels of political participation and political gain are largely independent by definition (as they were obtained by factor analysis), they lead to similar patterns in the redistribution of severance across different groups in society.

5 Conclusions and Further Work

This chapter examined the influence of political factors in the design of route alignments for a new road in an urban area, focusing on the little researched aspect of community severance impacts. The analysis used GIS methods to derive alignments based on the distribution of those impacts across populations with different political characteristics.

The findings add evidence to the study of trade-offs between efficiency and equity in public policy, by providing insights on the role of political factors in the choices over those trade-offs. The analysis showed that the main effect of political bias is at the level of efficiency, as higher levels of political bias are linked to increased deviations from the efficient route, not only in terms of aggregate severance effects but also in terms of land use costs. On the other hand, political bias lead to a more equal distribution of the costs of the project comparing with the efficient route.

The chapter also provides a basis for further exploration of the policy maker's motivations. The process leading to decisions on transport infrastructure is characterized by interactions between different levels of policy-making. Protests to decisions of national governments also tend to be lead by local leaders and as such the assessment of the political characteristics of local populations could include information on whether central and local governments are hold by different parties. The distribution of the time gains brought by the new road to different neighbourhoods, which were in this chapter treated as an indicator of the effects of politically-biased policies, may also be regarded as a factor influencing political decisions, due to its influence on voting behaviour (Hårsman and Quigley 2010).

The analysis of political influences on infrastructure planning tends to be constrained by the spatial scale at which electoral data is released, as electoral districts are usually larger than the scale of community severance and other environmental effects. In this chapter we assumed that the relationships of political and socio-economic variables found at the level of the civil parish also apply at a smaller scale. However, civil parish borders do not necessarily correspond to discontinuities in the spatial distribution of different types of voters. In addition, while we used socio-economic data to predict electoral behaviour in small areal units, aspects such as the communities' political power may also be contained within the residuals of the models—what Hamilton (1995) calls the "pure politics"

factor. Further developments in the estimation of spatially disaggregated distributions of political variables can therefore improve the analysis of political bias in transport policies with effects across different neighbourhoods.

Acknowledgments This research was supported by the Portuguese Foundation for Science and Technology. Thanks go to Giles Atkinson, Steve Gibbons, Helena Titheridge and Andrew Lovett for their comments.

References

Boschken HL (1998) Upper-middle-class influence on developmental policy outcomes: the case of transit infrastructure. Urban Stud 35(4):627–647

BrentRJ (1976) The Minister of Transport's social welfare function: a study of the factors behind railway closure decisions (1963–1970). PhD thesis, University of Manchester

Brooks N, Sethi R (1997) The distribution of pollution: community characteristics and exposure to air toxics. J Environ Econ Manag 32(2):233–250

Burstein P, Linton A (2002) The impact of political parties, interest groups, and social movement organizations on public policy: some recent evidence and theoretical concerns. Soc Forces 81(2):380–408

Camacho DE (1998) The environmental justice movement: a political framework. In: Camacho DE (ed) Environmental injustices, political struggles—race, class and the environment. Duke University Press, Durham, pp 11–30

Castells A, Solé-Ollé A (2005) The regional allocation of infrastructure investment: the role of equity, efficiency and political factors. Eur Econo Rev 49(5):1165–1205

Congleton RD, Bennett RW (1995) On the political economy of state highway expenditures: Some evidence of the relative performance of alternative public choice models. Public Choice 84(1/2):1–24

Earnhart D (2004) The effects of community characteristics on polluter compliance levels. Land Econ 80(3):408–432

Hamilton JT (1993) Politics and social costs: estimating the impact of collective action on hazardous waste facilities. Rand J Econ 24(1):101–125

Hamilton JT (1995) Testing for environmental racism: prejudice, profits, political power? J Policy Anal Manage 14(1):107–132

Hårsman B, Quigley JM (2010) Political and public acceptability of congestion pricing: ideology and self-interest. J Policy Anal Manage 29(4):854–874

Hillman M (1997) Public policy on the green modes. In: Tolley R (ed) The greening of urban transport, Chapter 6, 2nd edn. Wiley, Chichester, pp 71–79

Jussila Hammes J (2012) The political economy of infrastructure planning in Sweden: supporting analyses. Centre for Transport Studies Stockholm—Working papers in Transport Economics 2012, vol 21

Schade J, Schlag B (eds) (2003) Acceptability of transport pricing strategies. Elsevier, Oxford

Taylor BD, Kim EJ, Gahbauer JE (2010) The thin red line—a case study of political influence on transportation planning practice. J Plann Educ Res Dec 29(2):173–193

Walden ML, Eryuruk G (2012) Determinants of local highway spending in North Carolina. Growth Change 43(3):462–481

Errata to: Geographic Information Science at the Heart of Europe

Danny Vandenbroucke, Bénédicte Bucher and Joep Crompvoets

Errata to: D. Vandenbroucke et al. (eds.), *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography, DOI 10.1007/978-3-319-00615-4

In the chapter entitled "What You See is What You Map: Geometry-Preserving Micro-Mapping for Smaller Geographic Objects with MAPIT", the last sentence of the abstract reads "Please check and confirm the author names and initials are correct" should be dropped.

In this book, the name of the author is Chunyuan Cai (with Cai being the surname) instead of Cai Cai (which was the name in the START system wrongly filled by the main author)—it should read as Chunyuan Cai in three places "Table of contents, p. xi", "list of contributors, p. xv" and in the chapter entitled "What You See is What You Map: Geometry-Preserving Micro-Mapping for Smaller Geographic Objects with MAPIT", p. 3).

The online version of the original book can be found at 10.1007/978-3-319-00615-4.

D. Vandenbroucke (✉)
Spatial Applications Division, Katholieke Universiteit Leuven, Heverlee, Vlaams-Brabant, Belgium

B. Bucher
COGIT, Institut Geographique National France, Saint-Mandé Cedex, Paris, France

J. Crompvoets
Public Management Institute, Katholieke Universiteit Leuven, Leuven, Vlaams-Brabant, Belgium

One name of a co-author (Lukas Loos) was dropped due to wrong indications from the editors side. The order of the author should read as Mohamed Bakillah, Alexander Zipf, Steve H. L. Liang and Lukas Loos in the chapter entitled "Publish/subscribe System Based on Event Calculus to Support Real-Time Multi-Agent Evacuation Simulation", p. 337. Lukas Loos should be added in the "Table of contents, p. xiii", "list of Contributors, p. xvi" and in this chapter.