

10th International Conference on Geographic Information Science

GIScience 2018, August 28–31, 2018, Melbourne, Australia

Edited by

Stephan Winter

Monika Sester

Amy L. Griffin



Editors

Stephan Winter Infrastructure Engineering The University of Melbourne Australia winter@unimelb.edu.au	Monika Sester Cartography and Geoinformation University of Hannover Germany monika.sester@ikg.uni-hannover.de
---	---

Amy L. Griffin
Geospatial Sciences
RMIT University
Australia
amy.griffin@rmit.edu.au

ACM Classification 2012

Information systems → Location based services, Information systems → Geographic information systems,
Information systems → Personalization

ISBN 978-3-95977-083-5

Published online and open access by

Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Saarbrücken/Wadern,
Germany. Online available at <http://www.dagstuhl.de/dagpub/978-3-95977-083-5>.

Publication date

August, 2018

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed
bibliographic data are available in the Internet at <http://dnb.d-nb.de>.

License

This work is licensed under a Creative Commons Attribution 3.0 Unported license (CC-BY 3.0):
<http://creativecommons.org/licenses/by/3.0/legalcode>.



In brief, this license authorizes each and everybody to share (to copy, distribute and transmit) the work
under the following conditions, without impairing or restricting the authors' moral rights:

- Attribution: The work must be attributed to its authors.

The copyright is retained by the corresponding authors.

Digital Object Identifier: 10.4230/LIPICs.GIScience.2018.0

ISBN 978-3-95977-083-5

ISSN 1868-8969

<http://www.dagstuhl.de/lipics>

LIPICs – Leibniz International Proceedings in Informatics

LIPICs is a series of high-quality conference proceedings across all fields in informatics. LIPICs volumes are published according to the principle of Open Access, i.e., they are available online and free of charge.

Editorial Board

- Luca Aceto (*Chair*, Gran Sasso Science Institute and Reykjavik University)
- Susanne Albers (TU München)
- Christel Baier (TU Dresden)
- Javier Esparza (TU München)
- Michael Mitzenmacher (Harvard University)
- Madhavan Mukund (Chennai Mathematical Institute)
- Anca Muscholl (University Bordeaux)
- Catuscia Palamidessi (INRIA)
- Raimund Seidel (Saarland University and Schloss Dagstuhl – Leibniz-Zentrum für Informatik)
- Thomas Schwentick (TU Dortmund)
- Reinhard Wilhelm (Saarland University)

ISSN 1868-8969

<http://www.dagstuhl.de/lipics>

■ Contents

Preface	
<i>Stephan Winter, Monika Sester, and Amy Griffin</i>	xi:xii
Full Papers	
Early Detection of Herding Behaviour during Emergency Evacuations	
<i>David Amores, Maria Vasardani, and Egemen Tanin</i>	1:1–1:15
What Makes Spatial Data Big? A Discussion on How to Partition Spatial Data	
<i>Alberto Belussi, Damiano Carra, Sara Migliorini, Mauro Negri, and Giuseppe Pelagatti</i>	2:1–2:15
Intersections of Our World	
<i>Paolo Fogliaroni, Dominik Bucher, Nikola Jankovic, and Ioannis Giannopoulos</i> ..	3:1–3:15
Considerations of Graphical Proximity and Geographical Nearness	
<i>Francis Harvey</i>	4:1–4:18
An Empirical Study on the Names of Points of Interest and Their Changes with Geographic Distance	
<i>Yingjie Hu and Krzysztof Janowicz</i>	5:1–5:15
Outlier Detection and Comparison of Origin-Destination Flows Using Data Depth	
<i>Myeong-Hun Jeong, Junjun Yin, and Shaowen Wang</i>	6:1–6:14
Is Saliency Robust? A Heterogeneity Analysis of Survey Ratings	
<i>Markus Kattenbeck, Eva Nuhn, and Sabine Timpf</i>	7:1–7:16
Labeling Points of Interest in Dynamic Maps using Disk Labels	
<i>Filip Krumpal</i>	8:1–8:14
Improving Discovery of Open Civic Data	
<i>Sara Lafia, Andrew Turner, and Werner Kuhn</i>	9:1–9:15
Local Co-location Pattern Detection: A Summary of Results	
<i>Yan Li and Shashi Shekhar</i>	10:1–10:15
Detection and Localization of Traffic Signals with GPS Floating Car Data and Random Forest	
<i>Yann Méneroux, Hiroshi Kanasugi, Guillaume Saint Pierre, Arnaud Le Guilcher, Sébastien Mustière, Ryosuke Shibasaki, and Yugo Kato</i>	11:1–11:15
Heterogeneous Skeleton for Summarizing Continuously Distributed Demand in a Region	
<i>Alan T. Murray, Xin Feng, and Ali Shokoufandeh</i>	12:1–12:11
A Network Flow Model for the Analysis of Green Spaces in Urban Areas	
<i>Benjamin Niedermann, Johannes Oehrlein, Sven Lautenbach, and Jan-Henrik Haurert</i>	13:1–13:16
Continuous Obstructed Detour Queries	
<i>Rudra Ranajee Saha, Tanzima Hashem, Tasmia Shahriar, and Lars Kulik</i>	14:1–14:16

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Monika Sester, and Amy L. Griffin

Leibniz International Proceedings in Informatics



Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



Enhanced Multi Criteria Decision Analysis for Planning Power Transmission Lines <i>Joram Schito, Ulrike Wissen Hayek, and Martin Raubal</i>	15:1–15:16
FUTURES-AMR: Towards an Adaptive Mesh Refinement Framework for Geosimulations <i>Ashwin Shashidharan, Ranga Raju Vatsavai, Derek B. Van Berkel, and Ross K. Meentemeyer</i>	16:1–16:15
xNet+SC: Classifying Places Based on Images by Incorporating Spatial Contexts <i>Bo Yan, Krzysztof Janowicz, Gengchen Mai, and Rui Zhu</i>	17:1–17:15

Short Papers

A Critical Look at Cryptogovernance of the Real World: Challenges for Spatial Representation and Uncertainty on the Blockchain <i>Benjamin Adams and Martin Tomko</i>	18:1–18:6
Towards Optimal Deployment of a Sensor Network in a 3D Indoor Environment for the Mobility of People with Disabilities <i>Ali Afghantoloe and Mir Abolfazl Mostafavi</i>	19:1–19:6
Challenges in Creating an Annotated Set of Geospatial Natural Language Descriptions <i>Niloofer Aflaki, Shaun Russell, and Kristin Stock</i>	20:1–20:6
Improved and More Complete Conceptual Model for the Revision of IndoorGML <i>Abdullah Alattas, Sisi Zlatanova, Peter van Oosterom, and Ki-Joune Li</i>	21:1–21:12
Design for Geospatially Enabled Climate Modeling and Alert System (CLIMSYS): A Position Paper <i>Devanjan Bhattacharya and Marco Painho</i>	22:1–22:6
Geographical Exploration and Analysis Extended to Textual Content <i>Raphaël Ceré, Mattia Egloff, and François Bavaud</i>	23:1–23:7
Evaluating Efficiency of Spatial Analysis in Cloud Computing Platforms <i>Changlock Choi, Yelin Kim, Youngho Lee, and Seong-Yun Hong</i>	24:1–24:5
Towards the Usefulness of User-Generated Content to Understand Traffic Events <i>Rahul Deb Das and Ross S. Purves</i>	25:1–25:7
Unfolding Urban Structures: Towards Route Prediction and Automated City Modeling <i>Paolo Fogliaroni, Marvin Mc Cutchan, Gerhard Navratil, and Ioannis Giannopoulos</i>	26:1–26:6
Deconstructed and Inverted Multi-Criteria Evaluation for On-The-Fly Scenario Development and Decision-Making <i>Martin Geilhausen and Patrick Laube</i>	27:1–27:7
Space-Time Representation of Accessible Areas for Wheelchair Users in Urban Areas <i>Amin Gharebaghi and Mir Abolfazl Mostafavi</i>	28:1–28:6

Spatial Periodicity Analysis of Urban Elements Application to the Ancient City of Amida <i>Jean-François Girres, Martine Assenat, Robin Ralite, and Ester Ribo-Delissey ...</i>	29:1–29:6
Gaze Sequences and Map Task Complexity <i>Fabian Göbel, Peter Kiefer, Ioannis Giannopoulos, and Martin Raubal</i>	30:1–30:6
Facilitating the Interoperable Use of Cross-Domain Statistical Data Based on Standardized Identifiers <i>Jung-Hong Hong and Jing-Cen Yang</i>	31:1–31:7
Identification of Geographical Segmentation of the Rental Apartment Market in the Tokyo Metropolitan Area <i>Ryo Inoue, Rihoko Ishiyama, and Ayako Sugiura</i>	32:1–32:6
Automatic Wall Detection and Building Topology and Property of 2D Floor Plan <i>Hanme Jang, Jong Hyeon Yang, and Yu Kiyun</i>	33:1–33:5
Mapping Wildlife Species Distribution With Social Media: Augmenting Text Classification With Species Names <i>Shelan S. Jeawak, Christopher B. Jones, and Steven Schockaert</i>	34:1–34:6
Multimodal-Transport Collaborative Evacuation Strategies for Urban Serious Emergency Incidents Based on Multi-Sources Spatiotemporal Data <i>Jincheng Jiang, Yang Yue, and Shuai He</i>	35:1–35:8
A New Map Symbol Design Method for Real-Time Visualization of Geo-Sensor Data <i>Donglai Jiao and Jintao Sun</i>	36:1–36:6
How Do Texture and Color Communicate Uncertainty in Climate Change Map Displays? <i>Irene M. Johannsen, Sara Irina Fabrikant, and Mariele Evers</i>	37:1–37:6
An Analytical Framework for Understanding Urban Functionality from Human Activities <i>Chaogui Kang and Yu Liu</i>	38:1–38:8
Application of Style Transfer in the Vectorization Process of Floorplans <i>Seongyong Kim, Seula Park, and Kiyun Yu</i>	39:1–39:6
Estimating Building Age from Google Street View Images Using Deep Learning <i>Yan Li, Yiqun Chen, Abbas Rajabifard, Kouros Khoshelham, and Mitko Aleksandrov</i>	40:1–40:7
Center Point of Simple Area Feature Based on Triangulation Skeleton Graph <i>Wei Lu and Tinghua Ai</i>	41:1–41:6
The Use of Particle Swarm Optimization for a Vector Cellular Automata Model of Land Use Change <i>Yi Lu and Shawn Laffan</i>	42:1–42:6
Towards a Comprehensive Temporal Classification of Footfall Patterns in the Cities of Great Britain <i>Karlo Lugomer and Paul Longley</i>	43:1–43:6

Is This Statement About A Place? Comparing two perspectives <i>Alan M. MacEachren, Richard Caneba, Hanzhou Chen, Harrison Cole, Emily Domanico, Nicholas Triozzi, Fangcao Xu, and Liping Yang</i>	44:1–44:6
Geospatial Semantics for Spatial Prediction <i>Marvin Mc Cutchan and Ioannis Giannopoulos</i>	45:1–45:6
Docked vs. Dockless Bike-sharing: Contrasting Spatiotemporal Patterns <i>Grant McKenzie</i>	46:1–46:7
OpenPOI: An Open Place of Interest Platform <i>Grant McKenzie and Krzysztof Janowicz</i>	47:1–47:6
Exploring Shifting Densities through a Movement-based Cartographic Interface <i>Aline Menin, Sonia Chardonnel, Paule-Annick Davoine, and Luciana Nedel</i>	48:1–48:6
Geotagging Location Information Extracted from Unstructured Data <i>Kyunghyun Min, Jungseok Lee, Kiyun Yu, and Jiyoung Kim</i>	49:1–49:6
Linked Open Data Vocabularies for Semantically Annotated Repositories of Data Quality Measures <i>Franz-Benjamin Mocnik</i>	50:1–50:7
Need A Boost? A Comparison of Traditional Commuting Models with the XGBoost Model for Predicting Commuting Flows <i>April Morton, Jesse Piburn, and Nicholas Nagle</i>	51:1–51:7
Modeling Road Traffic Takes Time <i>Kamaldeep Singh Oberoi, Géraldine Del Mondo, Yohan Dupuis, and Pascal Vasseur</i>	52:1–52:7
Diversity in Spatial Language Within Communities: The Interplay of Culture, Language and Landscape in Representations of Space <i>Bill Palmer, Alice Gaby, Jonathon Lum, and Jonathan Schlossberg</i>	53:1–53:8
Flexible Patterns of Place for Function-based Search of Space <i>Emmanuel Papadakis, Andreas Petutschnig, and Thomas Blaschke</i>	54:1–54:7
Novel Models for Multi-Scale Spatial and Temporal Analyses <i>Yi Qiang, Barbara P. Battenfield, Nina Lam, and Nico Van de Weghe</i>	55:1–55:7
Geosocial Media Data as Predictors in a GWR Application to Forecast Crime Hotspots <i>Alina Ristea, Ourania Kounadi, and Michael Leitner</i>	56:1–56:7
Who Masks? Correlates of Individual Location-Masking Behavior in an Online Survey <i>Dara E. Seidl and Piotr Jankowski</i>	57:1–57:6
Dynamically-Spaced Geo-Grid Segmentation for Weighted Point Sampling on a Polygon Map Layer <i>Kelly Sims, Gautam Thakur, Kevin Sparks, Marie Urban, Amy Rose, and Robert Stewart</i>	58:1–58:7

The Landform Reference Ontology (LFRO): A Foundation for Exploring Linguistic and Geospatial Conceptualization of Landforms <i>Gaurav Sinha, Samantha T. Arundel, Torsten Hahmann, E. Lynn Usery, Kathleen Stewart, and David M. Mark</i>	59:1–59:7
Abstract Data Types for Spatio-Temporal Remote Sensing Analysis <i>Martin Sudmanns, Stefan Lang, Dirk Tiede, Christian Werner, Hannah Augustin, and Andrea Baraldi</i>	60:1–60:7
Towards Vandalism Detection in OpenStreetMap Through a Data Driven Approach <i>Quy Thy Truong, Guillaume Touya, and Cyril de Runz</i>	61:1–61:7
A Conceptual Framework for Representation of Location-based Social Media Activities <i>Xuebin Wei and Xiaobai Angela Yao</i>	62:1–62:7
Towards the Statistical Analysis and Visualization of Places <i>René Westerholt, Mathias Gröbe, Alexander Zipf, and Dirk Burghardt</i>	63:1–63:7
An Experimental Comparison of Two Definitions for Groups of Moving Entities <i>Lionov Wiratma, Maarten Löffler, and Frank Staals</i>	64:1–64:6
Extracting Geospatial Information from Social Media Data for Hazard Mitigation, Typhoon Hato as Case Study <i>Jibo Xie, Tengfei Yang, and Guoqing Li</i>	65:1–65:6
Propagation of Uncertainty for Volunteered Geographic Information in Machine Learning <i>Jin Xing and Renee E. Sieber</i>	66:1–66:6
Satellite Image Spoofing: Creating Remote Sensing Dataset with Generative Adversarial Networks <i>Chunxue Xu and Bo Zhao</i>	67:1–67:6
A Safety Evaluation Method of Evacuation Routes in Urban Areas in Case of Earthquake Disasters Using Ant Colony Optimization and Geographic Information Systems <i>Kayoko Yamamoto and Ximing Li</i>	68:1–68:7
Analysis of Irregular Spatial Data with Machine Learning: Classification of Building Patterns with a Graph Convolutional Neural Network <i>Xiongfeng Yan and Tinghua Ai</i>	69:1–69:7
Assessing Neighborhood Conditions using Geographic Object-Based Image Analysis and Spatial Analysis <i>Chi-Feng Yen, Ming-Hsiang Tsou, and Chris Allen</i>	70:1–70:7
Spatial Information Extraction from Text Using Spatio-Ontological Reasoning <i>Madiha Yousaf and Diedrich Wolter</i>	71:1–71:6
Scalable Spatial Join for WFS Clients <i>Tian Zhao, Chuanrong Zhang, and Zhijie Zhang</i>	72:1–72:6
Modelling Spatial Patterns Using Graph Convolutional Networks <i>Di Zhu and Yu Liu</i>	73:1–73:7

■ Preface

The Tenth International Conference on Geographic Information Science, GIScience, was held in Melbourne, Australia, 28–31 August 2018. Hosted at RMIT University in collaboration with the University of Melbourne, GIScience 2018, the flagship conference in the field of geographic information science continued the highly successful conference series, which started in 2000.

The conference regularly brings together more than 300 international participants from academia, industry, and government to discuss and advance the state-of-the-art in geographic information science. August 28, 2018 was dedicated to Workshops and Tutorials. The main conference took place from August 29 to 31, 2018, and consisted of two refereed paper tracks: Full papers and short papers/extended abstracts. The latter track allowed authors to choose between published short papers and unpublished extended abstracts; both types were treated equally in the review process, and both were presented orally at the conference. This volume of proceedings contains only the full papers and the short papers.

For GIScience 2018, we received 46 full paper submissions. Each full paper was read by at least three members of the international program committee. 17 of the full paper submissions made it through the selection process for presentation at the conference and publication in this volume (37%). We also received a total of 113 submissions in the short papers/extended abstracts track. These short papers and extended abstracts were reviewed by at least two members of same program committee. We accepted a total of 89 short papers and extended abstracts for presentation at the conference (75%). The authors of 56 of these selected submissions chose the option of having a short paper, which are published in the second part this volume.

Bringing the GIScience conference to Australia was also coupled with the hope of attracting more participants from the Asia-Pacific region, which was fulfilled: This year's conference saw strong participation from scientists based in Asia, with 22% of accepted presentations including at least one author who was affiliated with an institution in Asia – a significant increase compared to prior events in this series. Of course the decision to bring the conference to Australia came also with a price tag for the European and North American scientific community. The chairs are thankful for the effort this community made.

The GIScience conference series has always had a focus on fundamental research themes and questions. Papers advancing the field methodologically or theoretically are encouraged; papers strictly dealing with applications are discouraged. GIScience 2018 welcomed papers and proposals covering emerging topics and fundamental research findings across all sectors of geographic information science, including (but not limited to) the role of spatial information in geography, computer science, engineering, information science, linguistics, mathematics, cognitive science, philosophy, psychology, social science, and geostatistics. In GIScience 2018 a number of papers have been related to the emerging topic of (Deep) Learning.

GIScience is a community-run conference series, backed by a steering committee. The organizers of GIScience 2018 are grateful for the trust and advice by the committee. A number of people have contributed to the organization of this conference, many of them in chairing roles or on the program committees of the conference or its satellite events. This year's GIScience program committee had a record number of 127 volunteers from all over the world. We would also like to take the opportunity to thank all those people usually unnamed on websites and in proceedings: the staff managing registrations, the student volunteers helping with a smooth event, and the proceedings editor, Subhrashanka Dey. The true



makers of a conference are, of course, the authors and participants of the conference.

From 2018 onwards, GIScience conference proceedings will be published in LIPIcs, the Leibniz International Proceedings in Informatics series. LIPIcs volumes are peer-reviewed and published according to the principle of open access, i.e., they are available online and free of charge. Each article is published under a Creative Commons CC BY license (<http://creativecommons.org/licenses/by/3.0/>), where the authors retain their copyright. Also, each article is assigned a DOI and a URN. The digital archiving of each volume is done in cooperation with the Deutsche Nationalbibliothek/German National Library. We hope that this more community-spirited format helps further with the growth and impact of GIScience.

Stephan Winter, Monika Sester, and Amy Griffin
Program Committee Chairs, GIScience 2018

■ GIScience 2018 Program Committee

- Ola Ahlqvist, Ohio State University, USA
Jared Aldstadt, SUNY Buffalo, USA
Natalia Andrienko, Fraunhofer Institute, Germany
Jagannath Aryal, University of Tasmania, Australia
Andrea Ballatore, University of London, UK
Kate Beard, University of Maine, USA
Scott Bell, University of Saskatchewan, Canada
Itzhak Benenson, Tel Aviv University, Israel
David Bennett, University of Iowa, USA
Luke Bergmann, University of Washington, USA
Michela Bertolotto, University College Dublin, Ireland
Ling Bian, SUNY Buffalo, USA
Susanne Bleisch, University of Applied Sciences of Northwestern Switzerland, Switzerland
Boyan Brodaric, Geological Survey of Canada, Canada
Dan Brown, University of Washington, USA
Chris Brunsdon, National University of Ireland, Ireland
B n dicte Bucher, IGN, France
Maike Buchin, Ruhr Universit t Bochum, Germany
Barbara Buttenfield, University of Colorado at Boulder, USA
Jonathan Cinnamon, University of Exeter, UK
Christophe Claramunt, Naval Academy Research Institute, France
Eliseo Clementini, University of L'Aquila, Italy
Arzu Coltekin, University of Zurich, Switzerland
Tom Cova, University of Utah, USA
Clodoveu Davis, Universidade Federal de Minas Gerais, Brazil
Sytze de Bruin, Wageningen University, The Netherlands
Leila De Florian, University of Maryland, USA
Eric Delmelle, University of North Carolina at Charlotte, USA
Urska Demsar, University of St Andrews, UK
Somayeh Dodge, University of Minnesota, USA
Joni Downs, University of South Florida, USA
Sara Irina Fabrikant, University of Zurich, Switzerland
Carson Farmer, University of Colorado at Boulder, USA
Paolo Fogliaroni, Vienna University of Technology, Austria
Mark Gahegan, University of Auckland, New Zealand
Ioannis Giannopoulos, Vienna University of Technology, Austria
Daniel Goldberg, Texas A&M University, USA
Tony Grubestic, Arizona State University, USA
Diansheng Guo, University of South Carolina, USA
Torsten Hahmann, University of Maine, USA

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Monika Sester, and Amy L. Griffin



Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum f r Informatik, Dagstuhl Publishing, Germany



102:xiv Program Committee

- Lars Harrie, Lund University, Sweden
- Francis Harvey, Leibniz Institute for Regional Geography, Germany
- Jan-Henrik Haunert, Universität Bonn, Germany
- Stephen Hirtle, University of Pittsburgh, USA
- Hartwig Hochmair, University of Florida, USA
- Haosheng Huang, University of Zurich, Switzerland
- Qunying Huang, University of Wisconsin-Madison, USA
- Piotr Jankowski, San Diego State University, USA
- Krzysztof Janowicz, University of California at Santa Barbara, USA
- Bernhard Jenny, Monash University, Australia
- Bin Jiang, University of Gävle, Sweden
- Peter Johnson, University of Waterloo, Canada
- Christopher Jones, Cardiff University, UK
- Marinos Kavouras, National Technical University of Athens, Greece
- Carsten Keßler, Aalborg University Copenhagen, Denmark
- Peter Kiefer, ETH Zurich, Switzerland
- Brian Klinkenberg, University of British Columbia, Canada
- Alexander Klippel, Pennsylvania State University, USA
- Margarita Kokla, National Technical University of Athens, Greece
- Werner Kuhn, University of California at Santa Barbara, USA
- Mei-Po Kwan, University of Illinois at Urbana-Champaign, USA
- Phaedon Kyriakidis, Cyprus University of Technology, Cyprus
- Shawn Laffan, University of New South Wales, Australia
- Nina Lam, Louisiana State University, USA
- Patrick Laube, Zurich University of Applied Sciences, Switzerland
- Jiyeong Lee, University of Seoul, South Korea
- Agnieszka Leszczynski, University of Auckland, New Zealand
- Xia Li, Sun Yat-sen University, China
- Xiang Li, East China Normal University, China
- Steve Liang, University of Calgary, Canada
- Hui Lin, Chinese University of Hong Kong, Hong Kong
- Yan Liu, University of Queensland, Australia
- Yu Liu, Peking University, China
- Amy Lobben, University of Oregon, USA
- Jed Long, University of St Andrews, UK
- Feng Lu, Chinese Academy of Sciences, China
- Grant McKenzie, McGill University, Canada
- Liqui Meng, Technical University of Munich, Germany
- Jeremy Mennis, Temple University, USA
- Jennifer Miller, University of Texas at Austin, USA
- Harvey Miller, Ohio State University, USA
- Mir Abolfazl Mostafavi, Laval University, Canada
- Alan Murray, University of California at Santa Barbara, USA
- Tomoki Nakaya, Tohoku University, Japan
- Atsuyuki Okabe, University of Tokyo, Japan

- David O’Sullivan, University of California at Berkeley, USA
- Dimitris Papadias, Hong Kong University of Science and Technology, Hong Kong
- Edzer Pebesma, University of Muenster, Germany
- Karin Pfeffer, University of Amsterdam, The Netherlands
- Ross Purves, University of Zurich, Switzerland
- Martin Raubal, ETH Zurich, Switzerland
- Tarmo Rimmel, University of York, Canada
- Anne Ruas, IFSTTAR, France
- Simon Scheider, University Utrecht, The Netherlands
- Raja Sengupta, McGill University, Canada
- Shih-Lung Shaw, University of Tennessee, USA
- Takeshi Shirabe, KTH Royal Institute of Technology, Sweden
- Alex Singleton, University of Liverpool, UK
- Gaurav Sinha, Ohio University, USA
- Seth Spielman, University of Colorado at Boulder, USA
- Emmanuel Stefanakis, University of New Brunswick, Canada
- Monica Stephens, SUNY Buffalo, USA
- Kathleen Stewart, University of Maryland, USA
- Martin Swobodzinski, Portland State University, USA
- Gautam S. Thakur, Oak Ridge National Laboratory, USA
- Jim Thatcher, University of Washington, USA
- Jean-Claude Thill, University of North Carolina at Charlotte, USA
- Sabine Timpf, University of Augsburg, Germany
- Martin Tomko, University of Melbourne, Australia
- Guillaume Touya, IGN, France
- Ming-Hsiang Tsou, San Diego State University, USA
- Nico Van de Weghe, Ghent University, Belgium
- Maria Vasardani, University of Melbourne, Australia
- Monica Wachowicz, University of New Brunswick, Canada
- Shaowen Wang, University of Illinois at Urbana-Champaign, USA
- Robert Weibel, University of Zurich, Switzerland
- Nancy Wiegand, University of Wisconsin-Madison, USA
- John Wilson, University of Southern California, USA
- Matthew Wilson, University of Kentucky, USA
- Michael Worboys, University of Greenwich, UK
- Ningchuan Xiao, Ohio State University, USA
- Phil Yang, George Mason University, USA
- Eunhye Yoo, SUNY Buffalo, USA
- Bailang Yu, East China Normal University, China
- May Yuan, University of Texas at Dallas, USA
- Naijun Zhou, University of Maryland, USA
- Alexander Zipf, University of Heidelberg, Germany

Early Detection of Herding Behaviour during Emergency Evacuations

David Amores

Infrastructure Engineering, The University of Melbourne, Parkville, VIC 3010, Australia
damores@student.unimelb.edu.au

Maria Vasardani

Infrastructure Engineering, The University of Melbourne, Parkville, VIC 3010, Australia
mvasardani@unimelb.edu.au

Egemen Tanin

Computing and Information Systems, The University of Melbourne, Parkville, VIC 3010, Australia
etanin@unimelb.edu.au

Abstract

Social scientists have observed a number of irrational behaviours during emergency evacuations, caused by a range of possible cognitive biases. One such behaviour is *herding* — people following and trusting others to guide them, when they do not know where the nearest exit is. This behaviour may lead to safety under a knowledgeable leader, but can also lead to dead-ends. We present a method for the automatic early detection of herding behaviour to avoid suboptimal evacuations. The method comprises three steps: (i) people clusters identification during evacuation, (ii) collection of clusters' spatio-temporal information to extract features for describing cluster behaviour, and (iii) unsupervised learning classification of clusters' behaviour into 'benign' or 'harmful' herding. Results using a set of different detection scores show accuracies higher than baselines in identifying harmful behaviour; thus, laying the ground for timely irrational behaviour detection to increase the performance of emergency evacuation systems.

2012 ACM Subject Classification Information systems → Location based services, Computing methodologies → Spatial and physical reasoning

Keywords and phrases spatio-temporal data, emergency evacuations, herding behaviour

Digital Object Identifier 10.4230/LIPICs.GIScience.2018.1

1 Introduction

Certain cognitive biases may govern the way people react and move during emergency evacuations and may result in irrational behaviours that can hinder operations and lead to slower evacuation times, perhaps even endangering lives. An example of a common and well-known behaviour is herding – “when under highly uncertain and stressful situations, an individual tends to follow others almost blindly” [19]. This behaviour sometimes helps people exit a building safely when the leader knows the way out (benign herding), but may otherwise lead people to dead ends (harmful herding). Early identification of such behaviour can aid in more timely, orderly, and ultimately more successful evacuations.

Considering these benefits, this work proposes an automatic method for the early detection of harmful herding behaviour, based on features extracted from the spatio-temporal characteristics of people's group (cluster) movements during emergency evacuations. Figure 1 depicts snapshots of a moving cluster of people during a building evacuation at different times, which displays harmful herding behaviour. Figure 1b shows the point in time when



© David Amores, Maria Vasardani, and Egemen Tanin;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 1; pp. 1:1–1:15

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



(a) Group seemingly heading towards the exit.

(b) Display of harmful herding behaviour.

■ **Figure 1** Snapshots of group behaviour at two different time steps.

the group moves into a room instead of going for the exit. This is when a human observer with knowledge of the building layout would identify this behaviour as erratic and alert the people. The proposed method succeeds in analysing the group's movement trajectory and, more importantly, the group leader's trajectory, to make an *earlier* detection. The assumption is based on the herding behaviour's definition — people delegating wayfinding responsibility to the group's leader. If the leader's past trajectory displays erratic movement, chances that the group will head straight to the nearest exit decrease.

Our method comprises three steps. First, clusters of people traveling together are identified. Second, information about the identified clusters is collected, such as the cluster and cluster leader's moving trajectories, as well as the cluster's distance from the nearest exit. This information is compiled into a *feature vector*. Third, all feature vectors are classified as either *benign* or *harmful* behaviour, using an unsupervised learning classification method. The method is assessed against a ground truth, and also compared to human assessment. The ground truth knows at all times if the ultimate destination of each cluster is the exit or a dead end. The human assessment is performed by visually inspecting the cluster's trajectory and determining the point of wrong going (e.g., turning away from the exit). A set of scores is defined and used to assess the performance of the suggested method when detecting harmful behaviour.

Experiments based on simulated emergency evacuation scenarios show favorable results, as the method outperforms baseline cases and visual inspection in early detection of harmful behaviour. Using different cluster feature combinations, the results also allow for some interesting observations. For example, considering only the actual distance between the cluster and the nearest exit in fact hurts the classification, making it resemble a random one. Instead, the previous moving history of a cluster, rather than its mere distance from an exit, is a better indicator of harmful behaviour. Accordingly, the main contributions of this work are: (1) The identification of spatio-temporal cluster features that can be trusted to describe herding behaviour as either benign or harmful, and (2) a method that uses these features to early detect harmful herding behaviour during emergency evacuations, in an automated way.

The remainder of this paper is organised as follows. Section 2 summarises related research in behaviour detection including simulations, pattern recognition, and personalised evacuation systems. Section 3 discusses the concepts and previously defined behaviours on which our herding detection method is based. Section 4 presents the suggested methodology – clustering, feature extraction, and unsupervised behaviour classification – as a proof of concept for automatic herding behaviour detection. Section 5 discusses different experiments results using various spatio-temporal cluster feature combinations. In Section 6 we present the main findings and suggestions for future work.

2 Related work

Research on crowd behaviour and herding is extensive. A frequent outcome in such research is a simulation depicting more realistic behaviours. Movement patterns, such as hotspots, that arise because of people's biases are also analysed in both outdoor and indoor scenarios. State-of-the-art evacuation systems can use personalised warning messaging and routing directions. This section discusses literature in these areas.

2.1 Simulations displaying social behaviours

Most of the computer-related works that study herding behaviour have the goal of producing simulations. A number of simulations that take into account the microscopic interactions during an evacuation is proposed in the literature (e.g., [8]). Agent-based models are a popular way of creating simulations that include social interactions between the agents. Interactions such as negotiation, following, or collision avoidance can be coded to reproduce common behaviours like herding [19], while cellular automata are frequently used in simulating evacuations [30]. Behaviours such as “freezing by heating”, “faster is slower” and herding behaviour are identified in simulations using a social force model [11]. Although such models are successful in displaying social behaviours, including herding, their identification of such behaviours is done in a visual and manual manner. That is, there is a human checking for instances of behaviour, and papers usually include an image of the seen behaviour. Our method goes a step further by making the behaviour detection automatic.

2.2 Movement and behaviour detection

A number of methods are used for analysis and detection of movement and behaviour patterns. For example, trajectory prediction models using mobile data have been proposed in normal circumstances [17], and during disasters [23]. Such prediction is done with extensive prior knowledge about a person's movement habits. For example, they rely on social networking data to know a person's usual locations. Our model relies on real time and short trajectory knowledge for prediction, and focuses on specifically identifying irrational behaviours.

Hotspot detection is a useful mechanism for alerting stakeholders about people's concentrations. Many hotspot detection mechanisms have been developed for indoor evacuations [9] and crowd disasters [4]. While hotspot detection is useful, detection of other behaviours is rather scarce. The current work specifically targets the detection of harmful herding behaviour.

2.3 Personalised alert and evacuation assistants

Before the wide adoption of mobile technologies, alert systems targeted a large number of persons through mass media. An overview of past research regarding the warning stage of a disaster can be found in [7]. An overview of how warning response, adoption, and timing affects people's behaviours during disasters is given in [24]. With the rise of microblogging services, such as Twitter, further research was conducted in message personalisation. The proposed method aims at the wider use and integration of personalised alert messages produced by observing the real time behaviour of people during emergency evacuations.

Personalised warning messages and assistance is a possibility due to improved research on video tracking technologies and the use of mobile phones. A study in [3] underscores the research needed to send localised warning messages to people's cell phones during an imminent hazard. Furthermore, mobile phone sensors provide grounds for context-aware

indoor navigation. A routing system is proposed in [28] that exploits cell phone sensors in order to have context knowledge in real-time, for example blocked exits. A robot-assisted evacuation method is proposed in [25] improving evacuation times and is tested in a simulated shopping mall environment. These approaches fail to take into account people's beliefs and biases, which may affect their successful adoption. The work in this paper takes a first step into examining people's behaviours, and extracts characteristics that can detect potentially harmful herding caused by cognitive biases. Some relevant work has looked into the role of leaders during emergencies [29]. The authors argue for the optimal number and position of evacuation assistants. However, they only take into account formally defined leaders, rather than leaders that naturally arise in groups of people during emergency evacuations. The latter type of leaders and their behaviour is examined in this work.

3 Background

This section discusses the concepts inherent to herding behaviour, and describes certain methods used in each of the three steps of our methodology: people cluster identification in evacuations, feature extraction to describe cluster behaviour, and a learning model for cluster classification.

3.1 The problem with herding behaviour

Herding behaviour is a cognitive bias examined in early psychology and sociology research [18, 5] comprising different contexts of everyday life. In the context of evacuations, herding behaviour is exhibited when people follow others, without knowing with certainty where the group is heading to. Although herding can successfully lead people towards a safe place, it can also lead them to prevent successful evacuations, as evidenced by past studies in bushfires [1] and indoor evacuations [9]. A study in [10] considers a balance between individualistic behaviour and herding behaviour to be optimal for indoor evacuations. This research focuses on identifying harmful herding behaviour. For language consistency, we distinguish **benign** herding behaviour – when people follow others successfully to safety – from **harmful** herding behaviour – when the group fails to find an exit.

3.2 Moving people clustering

The first step in our method is cluster identification during evacuation. We borrow ideas from previous works that have studied crowd clustering [21, 20] and groups of points moving together [14]. Clustering methods often use Euclidean distance for assigning members in a cluster. Nevertheless, several applications, including this work, require non-traditional distance measures, such as graph distance or similarity measures. The suggested method clusters people in a floor setting; therefore, people separated by a wall should not be assigned to the same cluster, even if their Euclidean distance is short. Spectral clustering takes a similarity matrix as input for identifying clusters [15]. Such a similarity matrix can be computed from any pair-wise distance metric of the instances – persons in our case. The way the similarity matrix is built in this work is explained in Section 4.1.

3.3 Feature extraction for conveying herding behaviour

The second step involves the collection of spatio-temporal information from clusters previously defined, to be encoded into a feature vector. The set of these features is used to describe the

harming herding behaviour that a cluster might be displaying, and is one of this work's major contributions. Previous work in activity recognition and anomaly detection from trajectories provides inspiration for this model. As in many learning problems, feature engineering is a crucial step towards an effective model. Additionally, motion information representation is the basis in spatiotemporal analysis [13]. Consequently, several approaches encode trajectory information (e.g. distance between objects, acceleration) into their feature vector [32, 22].

The aim of this work is to produce a feature vector that describes herding behaviour. Relevant characteristics are:

- **Characteristic 1.** Forming groups.
- **Characteristic 2.** Moving towards or away from an exit.
- **Characteristic 3.** Delegating wayfinding to the leader and then moving collectively.

Characteristic 1 is achieved by the clustering step. Characteristics 2 is encoded into the feature vector by calculating the distance change from the cluster towards or away from exits. For satisfying characteristic 3, the trajectory from the cluster's leader is analysed from previous time steps. How these features are formally obtained is explained in Section 4.

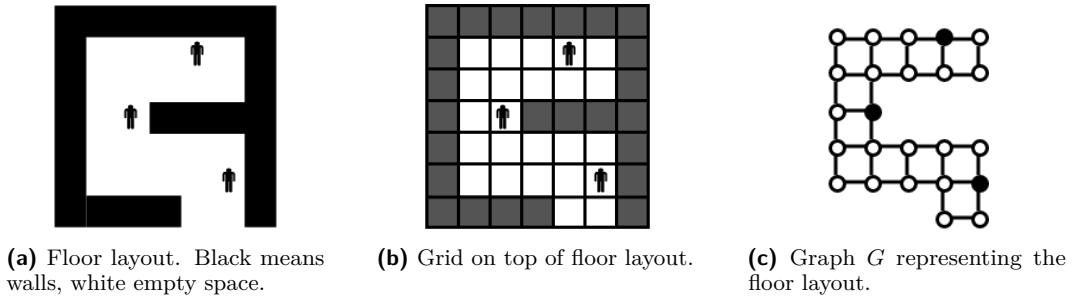
3.4 Learning model

The proposed approach uses an unsupervised learning method to identify the clusters heading towards a dead end. Machine learning methods are now a common practice for categorising a set of instances. Each instance comprises a set of features and may contain continuous or discrete values. As such, learning methods are used for the detection of differing behaviours or anomalies. Previous works for categorising trajectories and behaviours have used semi-supervised [22] and unsupervised [32] learning models by means of different clustering algorithms such as Gaussian mixture, or Latent Dirichlet Allocation (LDA).

The proposed method compares two different and widely used unsupervised learning algorithms: k-means clustering and hierarchical clustering. K-means clustering finds a centroid per cluster and uses a distance based metric to classify points based on the proximity to the centroid. Hierarchical clustering performs better on non-linear and high-dimensional data. Our method has high dimensionality as it uses up to 55 different features.

3.5 Data sources and simulation

The proposed method assumes known coordinate positions of each person for the duration of the emergency evacuation. As real data of this type are scarce, a simulation instead is used, while current complementary research efforts are developing technologies for real-time monitoring of evacuees [6]. Also, in order to focus on examining herding behaviour, the effects of indoor landmarks on way finding, or the limits of maximum evacuation times and multi-level building complexities are left for future consideration. The simulation is built based on the general guidelines provided in [19]. In that work, the authors construct a simulation that displays different "nonadaptive crowd behaviours", including herding behaviour. They build an agent-based model in which agents display social interactions, such as negotiation or people-following. They define a set of possible actions and different types of profiles. The simulation used in this paper uses a subset of those actions and profiles for displaying the expected behaviour (herding). The following set of possible actions is used: (i) **Random walk** - heads towards a random direction in sight, (ii) **Seek** - if the exit is known, heads to the exit; otherwise, keep looking for the exit by going towards doors, and (iii) **Target following** - follow the nearest group of people.



■ **Figure 2** A 7x7 floor layout discretization.

■ **Table 1** Similarity matrix of sample pair distances.

	p1	p2	p3
p1	0	-4	-9
p2	-4	0	-5
p3	-9	-5	0

Accordingly, three profiles are used for agents. The exact probabilities are not provided in [19], so they are based on evacuation behaviour findings in [31] and [16]. Each profile contains the probabilities for the actions it can take (probabilities must sum up to 1).

- Adult: $random_walk = 0.2$, $seek = 0.4$, $target_following = 0.4$
- Child: $random_walk = 0.3$, $seek = 0.2$, $target_following = 0.5$
- Elderly: $random_walk = 0.0$, $seek = 0.7$, $target_following = 0.3$

4 Method

The methodology used to detect harmful herding behaviour comprises three steps. The purpose of detecting herding behaviour is to know if people may be headed towards a dead-end, or taking a much longer evacuation route. In this case, the behaviour belongs to a group of people rather than to individuals. As such, the method first identifies clusters at each time step. A feature vector is extracted from each of these clusters and an unsupervised learning method is used to predict the ones displaying herding behaviour. The following subsections describe each step in the methodology.

4.1 Clustering

The floor layout is discretised into a grid and represented by a graph G . Each grid cell that is not a wall is a node of G . Figure 2a shows a sample 7x7 floor layout, figure 2b shows a grid on top of it, and figure 2c shows the respective graph G . Black dots represent people, and each vertex in G is connected to the nodes up, down, left and right.

At each time step, each person is located at a node of G (as in Figure 2c) and the distances between each pair of persons is computed into a *similarity matrix*. The sample persons in Figure 2 are located at (4, 5), (2, 3), and (5, 1), and we call them p_1 , p_2 , and p_3 respectively. The *similarity matrix* for the sample 3 persons is shown in Table 1.

The similarity matrix contains the distance of each pair of points multiplied by -1 , to represent a *similarity* rather than *dissimilarity*. Computing shortest paths in a graph at each time step can be time consuming. Therefore, the paths are pre-computed and stored

in a hash table *PATHS* in memory, such that for pair p_1 and p_2 we can obtain its graph distance by calling $PATHS(p_1, p_2)$. An additional variable, *EXITS* stores distances from each person to the nearest exit (e.g., $EXITS(p_1)$).

The *similarity matrix* is then input to the spectral clustering algorithm for cluster identification. The clustering step represents virtually the whole method's time complexity as $O(n^3)$, while next steps run in linear time or less. When clustering is performed at every time step, it might produce temporal errors. For example, people passing each other in opposite directions could temporarily be close together but shouldn't be considered part of the same cluster. To address this, we use a parameter τ that represents the number of time steps required for a group of people to be considered as 'traveling together'.

Over time, clusters may add members, lose members, split, or even completely dissolve. Therefore, identifying a cluster over time requires some flexibility about its members. Thus, we define the equivalence between cluster C_1 from time step t and cluster C_2 from time step $t+1$, if $|C_1 \cap C_2| \geq 2 \implies C_1 \equiv C_2$, and we define the age of cluster C as $T_C = |C_t, \dots, C_{t+n}|$ where $C_i \equiv C_{i+1} \forall i \in \{i | t \leq i \leq t+n\}$. With that, the age constraint for cluster C to be considered herding is $T_C \geq \tau$. For the experiments described in Section 5, a visual inspection of the moving clusters showed a τ value of 5 ensures a group of people are moving together. This paper doesn't cover the effect that varying values of τ can have on the discovery of moving clusters. For more thorough techniques on this area the reader is referred to [12].

4.2 Behaviour Definition and Feature Vector

Once the clusters of people traveling together are identified, a cluster feature vector is extracted from each. The method relies on the group's leader past trajectory as one of the most important features for behaviour description. So before listing the feature candidates, a formal spatio-temporal definition for 'leader' is provided.

Leader identification

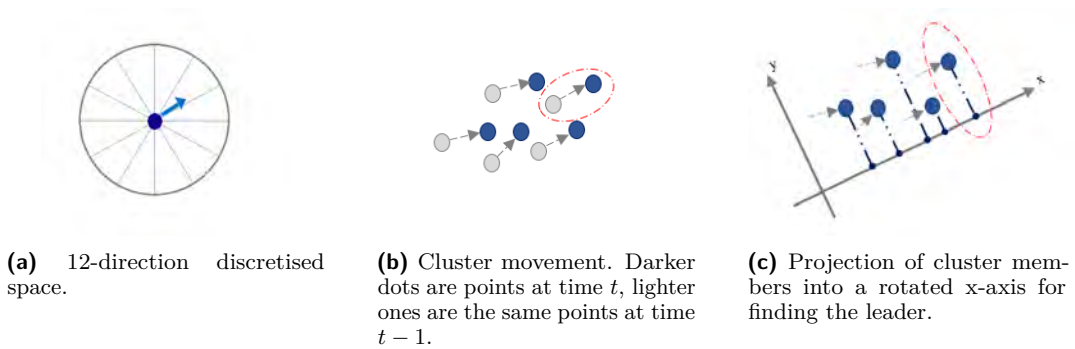
In plain terms, the leader is the person guiding the group. However, that definition is not enough for identifying the leader in spatio-temporal data. A simple definition of cluster leader is used where the leader is considered the most salient point in the cluster's moving orientation, as depicted in Figure 3b. More elaborate methods for leader identification fall out of the scope of this study, but the interested reader is directed to [2].

To obtain the cluster's orientation, a 12-direction discretised space is used (Figure 3a). The discretised angle of point m , is $\angle m$. Then, given a cluster C with n members m_0, \dots, m_n , the orientation of C is defined as the mode of the discretised angles of its members: $\angle C = Mo(\angle m_0, \dots, \angle m_n)$. Once $\angle C$ is computed, a plane rotation of $\angle C$ is performed, as shown in Figure 3c, and every member m_i is projected into the x-axis. From there, leader l_p is the member with the p -largest projection in the x-axis.

Feature candidates

In order to comply with the characteristics of herding behaviour listed in Section 3.3, three kinds of feature candidates (*FC*) are extracted from cluster C with members m_0, \dots, m_n :

- ***FC*₁ - Cluster's distance to exit** (*dist_to_exit*) – The average shortest distance to the closest exit for each member m of C . Distances from m to the nearest exit are stored in the *EXITS* hash table. So $dist_to_exit(C)$ determines this feature's value.



■ **Figure 3** Graphical definitions of orientation and cluster leader.

- **FC_2 - Cluster's distance change towards exit** (dist_change_i) – The change in the average shortest distance from the i previous time steps to the current one. If the change is negative it means the cluster is getting closer to the exit. This is computed by checking the positions of the members in the previous time step and using the stored distances in the *EXITS* hash table.
- **FC_3 - Leader's trajectory** ($\text{leader}_l\text{_away_steps}_i$) – This field refers to the number of steps the group leader l has taken away from the exit in the last i time steps. For instance, $\text{leader}_1\text{_away_steps}_5$ (i.e., $l = 1$ and $i = 5$) counts how many of the previous 5 steps leader l took away from the exit. The value would range from 0 to 5 in this example, and from 0 to i in general.

At every time step, clusters are identified and features extracted. The set of features to use can be the full set described, or a subset of it. The experiments in Section 5 use different subsets of the features explained here. Every feature set is stored and used in the unsupervised learning method explained in the next subsection.

4.3 Unsupervised Learning

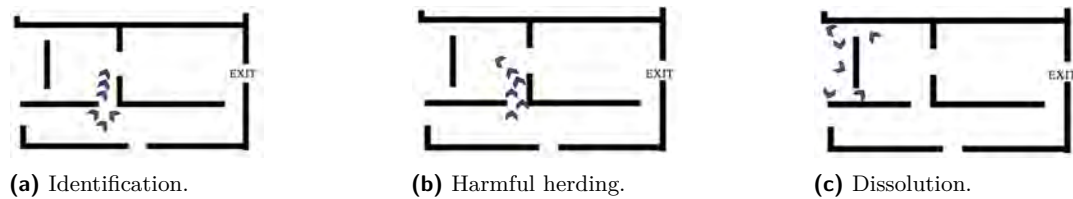
The final step of the method is applying a learning algorithm for classifying clusters displaying benign or harmful herding behaviour. Thus, two classes are defined: *benign* and *harmful*. As mentioned in Section 3.4, two unsupervised learning algorithms are used: k-means (KM) and hierarchical clustering (HC). Additionally, three baselines are used for thorough comparison:

- Zero rule: Classifies every instance as the most popular one. In this case, it will classify everything as harmful.
- Random: Classifies each instance randomly as either harmful or benign.
- Random with distribution: Classifies similar to the Random baseline but uses prior knowledge about the distribution of harmful and benign instances.

Comparing to a "dumb rule" classifier, such as Zero rule, ensures the proposed method meets minimum requirements, while comparing to the random baselines ensures it does not perform randomly. Comparisons with baselines ascertain credibility and robustness.

4.4 Evaluation Method

To evaluate the suggested method, every instance is associated with a label – *benign* or *harmful* – describing its behaviour. An instance refers to a cluster from its identification until its dissolution. Figure 4 shows a cluster in different stages of its lifespan.



■ **Figure 4** Three stages in the lifespan of a cluster of 6 persons: (a) the group is identified as such, (b) the human annotator identifies the group is taking the wrong turn, (c) the cluster dissolves after an exit is not found.

The ground truth holds every instance's label based on the *ultimate* cluster's destination. That is, if the cluster ends up in a dead-end or clearly goes in the wrong direction, it is labeled as *harmful*, whereas if it ends up exiting the building or closer to the exit, it is labeled as *benign*. The suggested method is evaluated on its ability to detect harmful herding behaviour but also on detection timeliness, as it is expected to make detections early on. Therefore, three checkpoints along the lifespan of a cluster are defined (Figure 4):

- **Checkpoint 1 (CP1)** – At cluster identification. This is when the cluster is identified by the clustering method defined in Section 4.1 (Figure 4a).
- **Checkpoint 2 (CP2)** – At human detection point. That is, when the human tester first realises that the cluster is headed towards the wrong direction (Figure 4b).
- **Checkpoint 3 (CP3)** – At cluster dissolution. This is when the clustering method defined in Section 4.1 stops identifying the former cluster members as one (Figure 4c).

Then, to assess detection timeliness, five scores – called **detection scores** – are defined using the checkpoints:

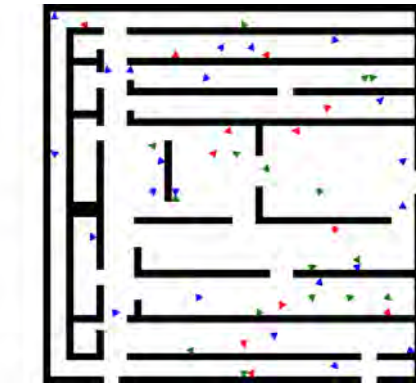
- **Early detection (ED)** – Number of harmful instances detected before CP1.
- **Detection (D)** – Number of harmful instances detected between CP1 and CP2.
- **Late detection (LD)** – Number of harmful instances detected between CP2 and CP3.
- **No detection (ND)** – Number of harmful instances not detected at all.
- **False warnings (FW)** – Number of benign instances detected as harmful at any time.

Additionally, **unified scores**, allowing a comparison between the method's detection times and the visual inspection (VI), are defined:

- **Before VI (BVI)** – The number of harmful instances detected before CP3 (faster than VI), plus the number of benign instances not identified as harmful. Formally, $BVI = ED + D + (TB - FW)$, where TB is the total number of benign instances in the ground truth.
- **After VI (AVI)** – The number of harmful instances detected after CP3 (slower than VI), plus the non-detected instances, plus the false warnings. Formally, $AVI = LD + ND + FW$

It is worth noting that false warnings tend to be sensitive, as a single harmful detection in a whole benign trajectory would yield a false warning. For that reason, a tolerance variable is introduced. Each of the detection scores checks for at least *one* harmful prediction. Using the tolerance variable t , the detection scores have to check for at least t harmful predictions, before classifying it as harmful.

The manual labeling is performed visually by a human observer. Although not optimal, this labeling allows for performing a proof-of-concept evaluation method against human judgment. In the future, a more thorough labeling mechanism such as domain expert labeling, or labeling from multiple annotators can be used.



■ **Figure 5** Initial setup of the simulation.

5 Experiments

Data for the experiments are generated by running the simulation described in Section 3.5. Location data for each agent at each time step are recorded in a text file. The text file is used as the input to the suggested method. The simulation is realised using the GAMA¹ simulation software. In the simulation, 50 agents are placed on a 50x50 grid. The layout of the grid resembles a building layout with walls and exit doors. Figure 5 shows the initial setting for the simulation to run.

Having obtained the simulation data, the main objective of these experiments is to test which features in the cluster feature vector describe best the herding behaviour. Feature sets are built using feature candidates (FC_{1-3}) described in Section 4.2, as follows:

- Feature set 1 (FS_1) – The information this feature set contains is the cluster’s distance to the exit (FC_1), the cluster’s previous movements (FC_2 with $i = 5$), and the trajectories of 3 leaders (FC_3 with $l = 3$ and $i = 20$).
- Feature set 2 (FS_2) – In this feature set, distance to the exit (FC_1) is not used, for checking its relevance. Considered are: cluster movement (FC_2 with $i = 5$) and leader trajectory (FC_3 with $l = 1$ and $i = 20$).
- Feature set 3 (FS_3) – Leader information (FC_3) is not considered, to check its relevance. Considered are only distance to exit (FC_1) and cluster movement (FC_2 with $i = 5$).

The values for the number l of leaders and number i of steps to check from past trajectory were chosen based on the behaviour definition and by performing several preliminary tests of the method with a number of combinations. Three experiments are performed, summarised in Table 2. Every experiment runs the classification step using both k-means (KM) and hierarchical clustering (HC):

- **Experiment 1** sets the tolerance value to 1, the number of instances to 31 and all three feature sets are compared.
- **Experiment 2** is similar to Experiment 1, but using a tolerance value of 2.
- **Experiment 3** is used to check whether the algorithm would benefit from having more instances to cluster by increasing the number N of instances and using the best performing feature set – FS_2 as seen later – with tolerance $t = 1$.

¹ <http://gama-platform.org/>

■ **Table 2** Parameters used in every experiment.

Experiment 1			Experiment 2			Experiment 3		
$tolerance = 1$			$tolerance = 2$			$tolerance = 1$		
FS_1	FS_2	FS_3	FS_1	FS_2	FS_3	FS_2		
$N = 31$			$N = 31$			$N = 21$	$N = 31$	$N = 52$

■ **Table 3** Results of Experiment 1, using tolerance $t = 1$. Showing detection scores (ED, D, LD, ND), false warnings (FW), and unified scores (BVI, AVI) of k-means (KM) and hierarchical clustering (HC) using different feature sets (FS_i). Baselines are shown beside them for comparison

	FS_1		FS_2		FS_3		Baselines		
	KM	HC	KM	HC	KM	HC	ZR	R	RD
ED	48%	52%	57%	76%	100%	100%	100%	86%	81%
D	24%	33%	43%	24%	0%	0%	0%	14%	19%
LD	19%	10%	0%	0%	0%	0%	0%	0%	0%
ND	10%	5%	0%	0%	0%	0%	0%	0%	0%
FW	36%	55%	36%	55%	100%	100%	100%	100%	100%
BVI	69%	72%	88%	81%	66%	66%	66%	66%	66%
AVI	31%	28%	12%	19%	34%	34%	34%	34%	34%

6 Results Analysis

Tables 3 and 4 show the complete results of Experiment 1 and 2, respectively. The tables contain detection and unified scores (Section 4.4) for a thorough comparison. The tables present the results of k-means and hierarchical clustering for all three feature sets (FS_1 , FS_2 , FS_3). The three baselines defined in Section 4.3 – Zero Rule (ZR), Random (R), Random with distribution (RD) – are placed next to the results for comparison.

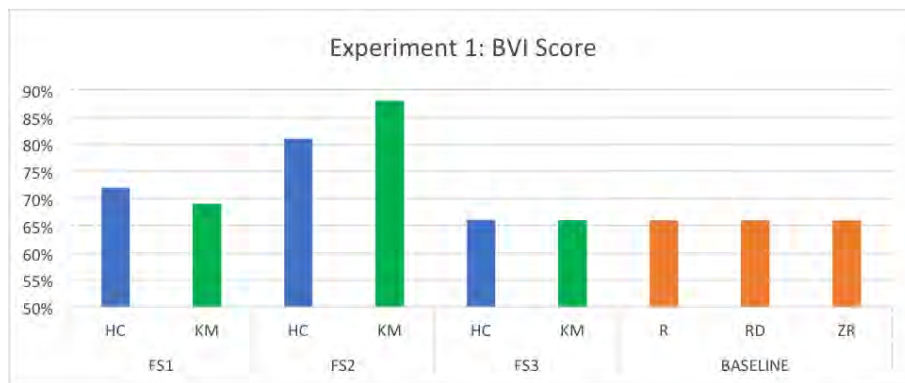
Ideally, a method would detect every harmful herding behaviour early on ($ED = 100\%$). Even though the baselines have a perfect or near-perfect ED score — since ZR classifies everything as harming ($ED = 100\%$) — they also have a 100% false warning rate (FW), which renders these baselines unreliable. Hence, the consolidated BVI score is a better indicator of overall performance, as it penalises either low detection, or high false warning rates. Figures 6 and 7 show the BVI score in Experiments 1 and 3, respectively, while the main findings of the analysis are as follows:

Leader trajectory is the best herding predictor. Overall, the best performing feature set is FS_2 with either k-means, or hierarchical clustering with a $BVI = 88\%$ and $BVI = 81\%$ with $t = 1$ (Figure 6) respectively, and $BVI = 81\%$ and $BVI = 84\%$ with $t = 2$. These algorithms all perform well above the baselines. These positive results suggest the features chosen, namely the leader trajectory and the recent cluster movement, were appropriate. When comparing Experiments 1 and 2, as tolerance increases, the FW score decreases as expected, but the overall BVI is not improved.

Distance to exit is not meaningful. Low results of FS_3 suggest the distance to the exit (the feature not present in FS_2) is not a trusting feature, as it makes the classifier act randomly. This is probably the reason for the lower performance of FS_1 compared to FS_2 , as it contains the $dist_to_exit$ feature. This observation is reasonable, given that long distance from the exit does not necessarily mean the group is lost or not heading towards the exit.

■ **Table 4** Results of Experiment 2, using tolerance $t = 2$. Showing detection scores (ED, D, LD, ND), false warnings (FW), and unified scores (BVI, AVI) of k-means (KM) and hierarchical clustering (HC) using different feature sets (FS_i). Baselines are shown beside them for comparison.

	FS_1		FS_2		FS_3		Baselines		
	KM	HC	KM	HC	KM	HC	ZR	R	RD
ED	33%	38%	29%	43%	62%	62%	62%	52%	57%
D	24%	29%	57%	52%	33%	33%	33%	38%	33%
LD	29%	24%	14%	5%	5%	5%	5%	10%	2%
ND	14%	10%	0%	0%	0%	0%	0%	0%	0%
FW	36%	36%	27%	36%	100%	100%	100%	100%	100%
BVI	59%	66%	81%	84%	62%	62%	62%	59%	59%
AVI	41%	34%	19%	16%	38%	38%	38%	41%	41%



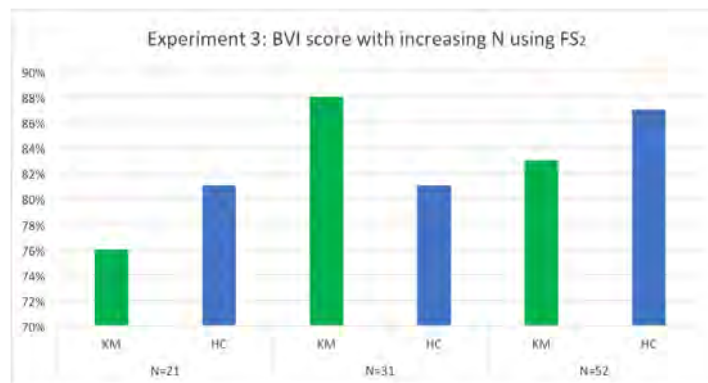
■ **Figure 6** Experiment 1 results summary. *BVI* score is displayed comparing the suggested method to the baselines.

Increasing number of instances improves performance. Figure 7 shows the results of Experiment 3, depicting how the scores change given an increasing number N of cluster instances. *BVI* score increases as N increases (except for k-means in $N = 31$), implying that the suggested method benefits from a higher number of instances. FS_2 is used in this experiment as it was the best performing feature set in the previous experiments.

7 Conclusions and future work

This paper presents a method for automatic, early detection of harmful herding behaviour using spatio-temporal information from clusters of people. The method comprises three steps. First, groups of people moving together are identified using clustering algorithms with added constraints. Second, relevant spatio-temporal information from the identified clusters is collected. Second, the extracted features are combined to spatially and temporally describe a herding behaviour. To achieve this, the position changes of the cluster and the cluster leader's movement trajectory are examined. The method assumes the leader's trajectory to be a most relevant feature for identifying the behaviour. Third, the observed clusters are classified as displaying either benign or harmful herding behaviour, using an unsupervised learning method.

The experimental results show promise towards advancing the understanding of herding behaviour effects. Seven different scores are defined to assess the method's ability to detect



■ **Figure 7** Experiment 3 results. FS_2 is used with an increasing number N of instances.

harmful behaviours and compare it to a human observer. In every experiment run, both algorithms (k-means and hierarchical clustering) are superior to the three baselines used. Different combinations of features were tested. The major findings are:

1. Features regarding leader trajectory and recent distance changes from the cluster to the exit best predict harmful herding behaviour, yielding above 80% of the BVI unified score in the experiments.
2. Distance to the exit (without considering movement) harms the prediction when added into the feature set, making it classify randomly.
3. Even though increasing the method's tolerance does not produce better results overall, it does decrease the amount of false warnings. This is useful in systems where issuing warnings is expensive, so additional confidence is needed.
4. The method benefits from large cluster instances in the data, which means that it scales well in environments with big crowds that need to evacuate in an emergency situation. Higher values of N , however, mean more time-consuming manual labeling for evaluation.

The harmful herding behaviour identification method can be further improved. Different graphs can be used, such as the visibility graph [26] or a bigraph [27] in the clustering step. The features extracted from the clusters can be improved by looking into more in-depth analysis of who the leader of a group is, rather than identifying the topmost one as such. A supervised learning method for behaviour classification can be compared to its unsupervised counterpart. An approach that would replace both learning approaches is a rules-based one where, given thorough domain knowledge, strict rules can be placed for the prediction of the harmful herding behaviour. Pertaining to the evaluation method, perhaps the most immediate step forward is the use of a real evacuation scenario datasets. Finally, herding is only one of several behaviours elicited by cognitive biases during disasters. Other biases such as the normalcy bias, confirmation bias, planning fallacy [1], may lead to equally harming behaviours during emergency evacuations. Consequently, future work may focus on identifying other behaviours, or even providing a bigger unified framework for irrational behaviour detection.

References

- 1 Carole Adam and Benoit Gaudou. Modelling human behaviours in disasters from interviews: Application to melbourne bushfires. *Journal of Artificial Societies and Social Simulation*, 20(3):12, 2017. URL: <http://jasss.soc.surrey.ac.uk/20/3/12.html>, doi:10.18564/jasss.3395.

- 2 Mattias Andersson, Joachim Gudmundsson, Patrick Laube, and Thomas Wolle. Reporting leaders and followers among trajectories of moving point objects. *GeoInformatica*, 12(4):497–528, 2008.
- 3 Hamilton Bean, Jeannette Sutton, Brooke F Liu, Stephanie Madden, Michele M Wood, and Dennis S Mileti. The study of mobile public warning messages: A research review and agenda. *Review of Communication*, 15(1):60–80, 2015.
- 4 Arianna Bottinelli, David TJ Sumpter, and Jesse L Silverberg. Emergent structural mechanisms for high-density collective motion inspired by human crowds. *Physical review letters*, 117(22):228301, 2016.
- 5 James Coleman. Foundations of social theory. *Cambridge, MA: Belknap*, 1990.
- 6 Davide Dardari, Pau Closas, and Petar M Djurić. Indoor tracking: Theory, methods, and technologies. *IEEE Transactions on Vehicular Technology*, 64(4):1263–1278, 2015.
- 7 Nicole Dash and Hugh Gladwin. Evacuation decision making and behavioral responses: Individual and household. *Natural Hazards Review*, 8(3):69–77, 2007.
- 8 Marcus Goetz and Alexander Zipf. Using crowdsourced geodata for agent-based indoor evacuation simulations. *ISPRS International Journal of Geo-Information*, 1(2):186–208, 2012.
- 9 Dirk Helbing, Lubos Buzna, Anders Johansson, and Torsten Werner. Self-organized pedestrian crowd dynamics: Experiments, simulations, and design solutions. *Transportation science*, 39(1):1–24, 2005.
- 10 Dirk Helbing, Illés Farkas, and Tamas Vicsek. Simulating dynamical features of escape panic. *Nature*, 407(6803):487, 2000.
- 11 Dirk Helbing, Illes J Farkas, Peter Molnar, and Tamás Vicsek. Simulation of pedestrian crowds in normal and evacuation situations. *Pedestrian and evacuation dynamics*, 21(2):21–58, 2002.
- 12 Panos Kalnis, Nikos Mamoulis, and Spiridon Bakiras. On discovering moving clusters in spatio-temporal data. In *International Symposium on Spatial and Temporal Databases*, pages 364–381. Springer, 2005.
- 13 Teng Li, Huan Chang, Meng Wang, Bingbing Ni, Richang Hong, and Shuicheng Yan. Crowded scene analysis: A survey. *IEEE transactions on circuits and systems for video technology*, 25(3):367–386, 2015.
- 14 Xiaohui Li, Vaida Ceikute, Christian S Jensen, and Kian-Lee Tan. Effective online group discovery in trajectory databases. *IEEE Transactions on Knowledge and Data Engineering*, 25(12):2752–2766, 2013.
- 15 Jialu Liu and Jiawei Han. Spectral clustering. In Charu Aggarwal and Chandan Reddy, editors, *Data Clustering: Algorithms and Applications*. CRC Press, 2013.
- 16 Min Liu and Siu Ming Lo. The quantitative investigation on people’s pre-evacuation behavior under fire. *Automation in construction*, 20(5):620–628, 2011.
- 17 Xin Lu, Linus Bengtsson, and Petter Holme. Predictability of population displacement after the 2010 haiti earthquake. *Proceedings of the National Academy of Sciences*, 109(29):11576–11581, 2012.
- 18 Alexander Mintz. Non-adaptive group behavior. *The Journal of abnormal and social psychology*, 46(2):150, 1951.
- 19 Xiaoshan Pan, Charles S Han, Ken Dauber, and Kincho H Law. Human and social behavior in computational modeling and analysis of egress. *Automation in construction*, 15(4):448–461, 2006.
- 20 Daniel Roggen, Martin Wirz, Gerhard Tröster, and Dirk Helbing. Recognition of crowd behavior from mobile sensors with pattern analysis and graph clustering methods. *arXiv preprint arXiv:1109.1664*, 2011.

- 21 Shobhit Saxena, François Brémond, Monnique Thonnat, and Ruihua Ma. Crowd behavior recognition for video surveillance. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 970–981. Springer, 2008.
- 22 Rowland R Sillito and Robert B Fisher. Semi-supervised learning for anomalous trajectory detection. In *BMVC*, volume 1, pages 035–1, 2008.
- 23 Xuan Song, Quanshi Zhang, Yoshihide Sekimoto, and Ryosuke Shibasaki. Prediction of human emergency behavior and their mobility following large-scale disaster. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 5–14. ACM, 2014.
- 24 John H Sorensen and Barbara Vogt Sorensen. Community processes: Warning and evacuation. In *Handbook of disaster research*, pages 183–199. Springer, 2007.
- 25 Bo Tang, Chao Jiang, Haibo He, and Yi Guo. Human mobility modeling for robot-assisted evacuation in complex indoor environments. *IEEE Transactions on Human-Machine Systems*, 46(5):694–707, 2016.
- 26 Alasdair Turner, Maria Doxa, David O’sullivan, and Alan Penn. From isovists to visibility graphs: a methodology for the analysis of architectural space. *Environment and Planning B: Planning and design*, 28(1):103–121, 2001.
- 27 Lisa A Walton and Michael Worboys. A qualitative bigraph model for indoor space. In *International Conference on Geographic Information Science*, pages 226–240. Springer, 2012.
- 28 Jing Wang, Haifeng Zhao, and Stephan Winter. Integrating sensing, routing and timing for indoor evacuation. *Fire Safety Journal*, 78:111–121, 2015.
- 29 Xiaolu Wang, Xiaoping Zheng, and Yuan Cheng. Evacuation assistants: An extended model for determining effective locations and optimal numbers. *Physica A: Statistical Mechanics and its Applications*, 391(6):2245–2260, 2012.
- 30 Song Wei-Guo, Yu Yan-Fei, Wang Bing-Hong, and Fan Wei-Cheng. Evacuation behaviors at exit in ca model with force essentials: A comparison with social force model. *Physica A: Statistical Mechanics and its Applications*, 371(2):658–666, 2006.
- 31 CM Zhao, Siu Ming Lo, SP Zhang, and M Liu. A post-fire survey on the pre-evacuation human behavior. *Fire Technology*, 45(1):71, 2009.
- 32 Bolei Zhou, Xiaogang Wang, and Xiaoou Tang. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2871–2878. IEEE, 2012.


What Makes Spatial Data Big?

A Discussion on How to Partition Spatial Data

Alberto Belussi

Department of Computer Science, University of Verona, Italy


alberto.belussi@univr.it

 <https://orcid.org/0000-0003-3023-8020>

Damiano Carra

Department of Computer Science, University of Verona, Italy


damiano.carra@univr.it

 <https://orcid.org/0000-0002-3467-1166>

Sara Migliorini

Department of Computer Science, University of Verona, Italy

sara.migliorini@univr.it

 <https://orcid.org/0000-0003-3675-7243>

Mauro Negri

Department of Electronics, Information and Bioengineering, Politecnico of Milan, Italy

mauro.negri@polimi.it

Giuseppe Pelagatti

Department of Electronics, Information and Bioengineering, Politecnico of Milan, Italy

giuseppe.pelagatti@polimi.it

Abstract

The amount of available spatial data has significantly increased in the last years so that traditional analysis tools have become inappropriate to effectively manage them. Therefore, many attempts have been made in order to define extensions of existing MapReduce tools, such as Hadoop or Spark, with spatial capabilities in terms of data types and algorithms. Such extensions are mainly based on the partitioning techniques implemented for textual data where the dimension is given in terms of the number of occupied bytes. However, spatial data are characterized by other features which describe their dimension, such as the number of vertices or the MBR size of geometries, which greatly affect the performance of operations, like the spatial join, during data analysis. The result is that the use of traditional partitioning techniques prevents to completely exploit the benefit of the parallel execution provided by a MapReduce environment. This paper extensively analyses the problem considering the spatial join operation as use case, performing both a theoretical and an experimental analysis for it. Moreover, it provides a solution based on a different partitioning technique, which splits complex or extensive geometries. Finally, we validate the proposed solution by means of some experiments on synthetic and real datasets.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases Spatial join, SpatialHadoop, MapReduce, partitioning, big data

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.2

Acknowledgements This work was partially supported by the Italian National Group for Scientific Computation (GNCS-INDAM). This work has been supported by “Progetto di Eccellenza” of the Computer Science Dept., University of Verona, Italy.



© Alberto Belussi, Damiano Carra, Sara Migliorini, Mauro Negri, and Giuseppe Pelagatti; licensed under Creative Commons License CC-BY

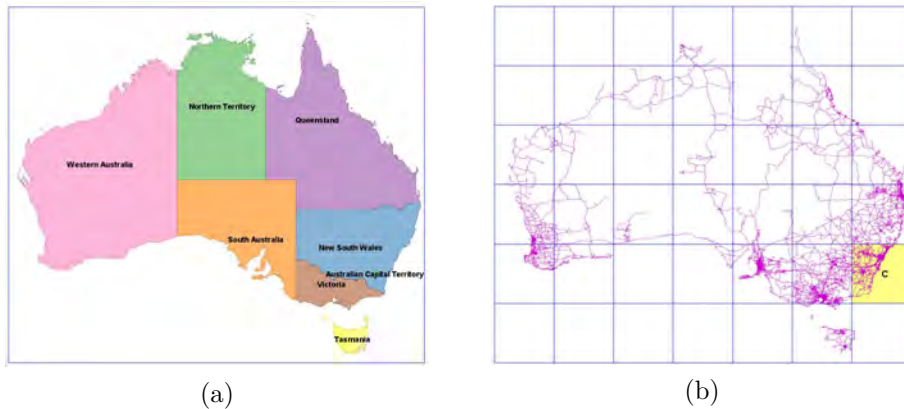
10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 2; pp. 2:1–2:15

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



■ **Figure 1** Example of unbalanced datasets between which a join has to be performed. (a) contains few geometries with a big extent described with a restricted number of vertices, while (b) contains many geometries with a small extent represented using several vertices.

1 Introduction

In recent years the amount of spatial data available to users have increased tremendously and the demand of resources for performing geo-spatial analysis on them cannot be satisfied any more by traditional GIS systems. As a consequence of this new scenario, in the last decade many efforts have been devoted to the extension of systems for big data processing based on the MapReduce paradigm, like Hadoop [15] or Spark [16], in order to make them able to deal with geo-spatial data. For instance, SpatialHadoop [7] is the result of one of these projects, it is an extension of Apache Hadoop which provides a native support for spatial data, in terms of spatial data types, operations and indexes. In particular, it provides various implementations of the spatial join, which is one of the most frequently used operation for analyzing spatial datasets and discovering connections between geo-spatial objects [2].

Various spatial join variants are available in literature [10] and some adaptations to the MapReduce context have been provided [8]. In particular, SpatialHadoop implements several spatial join algorithms which share the use of indexes for increasing their performance and avoiding a brute force approach that simply subdivides the Cartesian product of the two input datasets between tasks. As regards to the indexing techniques, all kind of indexes provided by SpatialHadoop are organized into two levels: (i) first data are physically partitioned in different blocks (usually called splits or partitions), producing a first level of index called *global index*, then (ii) in each block a specific index is built that works only on the data of the partition, producing a second level of index called *local index*. This indexing pattern directly derives from the way usually applied for organizing data inside the HDFS (Hadoop Distributed File System). In HDFS, a dataset is partitioned into splits whose size usually corresponds to the HDFS block size and each split represents the input for a single map task. This organization has been originally developed for processing large amount of mono-dimensional (textual) data where the execution time is directly affected by the number of bytes they occupy on the file system. This choice is justified by the observation that the amount of work to be performed on textual data usually depends on the data size (or number of records), thus partitioning data in blocks of the same size and assigning each block to a map task, produces a balanced work distribution among workers. This reasoning has been applied also to spatial data, since the physical partitioning induced by the global index uses again considerations based on the size in bytes of the dataset. However, geo-spatial objects

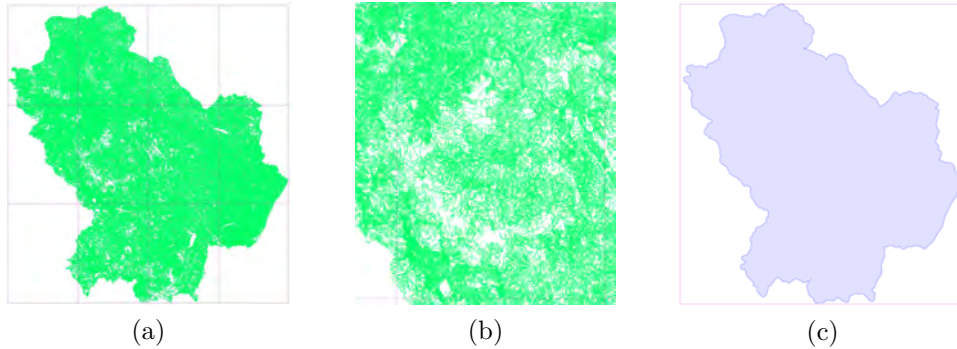
are also embedded in a 2D or a 3D reference space and their extension in these spaces is another dimension that can have an impact on the workload of many spatial operations. Notice that the portion of space occupied by a geo-spatial object on the Earth surface can be completely independent from the size in bytes of its physical representation as a file record. On the contrary the number of bytes may partially represent the complexity of a shape, in terms of number of vertices, but not its extent. During the execution of a spatial join, the average extent of the geo-spatial objects in both datasets affects their mutual selectivity (i.e., the ratio between the effective pairs produced by the join and the total number of possible pairs given by the Cartesian product) and thus it has an impact on the workload of the tasks devoted to its computation. The average extent of the geometries in a dataset can be approximated by means of the average area of the MBR (Minimum Bounding Rectangle) containing them and this parameter can be easily computed during the index construction.

The impact of the average geometry extent on the spatial join becomes particularly relevant when the two datasets are very unbalanced in terms of extent and size. Let us consider the case in which one dataset contains few simple geo-spatial objects with a large extent (possibly covering almost the whole reference space), while the other one is instead huge and contains a large number of geo-spatial objects with a small extent. The first dataset may be possibly stored in a single split, since only few vertices are required for describing the shapes of all objects, while the second dataset requires more splits to accommodate the numerous objects it contains. As a first example, let us consider the case illustrated in Fig. 1, where dataset D_s contains only seven polygons representing the Australian States (Fig. 1.a) while dataset D_r contains several complex linestrings representing the main road elements of the Australian transportation network (Fig. 1.b). A generic partitioning of the two datasets which is based only on their size in bytes will produce only one partition for D_s , since the whole set of geometries can fit in one split, while several different partitions will be built for D_r . In this case, a spatial join operation on them ($D_s \bowtie_{int} D_r$) can be divided into several tasks, but each one of them will work on a split of D_r and on the single global split of D_s , thus the Cartesian product is computed and no pruning effect is obtained by using the index. This means for example that all geometries in the cell with label c in the Fig. 1.b will be tested for intersection with all the states of Australia. Clearly, an efficient use of a local index can improve the performance and avoid some useless tests, but it will not affect the number tasks to be instantiated. Moreover, the problem worsens as the extent of the geometries in D_s enlarges, covering at the end the whole space.

The aim of this paper is to formalize and evaluate the problem discussed above and further explained in Sect. 1.1, called here *parallel execution of unbalanced spatial join*, in order to identify the characteristics that really represents the complexity of spatial data, making them “big”. In particular, Sect. 2 provides a formalization of the problem and a discussion of the limits of the current spatial join algorithms available in SpatialHadoop when applied to unbalanced cases. Sect. 2.4 illustrates by means of some experiments the behavior of spatial join algorithms when applied on synthetic datasets with increasing unbalanced characteristics. Then, Sect. 3 proposes a new approach for dealing with unbalanced spatial join that requires the implementation of an alternative kind of repartition which is based on the geometry extent instead of on the file size. In Sect. 4 some additional experiments show the effectiveness of the proposed approach when applied both to synthetic and real datasets in the execution of unbalanced spatial joins. Finally, Sect. 6 summarizes the obtained results and proposed some future work.

■ **Table 1** Some metadata about two real-world datasets representing the taxonomy of the soil usage (`cv_land`) inside the Basilicata region and its extent (`tot_reg`), respectively. The average number of vertices and the average extent area refer to each single geometry in the dataset.

Dataset	size	#splits	#obj	#vert ^{avg}	area ^{avg} (squared meters)
<code>cv_land</code>	1.5 (Gb)	12	913,428	70	10,550 (1e4)
<code>tot_reg</code>	263 (Kb)	1	1	8,000	10,589,998,917 (1e10)



■ **Figure 2** (a) Dataset `cv_land` with its grid. (b) A zoom on one cell of the `cv_land` grid. (c) Dataset `tot_reg` with its grid.

1.1 Motivating Example

The problem discussed in this paper originated from a real-world case regarding a collection of datasets about a region in Southern Italy, called Basilicata. In particular, we consider two datasets: the first one, called `cv_land`, contains several geometries representing the taxonomy of land usage inside the region, while the second one, called `tot_reg`, contains one object representing the whole territory of Basilicata. Tab. 1 reports some metadata of the two datasets: they greatly differ on the number of objects, their complexity (average number of vertices in each geometry), and their extents (average area of each geometry). The aim of the original task was to perform a qualitative evaluation by verifying the satisfaction of some spatial integrity constraints. In particular, one test has to check if the set of geometries belonging to `cv_land` represents a geometric partition of the whole territory of Basilicata. As shown in [12], the execution of this check by means of a sequence of SQL queries takes several days when executed in a PostgreSQL+PostGIS environment. Therefore, the introduction of a parallel execution has become soon necessary. One of the required query in the above cited sequence coincides with the spatial join between the two datasets. Given two datasets D_1 and D_2 , the spatial join determines the pairs $(d_1, d_2) \in D_1 \times D_2$ with an intersecting extent. This operation is usually performed exploiting a plane-sweep like algorithm, in order to reduce the number of required comparisons. Clearly, the case considered in this paper is particularly challenging, since as the extent of a geometry increases w.r.t. the other one, the number of comparisons increases. Similarly, the complexity of each comparison increases as the number of vertices describing each geometry becomes greater.

Fig. 2 illustrates the two datasets with the partitioning induced by the grid index of SpatialHadoop; the number of splits only depends on the dataset size in bytes, so `cv_land` is subdivided into 12 splits (Fig. 2.a), while `tot_reg` is contained into 1 split (Fig. 2.c). Moreover, datasets `cv_land` contains a great number of objects (see a zoom in Fig. 2.b), while the complexity of `tot_reg` is given by the average number of vertices in each geometry.

■ **Table 2** Comparison between the different spatial join algorithms provided by SpatialHadoop. The number of produced pairs is 913,428. The last two algorithms make use of indexes, but their performance are not greatly increased w.r.t. the first algorithm which works on non-indexed data.

Algorithm	# maps	Effective time (min)	Heap usage (MB)	HDFS reads (MB)	HDFS writes (MB)
DJNI	12	92.57	21.87	1.57	243.64
DJGI	18	88.73	30.90	1.66	243.64
DJRE	18	80.06	24.77	1.66	243.64

Tab.2 reports some data about the execution of the spatial join using the three main algorithms provided by SpatialHadoop, the distributed join with no index (DJNI), the distributed join with grid index (DJGI) and distributed join with repartition (DJRE), which will be briefly discussed in Sect.2.3. Notice that the time required to perform the join is very high and the execution does not benefit so much from the use of index.

2 Problem Statement

This section formalizes the problem presented in Sect.1 by discussing in details how data is traditionally partitioned in MapReduce environments (Sect.2.1) and how such techniques are adapted in SpatialHadoop for implementing spatial indexes (Sect.2.2). Finally, we introduce the problem of performing a spatial join and how this operation can be effected by the use of a spatial index (Sect.2.3), anticipating some limits of a size-based partitioning technique that will be discussed in more details in Sect.2.4.

2.1 Data Partitioning in MapReduce

Hadoop divides the input of a MapReduce job into fixed-size pieces called *splits* and creates one map task for each split. Each map task executes the user-defined (map) function on each record in its split. The main idea behind the MapReduce paradigm is that the time required to process each split individually is smaller than the time required to process the whole input. Therefore, the more such computation on each individual split can be performed in parallel, the more the process performance increases. The split size is generally set equal to the size of an HDFS (Hadoop Distributed File System) block, which is 128 Mbytes by default.

The partitioning of data into splits is a crucial operation for obtaining well balanced map tasks [3, 14, 13]. In particular, if the splits can be analyzed in parallel, the whole job is better balanced when the splits are small, since a faster machine will be able to process proportionally more splits during the map execution than a slower machine, while unbalanced tasks can frustrate the benefit of the parallelism, since a single heavy task can delay the end of the whole job. This observation tends to produce the conclusion that the smaller are the mappers the more the effective execution time of the job can be reduced; however, if the splits are too small, the overhead of managing the splits and creating map tasks begins to dominate the total job execution time. Thus, a tradeoff should be defined and the reference size of 128 Mbytes is the usual choice. Moreover, the partitioning of data is usually applied randomly and this might produce balanced tasks for uniformly distributed datasets, but not in general. In order to address this problem, when spatial data are analyzed, the introduction of auxiliary structures (indexes) is an option. Using a spatial index implies that a criteria based on spatial properties (i.e., closeness) will be used for grouping the records in the same split. The general structure of a spatial index is presented in the following subsection.

2.2 Spatial Indexes in SpatialHadoop

As discussed in the Sect. 1, SpatialHadoop has two level of indexes [5]: a global and a local one. The *global index* determines how data is partitioned among nodes, while the *local index* determines how data is stored inside each block. The construction of a global index g on a input dataset D causes that D is stored as a set of data files each one containing the records spatially belonging to one cell (or partition) of the grid g . More specifically, given a dataset representing the input data, a directory named `dataset.<index>` will be created containing several files: `_master.<index>`, `part-00000`, `part-00001`, `part-00002`, and so on, where `<index>` denotes the kind of global index (e.g., grid, quadtree, rtree). File `_master.<index>` represents the global index and it has one row for each partition containing the boundaries of the partition and the partition file name (e.g., -179.32, -54.93, 6.92, 71.28, `part-00000`). All the other files are data files containing the data records. For a grid index, each partition file is simply a text file containing one record for each row, conversely for a R-tree it has a more complex structure subdivided into two sections: the first one contains the tree structure in binary format, while the second one contains the data records.

As discussed in [5] despite the particular kind of index, the number n of desired partitions is computed considering only the file size and the HDFS block capacity which are both fixed for all partitioning techniques. Subsequently, the space is subdivided into n partitions and each record in the input dataset is assigned to one or more of them. Dependently on the fact that the index admits replication or not, geometries crossing partition boundaries can be assigned to more than one partition or to exactly one, respectively. The number of partitions n used for performing the subdivision is crucial in the identification of the number of mappers that will be executed in order to produce the result. As we will see in Sect. 2.4, if the determination of such number is computed considering only the file size and the HDFS capacity, we can obtain strange behaviours, as the one anticipated in Sect. 1.

2.3 Use of Spatial Indexes in Distributed Joins

SpatialHadoop provides five different alternatives of spatial join algorithm: distributed join with no index (DJNI), distributed join with grid-based index (DJGI), distributed join with repartition (DJRE), distributed join with direct repartition (DJDR), and the MapReduce implementation of the partition-based spatial merge join (SJMR). The main differences between them are: (i) the use of indexed or not-indexed data, (ii) the possibility to repartition one of the two datasets using the global index of the other, (iii) the execution of the intersection tests on the map or on the reduce side. All operators share the use of a plane-sweep like algorithm for checking the intersections between two list of geometries. The difference mainly resides in the way the two lists are built by the various operators. In particular, as regards to the map-side joins, the cardinalities of such lists and their composition directly depends on the content and size of each partition.

Since in this paper we are interested in analyzing the impact of data partitioning (i.e., global index) in spatial operators, such as the spatial join, we concentrate only on the map-side joins which exploit the use of indexes during the join computation. Therefore, in the following section we start by considering the behaviour of DJNI and DJGI in presence of unbalanced datasets, then we evaluate the possible positive effects of a repartition of the smaller dataset (in size) by evaluating the behaviour of DJRE. However, in all these algorithms, the extent of geometries and the geometry complexity (in terms of number of vertices) are not considered during the partition process. In DJNI the partition is performed considering only the size in bytes and the constraint of splits capacity, thus records are

grouped randomly in the necessary number of splits. In case of the DJGI, the datasets are indexed (i.e., partitioned) considering again the split capacity constraint, so that partitions have a homogeneous size in terms of bytes, but records are grouped according to their spatial closeness, which is evaluated in the space the geometries are embedded in. In case of the DJRE, one dataset is indexed while the one (usually the smaller in size) is repartitioned using the grid of the bigger one. The effect is that geometries of the smaller dataset are partitioned using the spatial closeness principle, but producing splits with potentially less records (i.e., size less than 128 Mbytes) and consequently reducing the cost of the map tasks.

All these spatial join variants perform the join inside the map tasks: each map receives a combined split built by a special reader that matches a split of the first dataset with a split of the second one. Moreover, in DJGI and DJRE a combined split is built combining only pairs of input splits that intersect (through the use of a filter). Therefore, the number of map tasks which will be instantiated is equal for DJGI to the number of intersecting partitions of the two global indexes, while for DJRE it is equal to the number of partitions of the bigger datasets that intersect the smaller one. Given a combined split, each mapper initially split its content into two lists (one for each dataset) and then executes a plane-sweep like algorithm on them in order to identify the pairs of intersecting geometries.

As discussed in [1], the cost of this plane-sweep phase depends on three factors: (i) the cardinality of the two lists (which depends on the partition size), (ii) the mutual dataset selectivity (which depends on the average extent/MBR of the geometries in the two datasets), and (iii) the average number of vertices of the geometries in the two datasets.

► **Definition 1** (Plane-sweep cost). Given two lists of geometries $n_i \subseteq D_i$ and $n_j \subseteq D_j$ coming from two input datasets D_i and D_j , whose geometries have an average number of vertices equal to v_i and v_j , respectively, and a selectivity $\sigma(A)$ computed w.r.t. a certain reference space A , the complexity of the plain-sweep phase can be formulated as:

$$ps(n_i, n_j, v_i, v_j, A) = n_i \log(n_i) + n_j \log(n_j) + (v_i + v_j) \log(v_i + v_j) \cdot n_i \cdot n_j \cdot \sigma(A) \quad (1)$$

where the first two components represent the sorting the two input lists, while the last component is due to the intersection test between pairs of geometries with intersecting MBRs. The selectivity $\sigma(A)$ can be estimated by applying the following formula, proposed in [10]:

$$\sigma(A) = \frac{1}{A} \cdot (area_{mbr}^{avg}(D_i) + area_{mbr}^{avg}(D_j) + len_x^{avg}(D_i)len_y^{avg}(D_j) + len_x^{avg}(D_j)len_y^{avg}(D_i)) \quad (2)$$

Eq. 2 requires that some estimates about the datasets content are available, in particular: the average area of the MBR of the geometries belonging to D_* ($area_{mbr}^{avg}(D_*)$) and the average length on the x axis and y axis of the same MBRs ($len_x^{avg}(D_*)$, $len_y^{avg}(D_*)$).

Introducing the necessary coefficients of proportionality for each operation, Eq. 1 provides an estimate of the cost of each mapper involved in the spatial join computation. Let us analyze the case of a sequential execution (“one task” case) and compare it with the three map-side spatial join algorithms provided by SpatialHadoop, DJNI, DJGI and DJRE. We can conclude that the benefits induced by the application of one of the MapReduce spatial join derive not only from the parallel execution of different portions of the whole job, but also as a consequence of the non linear behavior of the plane-sweep algorithm.

► **Observation 1** (Benefits of parallel execution of spatial join with DJNI). Given two datasets D_i and D_j in the reference space of area A , with cardinality N_i and N_j and an average number of vertices equal to V_i and V_j , respectively. The cost of a “one task” execution of the spatial join can be estimated using Eq. 1. Conversely, by applying algorithm DJNI having s_i

and s_j number of splits for D_i and D_j respectively, we obtain from Eq. 1 the estimate of the cost of each map task as follows, where a_1 e a_2 are the coefficients that are necessary for taking into account the cost of comparing two MBRs during the ordering phase and the cost of testing a geometry intersection during the last phase, respectively:

$$ps_{\text{DJNI}}\left(\frac{N_i}{s_i}, \frac{N_j}{s_j}, V_i, V_j, A\right) = a_1 \frac{N_i}{s_i} \log\left(\frac{N_i}{s_i}\right) + a_1 \frac{N_j}{s_j} \log\left(\frac{N_j}{s_j}\right) + \\ a_2(V_i + V_j) \log(V_i + V_j) \cdot \frac{N_i}{s_i} \cdot \frac{N_j}{s_j} \cdot \sigma(A)$$

Notice that $\sigma(A)$ does not change w.r.t the “one task” case, since the geometries in each split are randomly chosen, thus they cover the whole reference space. The cost of a map task is obviously reduced compared to the single process, in particular: (i) the ordering phases are reduced proportionally w.r.t. the input reduction with an additional cut of: $a_1 N_i \log(s_i)$ (or $a_1 N_j \log(s_j)$), (ii) the intersection testing phase is reduced by a significant factor: $s_i \times s_j$, since the number of pairs D_i considered each map task is a subset of the total amount of geometries. The total cost of DJNI is $ps_{\text{DJNI}} \cdot (s_i \times s_j)$, where $s_i \times s_j$ represents the number of mappers produced by DJNI. This is a greater cost compared to the “one task” case since the ordering phase of a split of D_i is replicated s_j times. However, under the hypothesis that we can execute in parallel all the map tasks, we can obtain a significant reduction of the effective time. Indeed, in this case the effective time will coincide with the execution time of the worst map task or to the average execution time of a map in a balanced situation.

► **Observation 2** (Benefits of parallel execution of spatial join with DJGI and DJRE). By applying the DJGI algorithm on datasets D_i and D_j , having both a grid index with a number of cells s_i and s_j , respectively, we can obtain the estimate of the cost of each map task by computing $ps_{\text{DJGI}}\left(\frac{N_i}{s_i}, \frac{N_j}{s_j}, V_i, V_j, A_{\text{cell}}\right)$ from Eq. 1. Notice that in this case the selectivity changes, since the geometries of a split are now spatially located only in a subset of the reference space, i.e. the space occupied by an index cell, namely A_{cell} (here the smallest cell of the two indexes is considered). A similar consideration holds for DJRE, even if in this case only a grid index is present, suppose the one of D_i , so s_j becomes the number of cells of s_i that intersect D_j .

The cost of each map task is reduced also for DJGI and DJRE. In particular, while the cost of the ordering phases is reduced as for DJNI, the intersection testing phase is more expensive since the selectivity is lower. This is the effect of the index that tends to balance the work among the map tasks and to reduce their number. The total cost can be obtained for DJGI by multiplying ps_{DJGI} by the factor $s_i \times \sigma(s_j)$, where $\sigma(s_j)$ is the number of cells of the index of D_j that are intersected by a cell of D_i , and for DJRE by the factor $\rho(s_i)$, where $\rho(s_i)$ is the number of cells of D_i that intersects D_j . In both cases it is in average less than the cost of the “one task” execution.

By applying the formulas in Def. 1 and Obs. 1-2 to the example in Sect. 1.1, we obtain as expected a lower cost for DJNI, DJGI and DJRE w.r.t. the “one task” case. However, we can also observe that with one big geometry the index does not have a significant impact on the execution time: both DJGI and DJRE do not reduce the cost of join w.r.t DJNI.

The following section shows the results of some experiments that have been performed with the aim to test how the number of geometries, the number of vertices and the selectivity can affect the effectiveness of the index partitioning in increasing the performance of a spatial join in MapReduce. As we will see, these factors can contribute in different ways, and their effect is not completely represented by the size in bytes of each split.

■ **Table 3** Metadata about the two datasets used for the experiments on MBR size. Notice that **# VertPerGeom** is the number of vertices describing each of the **# Geometries** in the dataset.

Dataset	Size	# Splits	# VertPerGeom	# Geometries	MBR ext
D_i	1152 (MB)	9	1,000	2,156	1e-8
D_j	1 (MB)	1	1,000	40	1 \rightarrow 1e-4

2.4 Experimental Analysis of the Problem

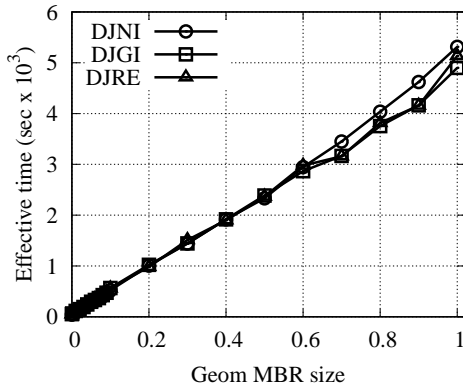
The previous section discusses the fact that a partitioning technique based only on the size in bytes of the input dataset does not properly capture the complexity of a spatial join operation, which instead depends on three factors (number of geometries, number of vertices and selectivity) that can be independent from the input size. More specifically, given a split s with a predefined size in Megabytes, its content can be very different: it can contain many simple geometries with a restricted number of vertices, or it can contain few complex geometries described by a huge number of vertices, again such geometries can have a very different extent which does not depend on the number of vertices. While the number of geometries contained in a split directly depends on the average number of vertices used to describe a shape, the extent is an independent aspect. Therefore, this section presents two kinds of experiments both performed by keeping constant the size in bytes of the two input datasets: (i) the average extent (MBR) of the second dataset is progressively augmented, producing a decrease of the selectivity (more join pairs), (ii) the number of vertices of the second dataset is progressively augmented, producing also a decreasing in the number of geometries contained inside a split.

2.4.1 Variation on the MBR Size

The first set of experiments tries to study the effect of the average geometry MBR size (namely the selectivity) on the join performance. In particular, we consider two synthetic datasets D_i and D_j with uniform distribution, D_i contains a huge number of geometries with a small extent, and D_j contains few geometries with a big extent. During the experiments the extent of geometries in D_j has been varied from 1 to 1e-4 w.r.t. to the overall dataset extent, namely initially the geometries occupy all the reference space, and this occupation is progressively decreased till a ratio of 1e-4. Tab.3 reports some metadata about the two datasets, such as the number of geometries and the number of vertices in each geometry.

We compute the spatial join between these datasets by considering the three algorithm variants presented in Sect. 2.3. Fig. 3 reports the results of such experiments. As you can notice, the time required to perform the spatial join depends linearly on the selectivity (i.e., the average MBR size of the geometries) as predicted by the formulas presented in Obs. 1. Moreover, the performances of the three algorithms are very similar to each other and this proves that a partitioning technique that takes care only of the size in bytes does not capture the real complexity of the dataset. Indeed, given the same input size, in this experiment the performances of the spatial join considerably worsen passing from a couple of minute to more than one hour. While the time remains acceptable in the first cases (till an MBR area of D_j equal to 1e-1), it becomes incredibly bad when the geometries of D_j occupy the whole reference space. In this last case, none of the available partitioning techniques are able to completely exploit the parallelism and the benefit of a MapReduce framework.

Coming back to the real-world case illustrated in Sect. 1.1, in Fig. 4 we consider a join between the dataset `tot_reg` and a polygon with an increasing MBR and a constant number



■ **Figure 3** Effective time taken by the three considered spatial join algorithms by varying the MBR size of dataset D_j from an area of $1e-4$ to 1 w.r.t. the area of the reference space.

tot_reg MBR	#Vert PerGeom	Join size	DJRE (sec)
$5e-2$	1,000	61,440	106
$1e-1$	1,000	104,934	145
$1.5e-1$	1,000	179,188	220
$2.2e-1$	1,000	262,311	329

■ **Figure 4** Effective time taken by the DJRE algorithm applied to the real-world case by varying the MBR size of the dataset `tot_reg`.

■ **Table 4** Metadata about the two datasets used for the experiments on the number of vertices. # VertPerGeom is the number of vertices describing each of the # Geometries in the dataset.

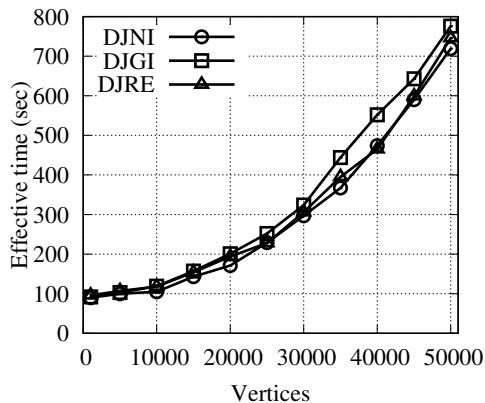
Dataset	Size	# Splits	# VertPerGeom	# Geometries	MBR ext
D_i	1152 (MB)	9	1,000	2,156	$1e-8$
D_j	$1 \rightarrow 75$ (MB)	1	$1,000 \rightarrow 50,000$	40	$1e-2$

of vertices (i.e., 1,000). In particular, the average MBR is changed from a radius of 15Km to a radius of 35Km, respectively. Again the time required for performing the join linearly depends on the MBR size, namely the resulting selectivity.

2.4.2 Variation on the Number of Vertices

The second set of experiments evaluates the effect of the number of vertices on the complexity of the spatial join. In particular, we consider two datasets D_i and D_j with a fixed size in terms of occupied splits and containing geometries with a fixed extent. In order to check how the number vertices affects the time required for performing the spatial join, we vary the number of vertices of all geometries in D_j , while maintaining constant the number of occupied splits, the number of geometries and their average MBR area. Tab. 4 shows some metadata about the considered datasets. In particular, for D_j the number of vertices describing each geometry is varied from 1,000 to 5,000.

Fig. 5 presents the result of these experiments, again the use of the partitioning induced by the spatial index does not considerably increase the performance of the join: the difference between the three algorithms is no more than few minutes and in some cases the execution time is less for the join without index than for the other twos. In this case the number of map tasks to be executed is the same for all three algorithms: DJNI considers a number of combined splits equal to the Cartesian product, namely 1×9 splits, which is equal to the number of combined splits computed by DJGI, since the 1 split of D_j intersects all splits of D_i . Similarly, for DJRE the repartition does not discard any cell of D_i ; moreover, the costs of performing a repartition phase is not compensated by its benefits, the number of repartitioned geometries is so very small. The trend of the curves in Fig. 5 completely adheres to the formulas presented in Obs. 1.



■ **Figure 5** Effective time taken by the three considered spatial join algorithms by varying the number of vertices of each geometry in dataset D_j from 1,000 to 50,000.

tot_reg # vert	Geom MBR	Join size	DJRE (min)
1,000	1e-1	103,683	13
5,000	1e-1	103,704	60
10,000	1e-1	103,707	118
15,000	1e-1	103,710	174
20,000	1e-1	103,707	231
25,000	1e-1	103,708	288
30,000	1e-1	103,707	342
35,000	1e-1	103,707	408
40,000	1e-1	103,707	463
45,000	1e-1	103,707	519
50,000	1e-1	103,707	578

■ **Figure 6** Effective time taken by the DJRE algorithms by varying the number of vertices of each geometry in dataset tot_reg from 1,000 to 50,000.

Referring to the real-world case introduced in Sect. 1.1, we evaluate the performance of DJRE by varying the number of vertices in dataset tot_reg while maintaining constant the characteristics of dataset cv_land . The results are reported in Fig. 6, again the time by the spatial join increases with the number of vertices in each geometry.

3 Proposed Solution and Discussion

Considering the experimental results presented in the previous section, we can conclude that the effective execution time of three spatial join algorithms, DJNI, DJGI and DJRE, are affected by both the selectivity of the datasets, which directly depends on the MBR area of the geometries they contain, and the average number of vertices of the same geometries. Indeed, in the experiments the size in bytes of the input file remained unchanged, while the selectivity and the number of vertices are varied, obtaining very different execution times. However, the partitioning techniques provided by SpatialHadoop are only based on the size in bytes of the input files, thus they cannot react to the variations of these parameters.

In order to avoid this problem we propose an alternative partitioning technique to be applied during the existing index building phase. Given the grid to be used for grouping the geometries of a dataset, the existing indexing phase scans the whole dataset and for each geometry g detects the subset of cells $S(g)$ that it intersects, then g will be inserted in the split of each cell in $S(g)$. This means that sometimes a geometry g can be replicated in more than one split. If g is relatively small w.r.t. the index cells, then the replication is not frequent, but when geometries are bigger, the repetition occurs more frequently. The replication rate has not been considered in the cost estimation (see Eq. 1), since it can be neglected in the considered experiments.

The proposed technique enriches the indexing phase with a splitting operation which should affect the partitioning result in particular when the MBR area increases or the number of vertices increases. In the first case, i.e. lower selectivity, we can split the geometries that cross two or more index cells, so that the average area of their MBR is reduced and thus their contribution to the join selectivity is reduced (see Eq. 2). Notice that in this case the number

of tested geometries does not change (in the original approach the whole geometry will be replicated in all combined splits), so the cost of a map task is reduced according to Eq. 1. In the second case, i.e. higher number of vertices, we can split the geometries when the number of their vertices exceed a threshold. In this way, a big geometry g can be substituted by a set of smaller geometries $\{g_1, \dots, g_n\}$, that represent a partition of g and have a number of vertices smaller than the number of vertices of g .

In order to combine the two cases, we consider a new partitioning technique where the dataset with the bigger (in terms of average area of their MBR) geometries are splitted by considering the grid of the other dataset. This *splitting phase* reduces both parameters discussed above, thus reducing the cost of the map tasks. The following proposition shows the effective cost reduction that the splitting phase introduces.

► **Observation 3** (Benefits of the splitting phase). Consider two datasets D_i and D_j in a reference space of area A , with cardinality N_i and N_j and an average number of vertices equal to V_i and V_j , respectively, and such that D_j is the dataset having the bigger geometries in terms of occupied area. The cost of the “one task” execution of the spatial join can be estimated by Eq. 1. Conversely, if we consider the application of DJGI on D_i and D_j in presence of grid indexes having respectively a number of cells s_i and s_j , and assuming that the geometries in D_j have been splitted so that they are spatially contained in one cell of D_i , the cost of each map task can be estimated from Eq. 1 as follows:

$$p_{SD_{DJGI}} \left(\frac{N_i}{s_i}, \frac{\alpha N_j}{s_j}, V_i, \frac{V_j}{\alpha}, A_{cell} \right) = a_1 \frac{N_i}{s_i} \log \left(\frac{N_i}{s_i} \right) + a_1 \frac{\alpha N_j}{s_j} \log \left(\frac{\alpha N_j}{s_j} \right) + a_2 \left(V_i + \frac{V_j}{\alpha} \right) \log \left(V_i + \frac{V_j}{\alpha} \right) \cdot \frac{N_i}{s_i} \cdot \frac{\alpha N_j}{s_j} \cdot \sigma(A_{cell}) \quad (3)$$

where α represents the average number of cells of D_i that are intersected by a geometry of D_j , namely the average number of small geometries obtained from each big geometry in D_j after the splitting phase. Accordingly to [1], it can be estimated as follows: $\alpha = \lceil \text{len}_x^{avg}(D_j) / \text{len}_x^{cel}(D_i) \rceil \cdot \lceil \text{len}_y^{avg}(D_j) / \text{len}_y^{cel}(D_i) \rceil + \beta$, where, considering the MBR of the geometries belonging to D_j , $\text{len}_x^{avg}(D_j)$ ($\text{len}_y^{avg}(D_j)$) is the average length on the x (y) axis of these MBRs, while $\text{len}_x^{cel}(D_i)$ ($\text{len}_y^{cel}(D_i)$) represents the average length on the x (y) axis of the index cells of D_i . β is an additional factor taking into account the displacement between MBRs and cells, namely it is a function of the probability that the MBR of a geometry of D_j crosses the boundaries of the cells of D_i .

Notice that, as shown in Obs. 1, the cost of a map task is obviously reduced compared to the “one task” case, in particular: (i) the ordering phases are reduced proportionally w.r.t. the input reduction with an additional cut quantifiable in: $a_1 N_i \log(s_i)$ for D_i (or $a_1 \alpha N_j (\log(\alpha) - \log(s_j))$ for D_j), (ii) the intersection testing phase is also reduced in two ways: by the reduction of the pairs of geometries to be considered, with a factor $(\alpha \cdot \sigma(D_j) / s_j)$, and also by the reduction of the cost for testing the intersection between two geometries, since the number of vertices is decreased by a factor α .

4 Validation of the Solution

This section presents some additional experiments that verify the theoretical behavior of the algorithms when the new splitting technique described in Sect. 3 is applied. In particular, we first consider the experiments related to the variation of the average MBR size (i.e., selectivity) and check the effect of splitting the geometries of D_j using the grid of the constant dataset D_i of size 9 splits. These results are reported in Tab. 5 where the first three columns

■ **Table 5** Comparison between the execution time of the three distributed join algorithms when performed considering the original and the modified synthetic datasets with a variable MBR size.

D_j Avg MBR size			DJNI		DJGI		DJRE	
Orig	Part	% Decr.	(sec)	% Improv.	(sec)	% Improv.	(sec)	% Improv.
1e-1	2.27e-2	77.25%	491	5.66%	527	7.17%	527	5.66%
2e-1	3.08e-2	84.58%	941	6.71%	934	8.72%	945	3.94%
3e-1	3.66e-2	87.79%	1,223	15.20%	1,285	10.95%	1,280	15.31%
4e-1	4.35e-2	89.12%	1,752	8.58%	1,633	14.91%	1,705	10.33%
5e-1	4.40e-2	91.19%	2,008	13.86%	1,814	23.99%	1,747	26.17%
6e-1	4.93e-2	91.79%	2,553	13.35%	2,586	9.81%	2,259	24.11%
7e-1	5.75e-2	91.78%	3,009	12.78%	2,650	16.28%	2,665	15.67%
8e-1	6.58e-2	91.77%	3,573	11.49%	3,570	4.86%	3,311	13.21%
9e-1	7.41e-2	91.76%	4,352	5.83%	4,021	3.37%	3,629	12.71%
1e+0	8.21e-2	91.79%	4,558	14.22%	4,473	8.73%	4,746	7.60%
Average			10.77%		10.88%		13.47%	

■ **Table 6** Comparison between the execution times of the three distributed join algorithms when performed on the original and the modified synthetic datasets with a variable number of vertices.

D_j # VertPerGeom			DJRE	
Original	Splitted	% Reduction	(sec)	% Improvement
1,000	149	85%	303	5%
10,000	1,545	84%	397	20%
20,000	2,924	85%	753	57%
30,000	4,512	85%	1,239	72%
40,000	6,020	85%	2,123	78%
50,000	8,272	83%	3,194	84%
Average			85%	
			52%	

contain: (a) the original MBR size, (b) the MBR size after the splitting and (c) the average percentage of decrease in the MBR size. The area of the used grid cells is $8.21e-2$, while the average percentage of decrease in the MBR size increases as the average MBR area of the original geometries increases. For each algorithm the table reports the execution time on the splitted geometries and the percentage of improvements w.r.t. the original situation. All three versions of the distributed join benefit from the partitioning with an average improvement of around 10-13% with respect to the previous executions.

As second set of validation experiments we consider the case in which the geometry MBR remains unchanged but we vary the number of vertices describing each geometry. In particular, we consider only the case of DJRE since the execution time of the various algorithms are not much different from each other and DJRE is on average the most efficient one. These results are reported in Tab.6 where the first three columns contain: (a) the original number of vertices in each geometry, (b) the number of vertices after the splitting (c) the average percentage of decrease in the number of vertices. The other two columns contain the execution time of DJRE on the splitted geometries and the percentage of improvement w.r.t. the original situation. The results of these experiments confirms what verified in Sect.2.4, namely the not negligible effect of the number of vertices on the spatial join execution time. Indeed, this time greatly decreased by decreasing the complexity of the geometries in terms of the average number of vertices in each geometry.

5 Related Work

A common strategy to reduce the cost of a spatial join is the filter-and-refine approach which consists on a filter phase that traditionally works on the MBR (minimum bounding rectangle) of the involved geometries, and a refining step which performs the actual test on the filtered pairs. The filter phase can usually benefit from the use of a spatial index while the identification of both the overlapping MBRs or geometries is performed using the plane-sweep algorithm [10]. In [17] the authors analysed the problem of how to partition spatial data in order to perform parallel spatial join. They promoted the use of spatial locality in task decomposition in order to speed-up the join computation. This partitioning reflects the way data is partitioned by SpatialHadoop during the construction of a global index. However, it is not effective in the case considered in this paper, since we assume the presence of some big and complex geometries which occupy the whole reference space. In the context of parallel spatial join execution, some research has been done in order to define partitioning techniques which produce balanced partitions even in presence of skewed data [4, 9, 11]. This paper does not consider the effect of the data distribution (skewed or uniform), but concentrates on the presence of big and complex geometries that do not allow to completely exploit the parallelism induced by the MapReduce approach.

In [5] the author analysed the various partitioning techniques available in SpatialHadoop and they experimentally studied the effect of such indexes on some operations, such as the range query and the spatial join. The work mainly evaluates such partitioning techniques based on four quality measures, but it assumed that the considered objects occupy a small space w.r.t. the reference space, so it did not consider the problem treated in this paper. The problem of processing big complex geometries together with small ones has been investigated for the first time in [12], where the author detected a difference in the performance of some Pigeon operations when performed on spatially equivalent datasets with different configurations for what regards the extent and complexity of the involved geometries. Pigeon [6] is an extension of Pig Latin for dealing with spatial data in SpatialHadoop.

6 Conclusion

This paper deals with the problem of identifying the characteristics that really represents the complexity of spatial data, making them “big” w.r.t. the most common operations. These characteristics have to play a central role in the definition of an effective partitioning technique able to exploits all potentiality of a MapReduce environment, like Hadoop. Traditionally, in such environments the partitioning of data is performed by subdividing the records in the original datasets so that each obtained split has an upper bound size given in terms of the number of occupied bytes. This kind of partitioning is used also in spatial-aware MapReduce systems, like SpatialHadoop, where the data partitioning, even the one induced by the construction of spatial indexes, is driven only by the data size in bytes. However, spatial data is characterized by other kinds of dimensions, such as the number of vertices (complexity) used to described a single geometry, or the average area of the MBR of the geometries (extent). These characteristics usually affect the cost of spatial analysis operations, such as the spatial join. Therefore, we can assume that what makes spatial data big is not only their size in bytes, but also their complexity and their extent. In order to validate such hypothesis, in this paper we analyse the behaviour of some distributed spatial join algorithms provided by SpatialHadoop when varying the average MBR size and the number of vertices, showing how such characteristics affect the performance of the spatial join and that they are not correctly captured by a partitioning technique based only on the size in bytes of the

input datasets. We propose the idea of a new partitioning technique which takes care of such characteristics by also performing a splitting of the original geometries in order to reduce their complexity and better exploit the parallelism induced by a MapReduce environment. Further improvements will regard the identification of the grid which is more appropriate on the base of the average MBR size of geometries and the average number of vertices.

References

- 1 A. Belussi, S. Migliorini, and A. Eldawy. A Cost Model for Spatial Join Operations in SpatialHadoop. Technical Report RR108/2018, Dept. of Computer Science, University of Verona, 2018. URL: <https://iris.univr.it/handle/11562/981957>.
- 2 Alberto Belussi, Sara Migliorini, Mauro Negri, and Giuseppe Pelagatti. Validation of spatial integrity constraints in city models. In *4th ACM SIGSPATIAL Int. Workshop on Mobile Geographic Information Systems*, pages 70–79, 2015. doi:10.1145/2834126.2834137.
- 3 Matteo Dell’Amico, Damiano Carra, and Pietro Michiardi. PSBS: Practical size-based scheduling. *IEEE Transactions on Computers*, 65(7):2199–2212, 2016.
- 4 D. J. DeWitt, J. F. Naughton, D. A. Schneider, and S. Seshadri. Practical skew handling in parallel joins. In *18th Int. Conf. on Very Large Data Bases*, pages 27–40, 1992.
- 5 A. Eldawy, L. Alarabi, and M. F. Mokbel. Spatial partitioning techniques in SpatialHadoop. *Proc. VLDB Endow.*, 8(12):1602–1605, 2015.
- 6 A. Eldawy and M. F. Mokbel. Pigeon: A spatial MapReduce language. In *IEEE 30th Int. Conf. on Data Engineering*, pages 1242–1245, 2014. doi:10.1109/ICDE.2014.6816751.
- 7 A. Eldawy and M. F. Mokbel. SpatialHadoop: A MapReduce framework for spatial data. In *2015 IEEE 31st International Conference on Data Engineering*, pages 1352–1363, 2015.
- 8 A. Eldawy and M. F. Mokbel. *Spatial Join with Hadoop*, pages 2032–2036. Springer International Publishing, Cham, 2017. doi:10.1007/978-3-319-17885-1_1570.
- 9 K. A. Hua and C. Lee. Handling data skew in multiprocessor database computers using partition tuning. In *17th Int. Conf. on Very Large Data Bases*, pages 525–535, 1991.
- 10 Edwin H. Jacox and Hanan Samet. Spatial Join Techniques. *ACM Trans. Database Syst.*, 32(1), 2007. doi:10.1145/1206049.1206056.
- 11 Masaru Kitsuregawa and Yasushi Ogawa. Bucket spreading parallel hash: A new, robust, parallel hash join method for data skew in the super database computer (SDC). In *16th Int. Conf. on Very Large Data Bases*, pages 210–221, 1990.
- 12 S. Migliorini, A. Belussi, M. Negri, and G. Pelagatti. Towards massive spatial data validation with spatialhadoop. In *5th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*, pages 18–27, 2016. doi:10.1145/3006386.3006392.
- 13 Giovanni Neglia, Damiano Carra, Mingdong Feng, Vaishnav Janardhan, Pietro Michiardi, and Dimitra Tsigkari. Access-time-aware cache algorithms. *ACM Transactions on Modeling and Performance Evaluation of Computing Systems (TOMPECS)*, 2(4):21, 2017.
- 14 Mario Pastorelli, Damiano Carra, Matteo Dell’Amico, and Pietro Michiardi. HFSP: bringing size-based scheduling to hadoop. *IEEE Trans. on Cloud Computing*, 5(1):43–56, 2017.
- 15 Tom White. *Hadoop: The Definitive Guide*. O’Reilly Media, Inc., 4th edition, 2015.
- 16 Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, and Ion Stoica. Apache Spark: A unified engine for big data processing. *Commun. ACM*, 59(11):56–65, 2016. doi:10.1145/2934664.
- 17 Xiaofang Zhou, David J. Abel, and David Truffet. Data partitioning for parallel spatial join processing. *GeoInformatica*, 2(2):175–204, 1998.

Intersections of Our World

Paolo Fogliaroni

Vienna University of Technology, Austria
paolo.fogliaroni@geo.tuwien.ac.at

Dominik Bucher

ETH Zurich, Switzerland
dobucher@ethz.ch

Nikola Jankovic

Vienna University of Technology, Austria
nikola.jankovic@geo.tuwien.ac.at

Ioannis Giannopoulos

Vienna University of Technology, Austria
igiannopoulos@geo.tuwien.ac.at

Abstract

There are several situations where the type of a street intersections can become very important, especially in the case of navigation studies. The types of intersections affect the route complexity and this has to be accounted for, e.g., already during the experimental design phase of a navigation study. In this work we introduce a formal definition for intersection types and present a framework that allows for extracting information about the intersections of our planet. We present a case study that demonstrates the importance and necessity of being able to extract this information.

2012 ACM Subject Classification Information systems → Geographic information systems, Information systems → Data analytics

Keywords and phrases intersection types, navigation, experimental design

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.3

1 Introduction

The street network of a city is a physical artifact embedded in the natural world. Most of the times, it consists of highways (i.e., streets meant for cars only), roads (meant for cars and pedestrians) and pathways (only for pedestrians). Sometimes these networks are following strict human design guidelines and sometimes they are bounded by natural constraints. Along with historical rationales, these constraints are the primary reasons that not all parts of a city follow a gridded design structure (e.g., curvilinear). This means that beside commonly encountered 3- and 4-way intersections, also more complex ones can exist.

But what are the main implications of this diversity of streets and intersections, and why is it important to know how a city, a country or even a continent are structured? What can we learn from this information and how can this information be useful?

In the following we will exemplify our work focusing on the area of navigation studies and experimental design. Independently of the research discipline, when planning an experiment there is a certain process that is followed in order to come up with a correct design. At the very beginning, information for the various relevant variables is collected that eventually will help to make the right choices.

In the case of navigation experiments, the relevant variables concern the subjects (e.g., gender or age), the type of navigation aid [13] and the timing of instructions [12], if any,



© Paolo Fogliaroni, Dominik Bucher, Nikola Jankovic, and Ioannis Giannopoulos;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 3; pp. 3:1–3:15

Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

and the environment (e.g., the route). When it comes to the environment, the relevant factors that have to be considered are numerous [15] and decisions can be made by taking into account possible interactions between the relevant subjects and the environment – e.g., previous experience of the subjects with the environment. Besides factors such as architectural differentiation and environmental landmarks [30], the types of intersections are highly relevant since they contribute to the complexity of a wayfinding decision [15]. A typical question during the design process is how the decision points along the designated route should be selected in terms of number of choices. How many and what kind of crossroads should the route encompass? Of course, the number of crossroads and their shape (e.g., T- or Y-intersections) on an experimental route is strongly related to the underlying research questions.

The aim of this work is to help answering this type of questions. We computed the number and type of all intersections on Earth and developed a web application that can be easily used to extract this precomputed information for any area in the world. Of course this work is not limited to navigation and experimental design. Next to researchers of various disciplines, industries related to the areas of transportation and urban planning can use our work for their decision making processes. For instance, by comparing the intersections of a street network between two areas, interesting correlations with other phenomena could be made, allowing to draw conclusions regarding the impact of the intersection types.

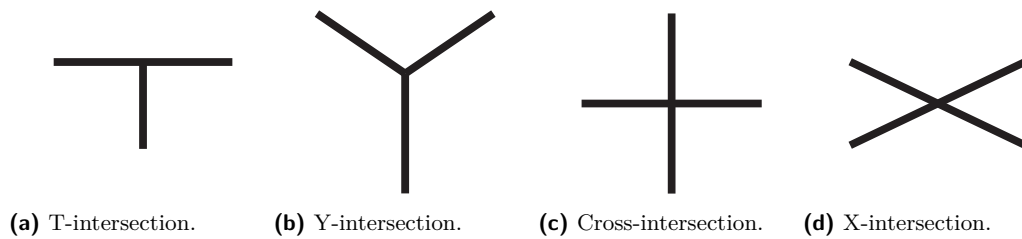
As a data source for our work we resorted to OpenStreetMap (OSM), that is one of the most commonly used source of volunteered geographical information (VGI). While approach we present does not require any particular form of road network data, the wide and free availability as well as the generally good quality of OSM [16] make it an adequate choice for intersection analysis. OSM data was analyzed in a multitude of studies before, not only in terms of quality and completeness [18], but also as a data source for answering questions about various environments [9], to determine the distribution of landmarks and points of interest [31, 28, 3], to build recommender systems [4] or as contextual enhancement for other types of data, such as Twitter posts [17].

In terms of intersection analysis, most previous work focuses around the automatic detection of roads and intersections from other sources of data, such as GPS traces [10, 7] or satellite imagery [6]. A variety of techniques exist, where intersection types are either implicitly learned using machine learning techniques (such as neural networks for satellite image analysis) [27, 32], or considered directly within the model [25]. To the best of our knowledge, in all of the automated detection methods the individual intersections are not classified in any way except based on the number of roads that lead up to them.

Intersections also play a central role in many routing applications [24]. Not only do red lights (commonly occurring at intersections) influence the driving time, behavior, and related emissions [23, 2], but even the difference between a right or left turn at an intersection incurs different penalties to route computations [20]. In addition, vehicular ad hoc networks (VANETs, which are used for inter-car communications) optimally also take intersections into consideration, as they provide data exchange points for cars driving on different routes and cars are likely to stop there [5, 1].

2 Types of Intersections

While the terms *junction* and *intersection* are commonly used interchangeably to refer street joints and crossings, they have slightly different meanings, with the term *intersection* referring to a specific type of junctions. According to the Oxford Dictionary, a *junction* is a place



■ **Figure 1** The most common types of prototypical named intersections.

where two or more roads or railway lines meet, while an *intersection* is a point at which two or more things intersect, especially a road junction.

The term *junction* unambiguously relates to the mobility infrastructure domain and denotes roads coming together but does not specify the exact nature of their connection (intersect, touch, meet at a square, etc.). Conversely, the term *intersection* has a broader scope – as it can refer to several domains. Yet, when it comes to the mobility infrastructure domain it clearly refers to the cases where two or more roads intersect with each other.

Intersections are mostly studied in the areas of Architecture, Civil and Traffic Engineering, as well as Urban Planning. Studies in these domains are concerned with intersection design and construction to optimize traffic load, road safety, and traveling time (e.g., [29]). Intersections are typically split into two main categories: *at-grade* and *grade-separated* (see, e.g., [8]). At-grade intersections consist of roads located at the same level (grade), while the roads creating a grade-separated intersection are at different levels (grades) and pass above or below each other. Grade-separated intersections are mostly used in highways and motorways, as they allow for a faster and smoother merging of car traffic but are not well suited for pedestrian navigation.

Both categories can be more finely classified. Grade-separated intersections can be divided into *interchanges* and *grade-separations without ramps*. Subcategories of at-grade intersections include *proper intersections*, *roundabouts*, and *staggered (or offset) intersections*, among others. Proper intersections are the most prototypical type of intersection for the layman: several road segments converge to meet at the same point. Roundabouts are circular intersections that cars can enter and exit smoothly and in which road traffic flows in a single direction. In Staggered intersections several (minor) roads meet a main road at a slight distance apart such that they do not all come together at the same point.

In the scope of this work we only take into consideration proper intersections and, marginally, staggered intersections (that we regard as a composition of proper intersections). The analysis of more more complex types of intersections such as, e.g., roundabouts will be investigated in future work.

In the following we will introduce relevant terminology and discuss properties of proper intersections. The most straightforward property to classify intersections is the number n of street segments stemming out of it. We call such street segments the *branches* of the intersection. An intersection I with n branches is called an *n -way intersection* and we denote it by I^n . Obviously, we need at least two street segments to meet in order to form an intersection. In this work we focus on the intersections which call for navigational decision making: given one street segment that is used to approach an intersection, there have to be at least two more street segments that can be used to leave that intersection (i.e., $n \geq 3$).

A second discriminant that we use to classify an intersection is its shape. That is, the angular arrangement of its branches. Typically, this is done by comparing the intersections at hand to some others that are generally accepted as prototypical ones [33, 22, 26, 14]. The

most common ones are reported in Figure 1: they are called T- and Y-intersections for $n = 3$; cross- and X-intersections for $n = 4$. Every intersection with more than four branches ($n > 4$) is typically referred to as a star-intersection.

There is evidence that these named intersection types are used very naturally by people when communicating route instructions verbally [33, 22] or schematically [33, 26, 14]. However, they suffer from two major drawbacks. First, these namings only exist for intersections with a small number of branches ($n \leq 4$). Second, they are often not precisely defined: for example, while most people would agree that a cross-intersection splits a revolution into four right angles, there might be a large disagreement on the skewness of an X-intersection.

For these reasons, we introduce the concept of *regular intersection*, whose branches divide a revolution into uniform parts. More formally:

► **Definition 1** (Regular n -way intersection). Let b_0, \dots, b_{n-1} be the branches of an n -way intersection enumerated in circular order. We define α_i as the angle formed by the pair $(b_{i-1}, b_{i \bmod n})$ for every $i \in \mathbb{N}$ such that $1 \leq i \leq n$. We say that a n -way intersection is *regular* if and only if $\alpha_1 = \alpha_2 = \dots = \alpha_n = 360/n$ and we denote it by R^n .

In general, to further characterize an n -way intersection we compare it to its regular counterpart, rather than to the aforementioned named intersection types. However, it has to be noted that regular 3- and 4-way intersections can be interpreted as exact definitions for Y- and cross-intersections, respectively. The arbitrary skewness of X-intersections makes them unsuitable to be taken as an objective reference for comparison. T-intersections, on the other hand, are well defined. For this reason, for 3-way intersections we also perform a comparison to T- intersections.

Finally, we define the angular distance $\Delta(I^n, R^n)$ among a generic n -way intersection I^n and its regular counterpart R^n as the minimum sum of angles that we have to rotate the branches of I^n to perfectly match R^n , while preserving the circular order of I^n 's branches. Note that there are $n - 1$ possible rotations that can be performed to match I^n to R^n (see Sec 3.2 for more details).

3 Intersections Framework

In the following, we present our framework that was implemented for the classification and analysis of intersections. As one of the goals was to make worldwide intersection data available, the presented framework is based on OpenStreetMap data and is publicly available¹. The framework is able to periodically process this data and writes the resulting intersection measures into a database, where they can be accessed through a web application.

3.1 Data Source

OpenStreetMap (OSM) is arguably one of the largest and most important volunteered geographic information (VGI) projects. As VGI is often not only the cheapest source of geographic information, but even the only one available in certain regions [16], it is an agreeable data source for a global intersection analysis. It needs to be noted that even though OSM data quality can be considered adequate for many purposes, its spatial distribution is not uniform, but depends on factors such as the information of interest or social events (e.g., an upcoming Football World Cup) in a region [18, 19, 11]. However, these quality

¹ See <http://intersection.geo.tuwien.ac.at>.

Analysis Class	Highway Tag Values	Description
Road	<i>living_street, primary, secondary, tertiary, unclassified, residential, service, primary_link, secondary_link, tertiary_link</i>	All ways that can be traversed by <i>both cars and pedestrians</i> , namely all normal roads.
Path	Road highway tag values plus <i>path, steps, bridleway, footway, track, pedestrian</i>	All ways that can be used by <i>pedestrians</i> . Including smaller tracks, hiking routes, etc. where cars cannot drive.
Car	Road highway tag values plus <i>motorway, motorway_link, trunk, trunk_link</i>	All ways that can be traversed by <i>car</i> . This additionally includes highways and motorways, where pedestrian access is usually forbidden.

■ **Table 1** Different highway tag values used within the intersection analysis framework.

issues often concern single newly built roads or geographical information unrelated to the road network, which make up for a negligible amount of data with respect to a regional intersection analysis.

The three primary data structures of OSM are *nodes*, *ways* and *relations*. Nodes represent single points in space (i.e., they have a *longitude* and *latitude*), such as points of interest or individual objects. Ways are ordered lists of nodes, and encode linear features (like roads or rivers) and boundaries of areas (when the first and last node are equal). Finally, relations describe relationships between multiple elements, e.g., a collection of ways which form a scenic *route*, or turn restrictions, which state that you cannot cross from one way into another at a certain intersection.

All the node, way and relation objects can have an arbitrary number of *tags*, which have a simple *key* → *value* form (both *key* and *value* are arbitrary strings). The tags themselves are not formally specified, but are chosen based on a consensus in the OSM community. For example, the very common tag *highway* is assigned to way objects which can somehow be used for travel, e.g., for walking or driving. It can take the values described in Table 1². Note that we distinguish between three analysis classes, one with ways solely accessible to pedestrians, and another two with ways accessible to cars (including resp. excluding motorways). To find intersections in the OSM data, it suffices to look at ways that carry a *highway* tag, and to determine which nodes are shared among several of these ways.

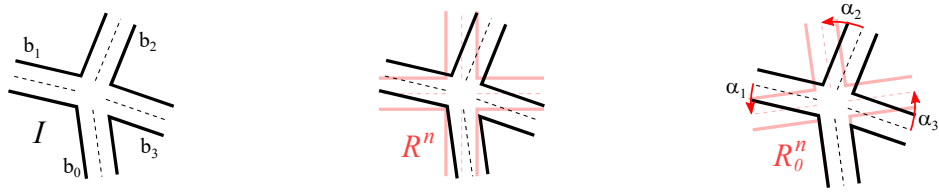
OSM data is available in different formats. As the whole uncompressed xml planet file is around 850 GB at the time of this writing, we opted for the protocol buffer binary format (PBF) instead, which is available as a 40 GB gzipped file³ and consists of around 4.3 billion nodes and 470 million ways.

3.2 Data Processing

After uncompressing the PBF file, we first search for nodes that should be considered intersections. As stated above, this corresponds to nodes which have more than two branches ($n \geq 3$). For each way in the OSM dataset that has one of the appropriate *highway* tag values

² For a detailed description of the individual values, and also additional ones that are not used in this framework, please consult the OSM documentation under wiki.openstreetmap.org/wiki/Key:highway.

³ For details see wiki.openstreetmap.org/wiki/PBF_Format.



(a) Original 4-way intersection I with branches b_0 - b_3 . (b) Overlay of perfect 4-way intersection R^4 . (c) Angles between original and perfect intersection.

■ **Figure 2** Computation of $\Delta(I^n, R^n)$, the sum of all angles that each branch b_i has to be rotated in order to produce a regular n -way intersection. Note that it suffices to align the regular intersection with each branch (as is done for b_0 in (c)), and take the minimal Δ of all possible alignments.

(cf. Table 1), we iterate through all the nodes making up this way, and build a mapping that stores all neighboring nodes of each node. To be able to distinguish the different analysis classes later on, the highway tag value is additionally stored for each neighboring node. In essence, we define intersections as a function mapping a center node p to a number n of adjacent nodes $p_{p,i}$, where for each $p_{p,i}$ in addition the highway tag value $t_{h,i}$ of the connecting way is stored:

$$I^n : p \mapsto \{(p_{p,i}, t_{h,i}) \mid 0 \leq i < n\} \tag{1}$$

As this is done for all nodes in the OSM dataset (irrespectively of n), in a second iteration, a final set of intersections $\{I_0, \dots, I_k\}$ has to be built by removing all nodes that dissatisfy the minimal number of branches condition (i.e., $|I(p)| < 3$). This set of intersections contains all the relevant OSM nodes for the purposes of the here presented framework. To compare each intersection to its regular counterpart (in the case of a 3-way intersection additionally to a perfect T-intersection), it is required to compute all angles between the different roads in a next step.

Thus, for the remaining intersections, a second pass through the OSM data collects the coordinates of the center node p itself, as well as the coordinates of all the neighboring nodes $p_{p,i}$ that can be reached by traversing its branches b_i . Using these coordinates, it is possible to compute all angles between the branches and the meridian passing through the center node. Figure 2 depicts a hypothetical 4-way intersection in black and, beneath it, the regular 4-way intersection, where the angles between branches are always 90° . In order to compute the angular distance $\Delta(I^n, R^n)$ to the regular intersection, we rotate the regular intersection n times, so that it always aligns perfectly with one of the branches b_i . Figure 2c shows one of the four possible alignments for a 4-way intersection. For each non-aligned branch, α_i denotes the required rotation to reach an alignment with the next “free” branch of the regular intersection (in this respect, “free” simply means that no two branches of the original intersection may be rotated to the same branch of the regular intersection). For any alignment with a branch of the regular intersection, a Δ' is computed as the sum of all α_i . The final Δ takes the value of the minimal Δ' over all n possible alignments. Note that this is a globally minimal Δ , even if arbitrary rotations of the regular intersection were allowed (and not just “snapping” to branches of the original intersection), as rotating the regular intersection monotonically increases or decreases Δ , until another alignment is reached. As such, all minima and maxima of Δ must occur at an alignment with the regular intersection.

All the intersections with their coordinates, the number of branches, as well as the computed $\Delta(I^n, R^n)$ are finally written to a PostGIS database⁴. Since it is required to know the analysis class of each intersection, an additional database field denotes if an intersection is valid for *road*, *path*, and *car*, or only any subset thereof.

3.3 Data Service

We provide public access to the intersection data computed with our framework through a web application that is accessible at intersection.geo.tuwien.ac.at.

The interface provides a map canvas with OSM as a basemap that can be used to freely browse the whole globe. With the current release of the application, the user is provided with a selection menu from where she can specify the type of intersections of interest (column “Analysis Class” in Table 1). We plan to extend this in future releases to allow the selection of combinations of the base intersection types.

We offer three possibilities to specify the region of interest: polygon drawing, viewport, and name search. In the first case, the user can specify a region by drawing a polygon on the map. With the canvas selection, the viewport currently shown on the map canvas is used to perform the database query. Finally, it is possible to look for named entities via a search box that provides a live interface to an OSM Nominatim⁵ server. After typing in the name of the searched feature, the user can ask the interface to draw the corresponding polygon on the map. Given the huge amount of intersection data available, we decided to limit the area of the search region to not overload the server. In future releases this limitation might be removed. Also, in order to promote interoperability, we plan to include the possibility of specifying custom geometries expressed in different type formats (e.g., KML, geoJSON, etc.).

The intersection type and the region specified are used to submit a query that returns a statistical summary for intersections of the given type in the provided region. This summary contains the number of occurrences for each n -way intersection, the average Δ from the corresponding regular intersection – for 3-ways, also the average Δ from the regular T intersection. Besides the statistical summary the user is also provided with a link to download the whole intersection data set for the specified region and type as a CSV file.

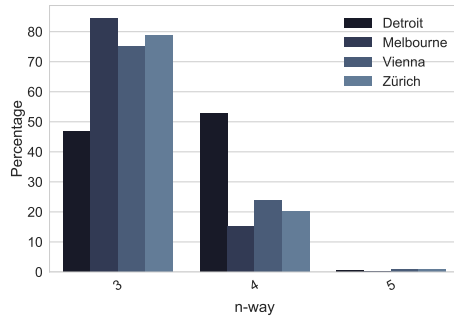
At the time of writing the CSV file only contains information about the intersection points that were computed from OSM nodes. Beside the geometric information (reported in WKT) each point is associated the following attributes: the *number of branches* and the *type* of intersection, and the *angular distance* Δ to the corresponding regular intersection.

Note that the intersection classes defined in Table 1 are not disjoint. This results in the same intersection occurring up to three times in our database, once for each category. Imagine the case of an intersection where both roads and paths converge. For example, we may have 3 roads and 1 path. This intersection appears twice in our dataset: as a path and as a road. Since roads are accessible by both pedestrians and cars but paths are only accessible by cars, we have a 4-way path intersection and a 3-way road intersection. A similar concept applies to the categories of road and car intersections. The relation of the number of ways (denoted as n_{class}) between the intersections that overlap is $n_{car} \geq n_{road}$ and $n_{path} \geq n_{road}$.

In our database we also keep trace of the ways that form intersection branches: their geometry (also converted to OGC standard), the original OSM highway tag, and a relation to the intersections that they generate. This information is not accessible through the current version of the application, but will be made available in future releases.

⁴ PostgreSQL 9.6 with PostGIS 2.3.2, the processing application is implemented in Rust 1.23.0.

⁵ Nominatim is a search engine for OSM data, see wiki.openstreetmap.org/wiki/Nominatim.



■ **Figure 3** Distribution of the intersections as the number of branches n varies.

	Detroit	Melbourne	Vienna	Zürich
3-way	46.76%	84.49%	75.16%	78.88%
4-way	52.84%	15.20%	23.74%	20.13%
5-way	0.36%	0.29%	0.93%	0.82%
6-way	0.04%	0.02%	0.13%	0.14%
7-way	0.002%	0.002%	0.02%	0.02%
8-way	0.002%	-	0.01%	0.004%
10-way	-	-	-	0.004%
Total	40929	191508	75644	26286

■ **Table 2** Distribution of intersections over number of ways for the four cities.

4 Use Case: Detroit, Melbourne, Vienna, and Zürich

In this section we present and discuss intersection data obtained with our framework for four exemplary cities and showcase how this data can be used during the design process of navigational experiments. In Section 4.1 we compare the four different cities, while in Section 4.2 we focus on local differences within a single city.

4.1 Comparative Study

We used our framework to extract intersection data for Detroit (USA), Melbourne (Australia), Vienna (Austria), and Zürich (Switzerland). While the framework allows for extracting intersection data concerning different types of streets (cf. Section 3.1), for this case study we focus on *paths* and *roads* (i.e., set of all walkable streets).

Table 2 reports the distribution (as percentages) of intersections as the number of branches n varies. From this data we can derive several interesting insights. First and foremost it has to be noted that for all the cities in exam almost the entirety of intersections are 3-ways and 4-ways. This becomes even more evident by looking at the graphical representation of the data reported in Figure 3. While this fact may seem trivial, it is still surprising the cumulative percentage that these two intersection categories reach together – ranging from 98.9% for Vienna to 99.7% for Melbourne. This pattern seems to recur everywhere in the world. Indeed, we found it in many other cities (Athens, Rome, Kathmandu, Washington DC, Paris, and London, among others) that we analyzed with our framework in a preliminary analysis for this work. This pattern consistently (only with minor differences) repeats across different cities, independently of their very heterogeneous morphology, history, and age.

The second insight that we can derive from this data relates to the ratio between the number of 3-way and 4-way intersections. In this respect, we notice that Melbourne, Vienna, and Zürich present a very similar trend with the majority of intersections being 3-ways, although with slightly different ratios between the number of 3- and 4-ways: approx. 5.5 for Melbourne, 3.2 for Vienna, and 3.9 for Zürich. Conversely, Detroit shows the opposite trend, with the number of 4-ways slightly bigger than that of 3-ways. This may indicate, for example, a more blocked structure of the city.

In the following we analyze the further discriminant introduced in this work to classify intersections: the similarity to *regular intersections* (see Definition 1). As discussed in Sections 2 and 3.2, we measure this by the angular distance $\Delta(I^n, R^n)$ between a generic n -way intersection I^n and the corresponding regular intersection R^n . For the case of 3-ways,

City	Min	P_{25}	P_{50}	P_{75}	Max	City	Min	P_{25}	P_{50}	P_{75}	Max
Det	~0%	0.58%	1.96%	15.98%	99.99%	Det	~0%	0.23%	0.59%	2.11%	50.00%
Mel	~0%	1.03%	4.31%	21.59%	99.65%	Mel	~0%	0.63%	2.37%	8.05%	83.69%
Vie	~0%	1.51%	6.08%	22.16%	99.87%	Vie	~0%	0.91%	3.17%	9.69%	85.81%
Zur	~0%	2.09%	7.54%	23.48%	99.76%	Zur	~0%	1.44%	4.45%	11.41%	64.12%
Δ -range: $[0^\circ, 180^\circ]$						Δ -range: $[0^\circ, 360^\circ]$					

(a) 3-way to regular T, delta percentiles.

(b) 4-way to regular 4-way, delta percentiles.

■ **Table 3** Distribution of 3-way (4a) and 4-way (4b) intersections for the four cities (normalized).

we compare against regular T intersection instead. Moreover, given that for the cities in exam 3-ways and 4-ways combined cover almost the totality of the number of intersections, we will only focus on those.

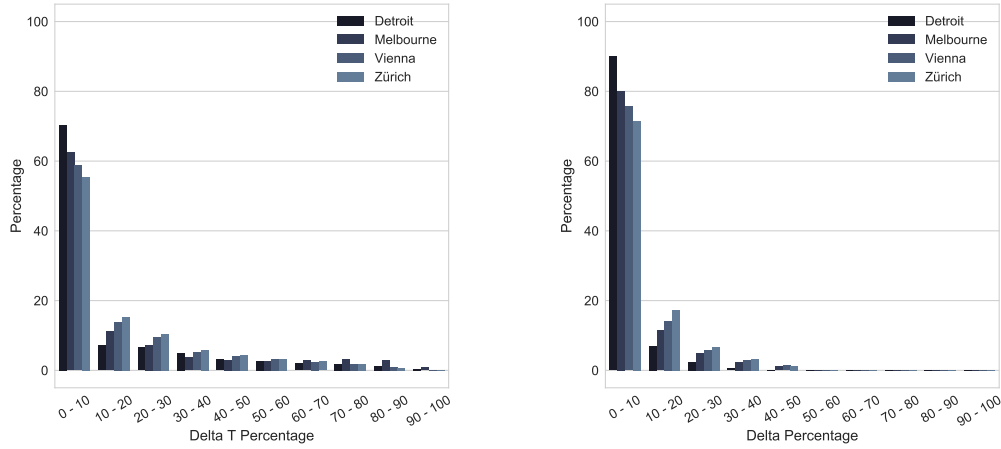
Tables 4a and 4b report descriptive statistics for 3-ways and 4-ways, respectively. The numbers reported are percentages referring to the value range that the angular distances can take on. This is called Δ -range and denotes the difference between the minimum (Δ_{min}) and maximum (Δ_{max}) angular distances from a generic intersection to its regular counterpart. Obviously, the minimum is always zero ($\Delta_{min} = 0^\circ$), which corresponds to a perfect match with the regular intersection. Conversely, Δ_{max} depends on the number of branches (n) of the intersection at hand and corresponds to the angular distance of the (theoretical) worst-case scenario where all the branches of an intersection collapse on top of each other:

$$\Delta_{max} = \sum_{i=1}^{\lfloor \frac{n-1}{2} \rfloor} (2i\alpha) + ((n-1) \bmod 2) \pi \tag{2}$$

For an understanding of this formula imagine to align any branch of the regular intersection to the first branch of the n -way at hand. Subsequently, take a pair of unmatched branches from the generic intersection and rotate them (one clockwise and the other counterclockwise) by $\alpha = \frac{360}{n}$ to match the first pair of unmatched branches of the regular intersection. Now repeat for the second pair of unmatched branches. In this case, we will have to rotate 2α in order to find the first pair of unmatched branches of the regular intersection. Generalizing this operation we obtain the formula in Equation 2. For 3-ways and 4-ways we have Δ -ranges equal to $[0^\circ, 240^\circ]$ and $[0^\circ, 360^\circ]$, respectively. The Δ -range for 3-ways when compared against the regular T intersection is equal to $[0^\circ, 180^\circ]$.

Figures 4a and 4b plot in greater details the distribution of 3-ways and 4-ways as the angular distance varies over the Δ -ranges for the regular T intersection and the regular 4-way, respectively. The figures show that the majority of the intersections are very similar to their regular counterparts (which aligns nicely with Klippel’s set of wayfinding choremes [22, 21]), with Detroit and Zürich representing extreme cases. The intersections of Detroit are the most regular, with approximately 70% of its 3-ways and 90% of its 4-ways showing an angular distance below 10% to the regular T intersection (i.e., 18°) and the regular 4-way (i.e., 36°), respectively. Conversely, Zürich is the least regular, with approximately 55% of its 3-ways and 70% of its 4-ways showing an angular distance below 10% to the regular T intersection (i.e., 18°) and the regular 4-way (i.e., 36°), respectively. Melbourne and Vienna are located in between these extremes, with Melbourne being slightly more regular than Vienna with respect to both 3-ways and 4-ways.

These findings can be used, for example, during the design of navigational experiments to select paths that adhere to the structure of the city where the experiments are to be



(a) 3-way to regular T intersection.

(b) 4-way to regular 4-way intersection.

Figure 4 Distribution of the angular distance (Δ) for 3-ways (a) and 4-ways (b) with respect to the regular T intersection and the regular 4-way intersection, respectively. The angular distance (on the x-axis) is reported as a percentage of the different Δ -ranges for 3-ways (i.e., $0^\circ - 180^\circ$) and 4-ways (i.e., $0^\circ - 360^\circ$). The percentage on the y-axis refers to the number of intersections in each bin with respect to the total number of intersections of that type (i.e., 3-way and 4-way). The smaller the value of Δ , the higher the similarity to the corresponding regular intersection.

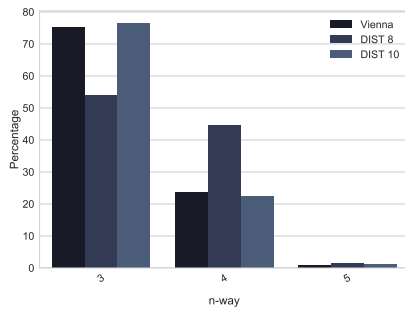
performed. In this way, we can avoid to select some *atypical* path that may lead to biased results. Assume that for our hypothetical navigational experiment we need a path that comprises 10 intersections. If we were to conduct the experiment with a path matching the characteristics of Detroit, we should select a path in the real world or in a virtual environment that encompasses, e.g., five 3-way and five 4-way intersections. Of the selected 3-ways (resp. 4-ways), three (resp. five) should present a maximum angular distance of 18° (resp. 36°) from the regular T intersection (resp. the regular 4-way). Conversely, if we were to conduct the same experiment with a path matching the characteristics of Zürich, our path should encompass eight 3-way and two 4-way intersections. Of the selected 3-ways (resp. 4-ways), four (resp. six) should present a maximum angular distance of 18° (resp. 36°) from the regular T intersection (resp. the regular 4-way).

Moreover, the availability of intersection data for the entire world easily supports comparative analysis that so far was difficult to control. Imagine to run the same spatial experiment in different cities or areas of the globe. The availability of this data may allow for comparing the different paths and, consequently, for relating and gaining insights on the possibly different experimental results obtained in different locations.

4.2 Local Differences

In this section we discuss local differences within the city of Vienna. We used our framework to run analysis on all 23 districts (DIST) and focus on the two with the highest variation, district 8 and 10.

Table 5 reports the distribution (as percentage) of the intersections as the number of branches n varies. This allows for easily comparing the statistics of the selected districts against the statistics extracted for whole Vienna. Both the selected districts adhere to



■ **Figure 5** Distribution of the intersections as the number of branches n varies.

	Vienna	DIST 8	DIST 10
3-way	75.16%	53.97%	76.46%
4-way	23.74%	44.63%	22.44%
5-way	0.93%	1.4%	1%
6-way	0.13%	-	0.07%
7-way	0.02%	-	0.03%
8-way	0.01%	-	-
Total	75644	428	7121

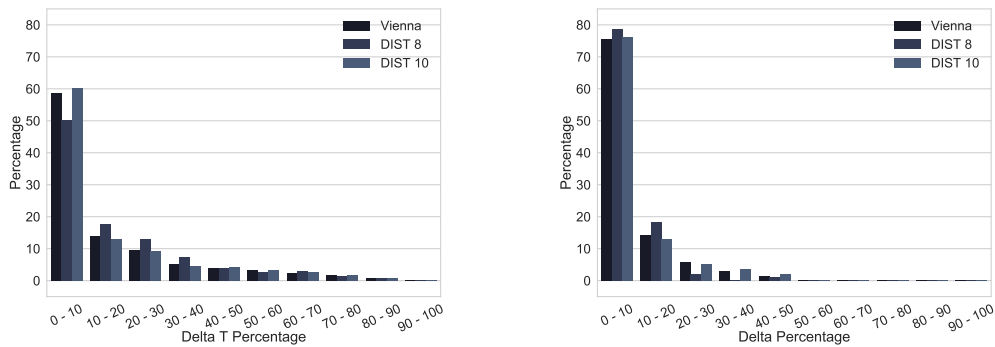
■ **Table 5** Distribution of intersections over number of ways for whole Vienna and the 2 districts in exam.

the overall distribution pattern that we discussed in Section 4.1, with almost the entirety of intersections distributed between 3-ways and 4-ways. The graphical representation of the data (see Figure 5) allows for glimpsing different local patterns for the two districts. Specifically, district 10 exhibits a distribution almost identical to whole Vienna. In contrast, district 8 exposes different distributions, with approximately 20% less 3-ways (resp. 20% more 4-ways) than whole Vienna.

The distribution of 3-way and 4-way intersections can be seen in Figures 6a and 6b as their normalized angular distance varies in the corresponding Δ -ranges – i.e., $[0^\circ, 180^\circ]$ and $[0^\circ, 360^\circ]$, respectively. As for 3-ways, district 8 is the most dissimilar with respect to Vienna, while district 10 exhibits only a small deviation from the distribution of the whole city. The same pattern emerges also for 4-ways.

Assume that we want to replicate in Vienna the navigational experiment discussed at the end of Section 4.1 for which we need to select a path encompassing 10 intersections. If we were to conduct the experiment in district 10, according to the intersection distribution reported in Figure 5, approximately 76% (resp. 22%) of these intersections should be 3-ways (resp. 4-ways). Say, for example, that we choose a path consisting of eight 3-ways and two 4-ways. According to the distribution of Δ s in Figures 6a and 6b, of the selected 3-ways (resp. 4-ways), five (resp. 2) should present a maximum angular distance of 18° (resp. 36°) from the regular T intersection (resp. the regular 4-way).

If the experiment was to be conducted in district 8 we could either decide to stick to the statistics of whole Vienna or to the statistics of the district. In the first case we would end up with a selection similar to that of district 10. In the second case we would have to choose differently. If we opt for the first alternative the findings that relate to the structure of intersections could be considered as a step towards generalization to whole Vienna but might apply more loosely to district 8. More generally, the statistical data provided by our framework can be used to find out areas all over the world that expose an intersection structure similar to that of a given area where, e.g., we performed an experiment. This information can be used to replicate the experiment in any of these areas and identify which of the insights we derive from the experiment results are invariant with respect to the intersection structure of the path.



(a) 3-way to regular T intersection.

(b) 4-way to regular 4-way intersection.

■ **Figure 6** Bar plot visualization of the distribution of the angular distance (Δ) for 3-ways (a) and 4-ways (b) with respect to the regular T intersection and the regular 4-way intersection, respectively. See Figure 4 for reading instructions.

5 Discussion and Conclusion

The framework presented in this work can be considered as an important asset during the design of spatial experiments and to perform spatial analysis. As shown through the case study in Section 4, the framework can be easily used to partially validate a selected route with respect to generalization issues. Since local differences can be found in an urban environment that do not adhere to the overall structure of a city, a country, or even a continent, the choice of a route has to be considered very carefully. Furthermore, by identifying similarities of the selected route at different scales (i.e., from district up to continent scale), one can go a step further and carefully interpret the findings of the experiment (at least those related to features of the intersection distributions) and draw conclusions concerning the reproducibility and comparison with experiments performed in different areas. Of course looking only at the intersections of a route is not sufficient, but necessary. This work can be considered as a further step towards interpreting the results of an experiment concerning generalizability aspects.

Next to the scenario used throughout this paper to exemplify how the results of this work can be utilized, this type of quantitative data can also be useful for a multitude of other purposes. For instance, machine learning approaches could profit from this framework, generating relevant features that can help to describe the spatial phenomena of interest. Another example would include work in the area of transportation, trying to model the access and demand or relevant work in the area of urban planning. Furthermore this framework could also easily be used as part of city modeling softwares, e.g., Esri CityEngine⁶, helping to automatically generate look-alike urban environments.

In this paper we presented the raw intersection data that we generated from OSM data through the procedure described in Section 3.2 and show an example of how this data can be used for the design and comparison of navigational experiments. However, according to the specific experiment at hand it might be necessary to clean the raw data in order to accommodate geometrical and perceptual aspects. We identified two cases where the raw data may need to be cleaned before usage. Both cases concern scenarios where two or more

⁶ See <http://www.esri.com/software/cityengine>.

intersections are located very close to each other. If the intersections under consideration are of the same type, this may denote a mapping issue: due to accuracy problems a single intersection in the real world is actually reported as several in OSM. Alternatively, the intersections might actually be correctly reported in OSM, but we may have a perception issue: although we physically have several intersections, they are so close to each other that a person could perceive them as a single intersection.

The other scenario concerns the case where intersections of different types are very close to each other. Specifically, we identified a somewhat problematic pattern where a road intersection is surrounded by a set of path intersections representing sidewalks and zebra stripes. In such situations, we actually have a single intersection in the real world that is identified as several by our framework. This issue is due to the fact that in OSM, sidewalks can either be mapped as separate ways or denoted with an apposite tag on the corresponding road. This means that we cannot know in advance how many times this scenario appears in our data. For this reason we performed a simple buffer and cluster analysis on Vienna to find out the amount of groups of intersections in our data that should actually be considered as a single intersection. We used buffer of different sizes (ranging from 1m to 10m) to identify clusters corresponding to both scenarios: intersections of same type and one road surrounded by path intersections. For the first scenario we found that the number of clusters ranges from 0.04% to 4.8% (resp. from 0.2% to 12.4%) of the road (reps. path) intersections, as we increase the buffer radius from 1m to 10m. For the road-to-path scenario, the number of clusters ranges 0 to 5.7% of the road and path intersections.

Finally, it has to be noted that the implementation of our framework does not compute the data on the fly from OSM data. Rather, a snapshot of the OSM database is taken and intersection data is generated from there. This means that the data provided on the website might not be completely actual, although we do not expect huge discrepancies.

6 Outlook

Since in our work we focused on regular intersections, we omitted analyses of roundabouts. In the underlying OSM data, roundabouts are modeled as multiple 3-way intersections. Although this might look correct at a first glance, one can argue that roundabouts form a category of its own, or even an n-way intersection, with n equals the number of ingoing and outgoing branches. As this is an open question that needs further investigation and probably a user study to understand how humans perceive roundabouts, we will focus on this problem in the future. Since this framework is not only intended to be used for experimental design, a possible solution could be to transfer the choice to the users of this framework, by providing multiple options on how to handle roundabouts during runtime.

Also, in this work we did not perform any scale-based aggregation of the street geometries (e.g., aggregating two lanes of a street into a single line). Therefore, the results presented in this paper are at the finest level of details allowed by data source. Street aggregation will also yield a reduction in the number of detected intersections as well as a simplification of the resulting intersection network. Future work along this direction may potentially lead to a hierarchical organization of the data that, in turn, may allow for further types of uses and analyses of the intersection data.

In future work we will also focus deeper on network patterns. For instance, what is the most common sequence of intersections for a given length (number of intersections)? What is the typical distance between intersections or intersection types (segment length)? Being able to extract this type of information will further improve the goals set in this paper, allowing to draw even better conclusions and automatically create even more realistic look-alike cities.

References

- 1 Nizar Alsharif, Sandra Céspedes, and Xuemin Shen. icar: Intersection-based connectivity aware routing in vehicular ad hoc networks. In *Communications (ICC), 2013 IEEE International Conference on*, pages 1736–1741. IEEE, 2013.
- 2 Behrang Asadi and Ardalan Vahidi. Predictive cruise control: Utilizing upcoming traffic signal information for improving fuel economy and reducing trip time. *IEEE transactions on control systems technology*, 19(3):707–714, 2011.
- 3 Mohamed Bakillah, Steve Liang, Amin Mobasheri, Jamal Jokar Arsanjani, and Alexander Zipf. Fine-resolution population mapping using openstreetmap points-of-interest. *International Journal of Geographical Information Science*, 28(9):1940–1963, 2014.
- 4 Andrea Ballatore, Gavin McArdle, Caitriona Kelly, and Michela Bertolotto. Recomap: an interactive and adaptive map-based recommender. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, pages 887–891. ACM, 2010.
- 5 Jin-Jia Chang, Yi-Hua Li, Wanjiun Liao, and Chau Chang. Intersection-based routing for urban vehicular communications with traffic-light considerations. *IEEE Wireless Communications*, 19(1), 2012.
- 6 Dragos Costea and Marius Leordeanu. Aerial image geolocalization from recognition and matching of roads and intersections. *arXiv preprint arXiv:1605.08323*, 2016.
- 7 Ole Henry Dørum. Deriving double-digitized road network geometry from probe data. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL’17, pages 15:1–15:10. ACM, 2017.
- 8 Said Easa. *Civil Engineering Handbook*, chapter Chapter 63: Geometric design. CRC Press, Boca Raton, FL, 2002.
- 9 Jacinto Estima and Marco Painho. Exploratory analysis of openstreetmap for land use classification. In *Proceedings of the second ACM SIGSPATIAL international workshop on crowdsourced and volunteered geographic information*, pages 39–46. ACM, 2013.
- 10 Alireza Fathi and John Krumm. Detecting road intersections from gps traces. In *International Conference on Geographic Information Science*, pages 56–69. Springer, 2010.
- 11 Mohammad Forghani and Mahmoud Reza Delavar. A quality study of the openstreetmap dataset for tehran. *ISPRS International Journal of Geo-Information*, 3(2):750–763, 2014.
- 12 Ioannis Giannopoulos, David Jonietz, Martin Raubal, Georgios Sarlas, and Lisa Stähli. Timing of pedestrian navigation instructions. In *LIPICs-Leibniz International Proceedings in Informatics*, volume 86. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.
- 13 Ioannis Giannopoulos, Peter Kiefer, and Martin Raubal. Mobile outdoor gaze-based geohci. In *Geographic Human-Computer Interaction, Workshop at CHI 2013*, pages 12–13. Citeseer, 2013.
- 14 Ioannis Giannopoulos, Peter Kiefer, and Martin Raubal. Gazenav: Gaze-based pedestrian navigation. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 337–346. ACM, 2015.
- 15 Ioannis Giannopoulos, Peter Kiefer, Martin Raubal, Kai-Florian Richter, and Tyler Thrash. Wayfinding decision situations: A conceptual model and evaluation. In *International Conference on Geographic Information Science*, pages 221–234. Springer, 2014.
- 16 Michael F Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221, 2007.
- 17 Stefan Hahmann, Ross S Purves, and Dirk Burghardt. Twitter location (sometimes) matters: Exploring the relationship between georeferenced tweet content and nearby feature classes. *Journal of Spatial Information Science*, 2014(9):1–36, 2014.
- 18 Mordechai Haklay. How good is volunteered geographical information? a comparative study of openstreetmap and ordnance survey datasets. *Environment and planning B: Planning and design*, 37(4):682–703, 2010.

- 19 Mordechai Haklay, Sofia Basiouka, Vyron Antoniou, and Aamer Ather. How many volunteers does it take to map an area well? the validity of linus' law to volunteered geographic information. *The Cartographic Journal*, 47(4):315–322, 2010.
- 20 Ronald F Kirby and Renfrey B Potts. The minimum route problem for networks with turn penalties and prohibitions. *Transportation Research*, 3(3):397–408, 1969.
- 21 Alexander Klippel. Wayfinding choremes. In *International Conference on Spatial Information Theory*, pages 301–315. Springer, 2003.
- 22 Alexander Klippel, Heike Tappe, Lars Kulik, and Paul U Lee. Wayfinding choremes—a language for modeling conceptual route knowledge. *Journal of Visual Languages & Computing*, 16(4):311–329, 2005.
- 23 Chunxiao Li and Shigeru Shimamoto. A real time traffic light control scheme for reducing vehicles CO₂ emissions. In *Consumer Communications and Networking Conference (CCNC), 2011 IEEE*, pages 855–859. IEEE, 2011.
- 24 Dennis Luxen and Christian Vetter. Real-time routing with openstreetmap data. In *Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems*, pages 513–516. ACM, 2011.
- 25 Helmut Mayer, Stefan Hinz, Uwe Bacher, and Emmanuel Baltsavias. A test of automatic road extraction approaches. *International Archives of Photogrammetry, Remote Sensing, and Spatial Information Sciences*, 36(3):209–214, 2006.
- 26 Tobias Meilinger, Christoph Hölscher, Simon J Büchner, and Martin Brösamle. How much information do you need? Schematic maps in wayfinding and self localisation. In *International Conference on Spatial Cognition*, pages 381–400. Springer, 2006.
- 27 Volodymyr Mnih and Geoffrey E Hinton. Learning to detect roads in high-resolution aerial images. In *European Conference on Computer Vision*, pages 210–223. Springer, 2010.
- 28 Eva Nuhn, Wolfgang Reinhardt, and Benjamin Haske. Generation of landmarks from 3d city models and osm data. In *Proceedings of the AGILE'2012 International Conference on Geographic Information Science, Avignon, France*, pages 24–27, 2012.
- 29 American Association of State Highway and Transportation Officials. A policy on geometric design of highways and streets, 2011.
- 30 Martin Raubal and Stephan Winter. Enriching wayfinding instructions with local landmarks. In *International conference on geographic information science*, pages 243–259. Springer, 2002.
- 31 Kai-Florian Richter and Stephan Winter. Harvesting user-generated content for semantic spatial information: The case of landmarks in openstreetmap. In *Proceedings of the Surveying and Spatial Sciences Biennial Conference*, pages 75–86, 2011.
- 32 Shunta Saito and Yoshimitsu Aoki. Building and road detection from large aerial imagery. In *Image Processing: Machine Vision Applications VIII*, volume 9405, page 94050K. International Society for Optics and Photonics, 2015.
- 33 Barbara Tversky and Paul U Lee. Pictorial and verbal tools for conveying routes. *Spatial information theory. Cognitive and computational foundations of geographic information science*, pages 51–64, 1999.


Considerations of Graphical Proximity and Geographical Nearness

Francis Harvey

Leibniz Institute for Regional Geography and University Leipzig, Schongauerstr. 9, 04275

Leipzig, Germany

f_harvey@ifl-leipzig.de

 <https://orcid.org/0000-0003-0027-8320>

Abstract

“Near things are more similar than more distant things” states Tobler’s first law of geography. This seems obvious and is part to much cognitive research into the perception of the environment. The statement’s validity for assessments of geographical nearness purely from map symbols has yet to be ascertained. This paper considers this issue through a theoretical framework grounded in Gestalt concepts, behavioral ecological psychology and information psychology. It sets out to consider how influential experience or training may be on the association of graphical proximity with geographical nearness. A pilot study presents some initial findings. The findings regarding the influence of experience or training are ambiguous, but point to the rapid acquisition of affordances in the survey instruments as another factor for future research.

2012 ACM Subject Classification Human-centered computing → Empirical studies in visualization

Keywords and phrases proximity, nearness, perception, cognition

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.4

Acknowledgements Thanks to many colleagues at IfL who discussed various aspects of this paper over the course of its development, especially Lea Bauer, Natalia Ipatow and Eric Losang, I want to especially thank Marie Windhorst for her help with the literature review. The helpful comments from anonymous reviewers were of great help in clarifying and refining the presentation of the research.

1 How is geographical nearness related to graphical proximity

Tobler’s first law of geography states it plainly: near things are more similar than more distant things [18, 19, 20]. It has become an anchor of a general understanding of space/time phenomena and a primary reference in teaching and research. It expresses a truth about our perceptions of the environment around us [15] that is a cardinal rule for evaluating geographic information. But a question remains to be asked: How does it apply to people’s perception of nearness using only map symbols? The answer is surprising: We do not know. There has been just no internationally published research to-date that assesses how the graphical proximity among map elements corresponds to geographical nearness in Tobler’s sense. This paper draws on socio-cognitive approaches from psychology since the early 20th century that considers how people rely on both cognitive and social faculties and knowledge in spatial comprehension. It offers a theoretical foundation for the exploratory study of how people understand the graphical proximity of spatial representations – geovisualizations, usually maps. This question is relevant for GIScience and the many daily and emergency applications of geographical information. Geovisualizations make up the dominant form for the representation of spatial information of geography [11]. Improving map-based geovisualization tools and augmented



© Francis Harvey;

licensed under Creative Commons License CC-BY

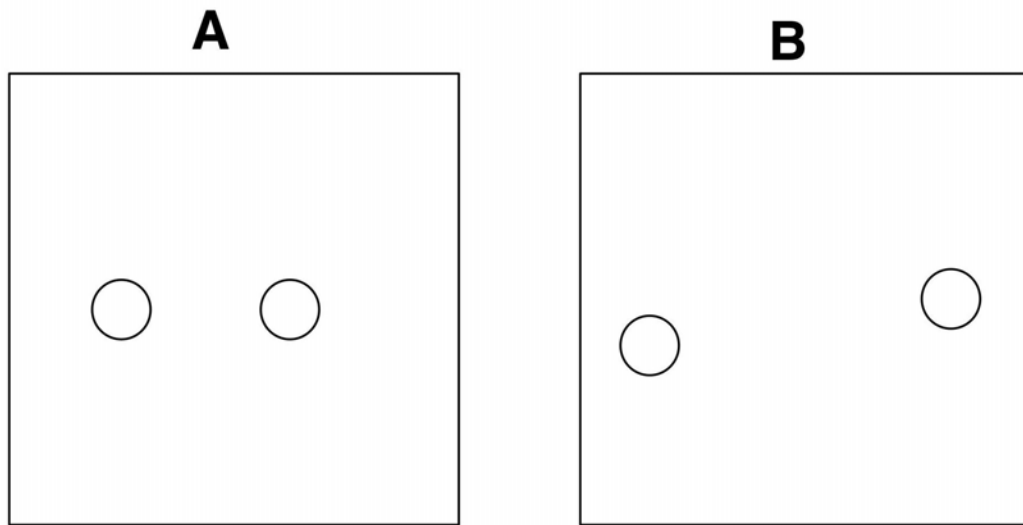
10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 4; pp. 4:1–4:18

Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

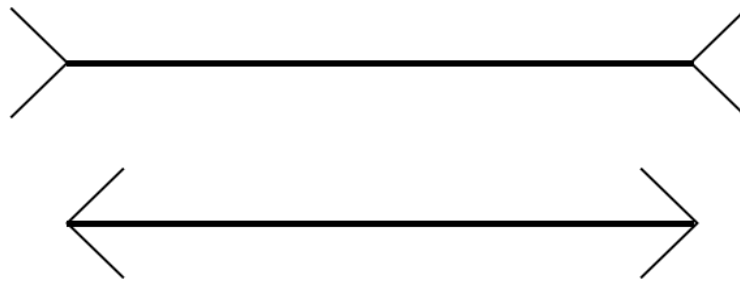


■ **Figure 1** Already recognized over 100 years ago in the first Gestalt studies our minds identify without conscious effort graphic elements that are closer. This A/B situation of object pairs near/further is used in the pilot study reported later in this paper

reality used for mobility and navigation presents multiple possibilities for improvements. Errors and distortions are also commonplace. To show things and events [10] at all geographic scales, from an aerial photo of an urban block to an animation showing wind speeds and directions around the earth, we rely on graphical representations. Maps, with their evolving meanings that reflect changes in technologies and media [3, 13, 4], remain GIScience's most common form of representation. The graphical proximity of things or events seems to preattentively (perception before conscious knowledge) indicate a geographical nearness. But does that perception and the following understanding come from intuitive understanding alone? How do training and experience impact the perception of graphical proximity and its translation into geographical nearness? A common understanding is that we see proximity immediately and intuitively – which means even before we think about what we see our mind assess the image and is aware of their proximity [26].

It is powerful and useful capacity of the human mind's visual faculties. However, to place the abstract question's relevance more clearly for GIScience, how does the preattentive perception of proximity among elements in a geovisualization reliably correspond to geographic nearness? We learn perhaps much of this, but how much of this is nurture and how much is nature? Indeed, the visual perception of the mind can be readily fooled. Illusions (such as shown in Figure 2) highlight that our perception is continuously subject to biases. Cognitive and social psychology have since the Gestalt psychology studies almost 100 years ago begun to shed considerable insight into the perception of graphical phenomena. From these studies, we know that proximity is a vital aspect of the mind's faculty in preattentively comprehending the world around us. This finding is evident from behavioral studies of visual perception. However, the many studies have focused very little attention looking into how graphical proximity corresponds to an understanding of geographical nearness.

Based on research into socio-cognitive processes of visual perception, this paper sets out to offer a tentative framework and some initial empirical evidence for the relationship between graphical proximity and geographical nearness to suggest the importance of distinguishing



■ **Figure 2** Example of preattentive vision capabilities prone to biases. We see that the upper line is longer, but both lines are actually the same length. The Müller-Lyss illusion shown here belongs to the many visual illusions studied over more than 100 years. This well-known illusion was first published 1889 by Franz Müller-Lyer in *Optische Urteilstäuschungen*. *Archiv für Physiologie Supplemental* pp. 263–270.

graphical perceptions of nearness from visual perceptions of nearness in GIScience research. The focus of this paper is on the conceptual review with some preliminary study data that offers a basis for further research and a test of an empirical Bayesian approach to analyzing experimental data. This paper is an initial foray into a complex area that multiple fields of science (psychology and neuroscience perhaps most commonly) have considered over more than 100 years. The concepts for this paper are rooted distinctly in research associated with behavioral psychology but follow concepts from informational psychology [5]. Presentation of these concepts make up a significant share of this paper and determine its structure. The next section of the paper describes the theoretical background in work on socio-cognitive studies of spatial perception and reviews research that refines the Gestalt concept of proximity. The following section provides a detailed presentation of the methods used in this study. The results of a small test are presented, analyzed and reviewed in the next section. The conclusion summarizes the findings of this study and points to a future avenue of research to better understand how people perceive graphical proximity in geovisualizations and comprehend it as geographical nearness.

2 Some Background: Nearness, Proximity, Biases and Ecological Psychology

Considering how people understand geographical nearness through representations that render geographical things and events as graphical elements and compositions build on psychological/behavioral research conducted over more than the past 100 years. Summarizing the breadth and depth of that research lies plainly beyond the scope of this paper. It provides a more limited literature review of relevant work in GIScience and cognitive psychology with some pointers to older seminal research. Central concepts of Gestalt, behavioral, ecological and informational psychology complement this work by taking up the concepts of affordance, visual clustering/patterns in visual comprehension, and pre-attentive patterns in that broader sense. Informational psychology takes this research and situates it in a contemporary information processing framework. The emphasis of this article on the connection of a theoretical framework to an empirical study also facilitates development in further studies.



■ **Figure 3** Figure 5 from Wertheimer's 1923 paper shows one of the more simpler configurations that illustrate the Gestalt concept of proximity. Thirty additional figures increase complexity and broaden the focus to issues of patterns and isomorphism in his search for laws for the whole (Figure available at <http://psychclassics.yorku.ca/Wertheimer/Forms/forms.htm>).

2.1 Tobler's law review

The concept behind Tobler's law is to develop a maximum of scientific utility from the simplest as possible statement [20]. Much has been considered about the philosophical issues implicit in this approach. For the intents of paper, drawing on research published by the first Gestalt theorists, the emphasis on simplicity is seen to have considerable value in its lucidity. As Tobler's law of geographical nearness appears to find valence for many users of geovisualizations, the parallel is relevant in seeking to evaluate the question how perceptions of graphic proximity correspond to the understanding of geographical nearness. This research focuses on understanding how people relate graphic proximity to geographical nearness in the use of geovisualizations.

2.2 Gestalt principle of proximity review

To ground an empirical understanding of graphical proximity and geographical nearness, graphical proximity, a central Gestalt concept, proximity, then in the positivist spirit referred to as a law, but now seen more as a rule, requires some review. Widely known work by Wertheimer, Koffka, and Köhler on Gestalt proximity belongs to the foundational work on Gestalt theories. The concepts have since been studied and further elaborated in neuropsychology [27]. Wertheimer's seminal work on Gestalt concepts, begun in 1912 and primarily published in the 1920s sought to define how visual perception is organized. His 1923 paper established several rules, including the gestalt law of proximity. In its standard formulation, perceptual proximity led to the mental association that closer elements are more similar than distant elements. I call it in this paper graphic proximity, as this and other Gestalt factors interact. It also stands in conjunction with Gestalt rules regarding grouping and similarity. Wertheimer is focussed in his paper on scaling empirically established observations to more general laws about patterns and assertions about the whole's relationship to the part. These later points were already considered more tenuous, and many derivatives of this work ensued soon after their publication. As they all bear the name Gestalt theory, without detailed archaeology of their development and distinctions to clarify their development, an overly complex and also contradictory body evolved.

Over time, neuroscientific and psychological behavioral research evolved from the original Gestalt rules. These developments are relevant to the theoretical framework used in this research. Although at first the search for general laws of form dominated Gestalt research, after its original widespread acceptance of its insights, critiques from positivists, dialecticians, and materialists regarding its mentalist and idealist tendencies led to Köhler's explanation of gestalt research as the isomorphism based on electromagnetic and thermodynamic theories [27]. The conception of an isomorphism between brain states and perception remained dominant later in the 20th century, but more recently has since become far more sophisticated, although it generally follows behavioralist traditions. Although Gestalt research was not of much significance following World War II, due to the failure to establish clear and workable rules

of visual perception, it remained significant in many professional fields. It became in various interpretations a tractable framework, for example, to introduce visual design concepts or holistic dimensions of cartography without embarking into the theosophic beliefs that came to dominate modern art, e.g., as seen in the writings of Paul Klee and Wassily Kandinsky [2].

In behavioral psychology, Gestalt theory remains a conceptual basis for studies of visual clustering and patterns, and in neuroscience, studies of preattentive patterns have influenced recent developments of more advanced theories of visual perception [25]. Ann Treisman has conducted many studies that examine how attention impacts grouping and binding [22, 21]. Other neuroscience studies regarding these matters, for example [17], provide empirical evidence that attention is not decisive in preattentive grouping. Colin Ware in his information visualization research has built nonetheless on the preattentive concepts to advance visual designs for tunable action maps that take human's innate grouping of visual elements in the environment to assist augmented reality navigation [26].

While the origins of these approaches to visualization go back to insights from the original Gestalt theorists, it is worth pointing out that Gestalt research on behavioral matters was also influential for later work of Daniel Kahneman's and Amos Tversky's that led to the development of prospect theory [12, 24]. Its relevance can be seen in the influence of Gestalt researchers work identifying visual illusions that reflect biases in vision (see Figure 1) and their research into biases in cognitive decision making. Gestalt theory also was a starting point for other influential studies. Research into cognitive mechanisms of visual perception remains an active scientific field.

2.3 Gibson and ecological psychology

Among psychologists focused on visual perception following on Gestalt research, J. J. Gibson's lifetime work in this area has had perhaps the largest impact. While impossible in the scope of this paper to consider its breadth and depth, it is relevant here to point out how Gibson's research starts with the behavioral insights of Gestalt research with a more thorough experimental-based development of theories, here labeled ecological psychology, which advanced the understanding of visual perception in a significant degree. J. J. Gibson's affordance concept has been widely used in GIScience and other fields [9, 8]. The nuanced way it is conceived of in his ecological psychology is central to understanding how people associate intuitively perceived graphical proximity with geographical nearness. In agreement with Kahneman and Tversky's behavioral psychology framework, its broadening of cognitive considerations situates vision in a system which includes ambient, accessible information in an ecological sense: "We are told that vision depends on the eye, which is connected to the brain. I shall suggest that natural vision depends on the eyes in the head on a body supported by the ground, the brain being only the central organ of a complete visual system." (Gibson, 1984, p. xii). While at the neuroscientific level of analysis cognitive processes of vision are encapsulated by brain activity, this approach broadens the scope to consider both nurture and nature factors. Vision, following Gibson, is an information-based process with vision central to activities that implicitly and explicitly involve assessments of the perceivable opportunities for action in the environment. Affordances are these opportunities. We have come to tend to think of them concerning product design and unobtrusive, even invisible, the inclusion of abilities for the pragmatic implementation [16]. Following Gibson's ecological approach, they are conceptual instantiations of specifying information that a viewer draws on to perceive and comprehend an image. In an example related to geovisualization, a hypothetical map showing the air freight volumes at the 50 largest airports of the world, would likely show the freight volumes (after standardization of the data using the graphical variable size [1]) with outlines

of national political borders including their names to facilitate speedy identification. This symbolization provides an affordance for readers of the map. The projection chosen for this map would be relevant as distortions common in the widely misused Web Mercator projection could impact the geographic associations that readers make about distances and areas. In this sense, the perception of geographic nearness through graphic proximity involves how people preattentively see and how we passively and actively come to understand the affordances in a geovisualization. Improved processes of intuitive visual perception coupled with a reduced effort of making sense [23] make for useful geovisual affordances. How graphical proximity, a fundamental Gestalt rule, affords the visual understanding of geographical nearness is a good starting point for distinguishing nurture and nature factors in visual perception. Developing an ecological understanding of visual biases and mental biases through the study of affordances used in geovisualization can aid in understanding the influence of preattentive and learned factors in the visual perception of geographic nearness in geovisualizations. Graphics-based visualization, affordances on which geographic understanding of nearness is based, should better be called geo-graphics in this sense.

2.4 Towards a theory of geo-geographical nearness

Building on these conceptual considerations regarding visual perception, the tentative theoretical framework advanced in this article is related to the research question about the relationship between perceptions of graphical proximity and geographic nearness. Its formulation in this first iteration, following Tobler's reflections, begins with an acknowledgment that theories are tools [7] that reflect the scientific, social and cultural contexts that they are created in. Acquired geographical understanding that conflicts with mapped representations is a different issue that rests on similar roots and concepts as this study focused on proximity/nearness. In any case, considering how a geographical understanding of things or processes achieved through understanding geovisualizations is different, from the understanding arrived at through direct experience, remains a relevant and valuable topic. Biases in geo-geographical understanding are complicated, lurking in all mental aspects of perception and comprehension. This work, given its narrow empirical basis, might be seen as a first and tentative attempt to identify biases and from their resolution advance GIScience in this areas. Since the data and theoretical concepts reported here are underdefined, the methodological implementation in the explorative study presented in section four of this paper relies on Bayesian statistics to consider and evaluate possible relationships between graphical patterns and other factors. The conceptual underpinnings build on information-based approach in psychology with a similar methodological process approach that Bayesian statistics align with. Proximity, in this sense, can be understood in the context of visual clustering, esp. Into patterns (Ware 2010, p. 58). De Wit et al. (2015), working directly from an information psychological refinement of Gibson's work established in their empirical work with visual illusions that selective attention influenced the use of specifying (task-specific) and non-specifying information in the perception of visual illusions. In summary, concepts of Gestalt rules, empirical behavioral, ecological and informational psychology ground concepts of affordance, visual clustering/patterns in visual comprehension, and pre-attentive patterns to develop a socio-cognitive understanding of visual perception and visual comprehension. Informational psychology concepts move this understanding and situate the framework for understanding how people perceive graphical proximity as geographical nearness to a cognitive information processing framework. Reformulating the research question based on this theoretical framework, the conceptual distinctions between perception and comprehension are relevant for examining how training or experience leads to recognizable specifying information that viewers draw on

in the perception of graphical proximity as geographical nearness. A simple assessment of the validity of this connection in the pilot study seeks to establish whether a relationship between training or experience to the perception of geographic nearness is evident.

3 Experimental Design and Methods

The explorative research for this paper relies on Bayesian methods. Bayesian methods have found in GIScience interest and uptake for research in semantics, land use modeling, and spatial statistics. This paper utilizes Bayesian methods with their strengths for explorative research to operationalize the tentative theoretical framework presented in the previous section, an application more often seen in psychology and other social sciences.

3.1 Bayesian Methodology Background

For this explorative research Bayesian methods are well suited in comparison to statistical techniques relying on frequentist statistics. Bayesian methods are ideally suited to identify and evaluate relationships in data as it produces results that reflect additional data. At the risk of oversimplification, Bayes Theorem focuses on elaborating an understanding through the evolving statistical testing of data that refines understanding of the phenomena. In statistical terms, Bayesian methods produce posterior distributions that researchers have understood speak to issues with polling, confidence intervals and p-values in classical frequentist statistics [6]. The prior belief about the data $P(A)$ is calculated with the normalization constant $P(B)$ and the conditional probability $P(B|A)$, as in Bayes' Rule:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \quad (1)$$

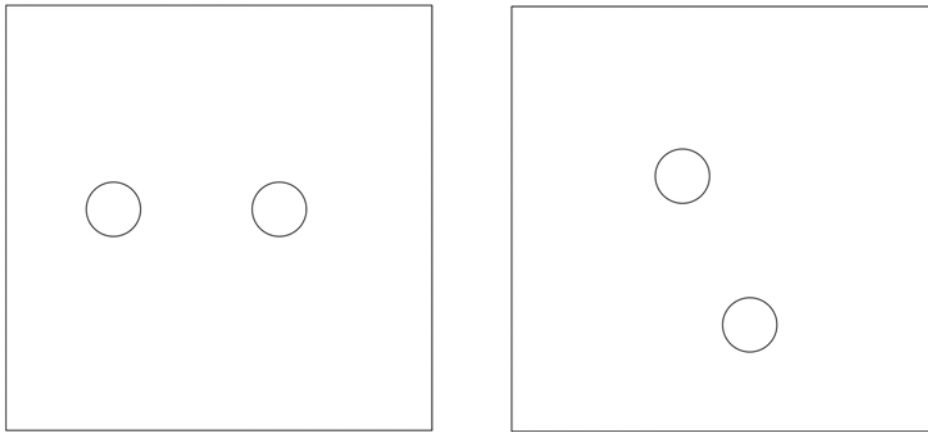
In other words, Bayes' Rule is the posterior probability equal the likelihood times the prior divided by the normalization constant. The application is wide-reaching in helping develop statistically grounded insights. For instance, this allows new observations, implemented as measures, to be introduced and modify the prior known distribution assuming a meaningful relationship between prior data and new data.

3.2 Explorative Study and Bayesian Methods Used

The use of Bayesian methods will be helpful in assessing the instruments and for further refinement of this data. The data collected for the study comes from an online survey developed with the software suite Lime Survey (limesurvey.org). It was analyzed using classical and Bayesian methods available in the open software package JASP from the University of Amsterdam (JASP-stats.org). Data were collected in October 2017 from students in the Master's level Critical Cartography seminar who had received an invitation with the URL. 10 students, 5 male, 3 female and 2 unknown, completed the survey. The ages of the participants ranged from 24 to 51, the mean age of the group was 30.4. The limits of the sample size are significant. Nonetheless, given the focus of this article, the small n remains helpful for a pilot study. The survey consists of five background questions and eight A/B comparisons using pairs of images. Each A/B test consisted of two identical images except for the location of two circles or diamonds in each. These symbols were placed in different distance to each other. Participants were requested to determine in which image the two circles were closer together without using additional aids. They were instructed at the beginning of the survey that the time for them to indicate a response was relevant to the study.

4 Explorative Study Analysis and Critique

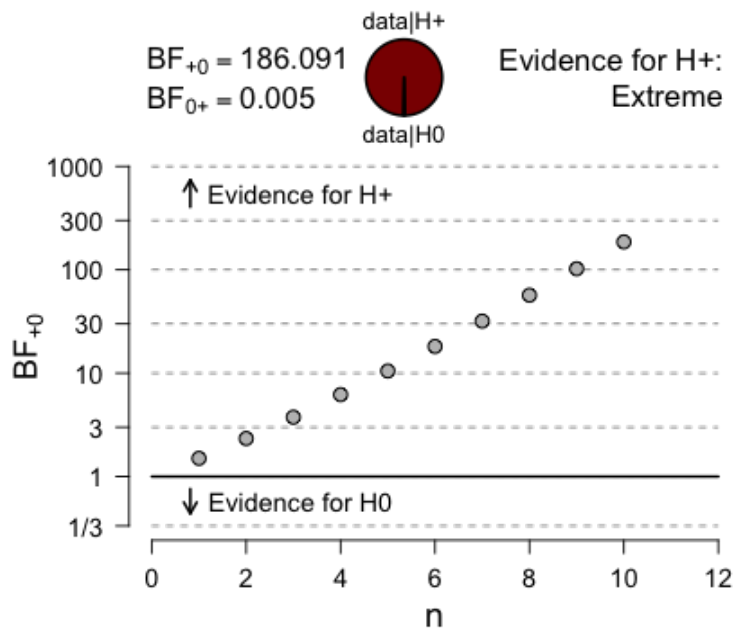
The images in the online survey instrument for an invited group of participants presented pairs of images that varied in graphical complexity and geographical context. Some included random graphic elements, others used outlines of countries, and some included both random graphic elements and country outlines (see Figure 8). Following the central research question, the survey focused on assessing differences in the times it took participants to identify the closest graphic proximity between two elements from each pair of images while increasing the number of other elements and adding a geographical dimension to the graphic. The key instrument operationalized in the survey is if graphical proximity affords the visual understanding of geographical nearness. The results can be analyzed with data on experience and cartographic/GIS education. Additional questions about training and experience can shed insights into these factors influence on the experimentally established behavior and help understand ecological visual biases and mental biases.



■ **Figure 4** Image pair used for the first comparison by study participants.

The presentation of the study commenced with an overview and presentation of some summary statistics. Most of the ten survey participants had cartography/GIS courses or experience. Two indicated have more than 3 years experience, and six had 2–3 years experience. The other two respondents had no or 1 year of experience. As to be expected, most of the students (5) had 2–3 courses in cartography/GIS; two had one course, and three had had 3 or more courses. Correspondingly, the majority of participants recognized a set of widely-used terms in cartography/GIS. Interesting is the ambiguity in the relationship between experience and courses to the knowledge of these terms despite general indicated knowledge of the terms. Of 14 terms, only the terms reliability and visual variables were known by two of the respondents. Typography and raster were the only terms known by 7 of the participants. The contingency table analysis of these responses regarding years of experience also showed no strong positive relationship. Curiously, the contingency table analysis of the responses regarding the number of cartography/GIS courses taken pointed to an increased lack of knowledge of these four terms among participants who had taken 1–2 courses in contrast to the students who had taken 1 or 3 or more courses.

The times taken by participants to identify the image out of two that showed the more considerable distance between two circular or two diamond shaped elements is the key attribute of the studies to consider. While the variance in the response time is noticeable, generally, even with increasing complexity of the images to compare, the response time



■ **Figure 5** Graphic summarizing the Bayesian sequential analysis for figure pair one from JASP. The result offers extreme support for the alternative hypothesis, increasing in reliability as each response is calculated.

declines. The minimum response time for all eight image pairs drops from 9.4 to 5.6 seconds and the maximum decreases from 40.68 to 16.81. The means and standard deviation values also decline. These results suggest a learning effect having a substantial impact on the results. This issue is significant and considered later in greater detail.

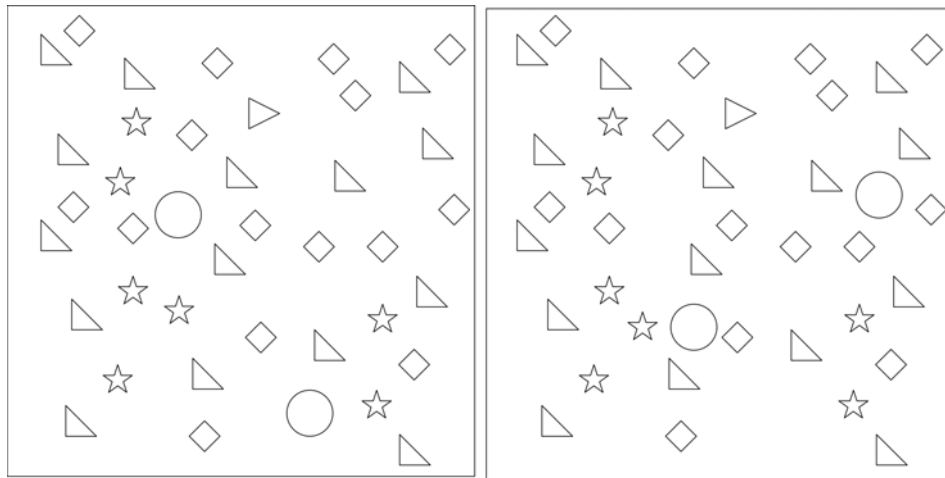
4.1 Analysis of the Pilot Study Data

A Bayesian binomial analysis was applied to the responses to establish whether participants relied on guessing to complete the survey. The majority of responses were correct, but some figure pair comparisons of up to two participants failed to identify the figure with the closest pair of objects. First, though, the first pair of figures, which all participants identified correctly (see Figure 4) offers a benchmark of Bayesian binomial statistic for assessing this statistic for the figure pairs that were not consistently correctly interpreted. The very high Bayes Factor BF_{+0} of 186.091 is substantial support for the alternative hypothesis that participants were not guessing in their interpretations.

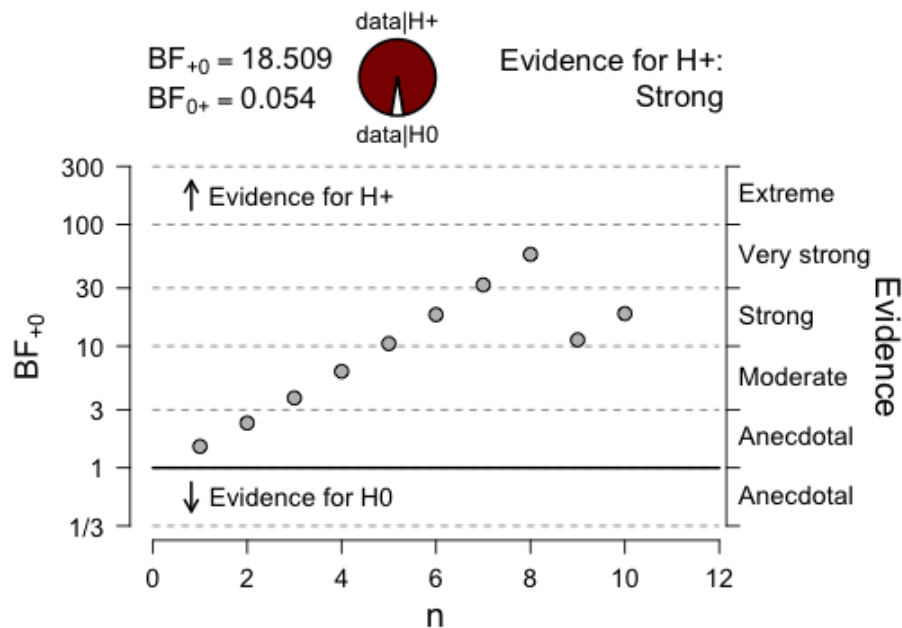
In image pair three, one person incorrectly misinterpreted the closer pair of elements or incorrectly chose the image, which placed the two elements further apart. The Bayesian sequential analysis of the Bayesian binomial statistic shows this in the chart and in the Bayes Factor BF_{+0} of 18.509. This lower Bayes Factor suggests this figure pair could be more challenging to interpret.

The final Bayesian binomial analysis considers the responses to image pair 5. Analogous to the analysis of figure pair 3, the sequential analysis chart shows how incorrect interpretations of proximity have a negative impact and strength of the Bayes Factor.

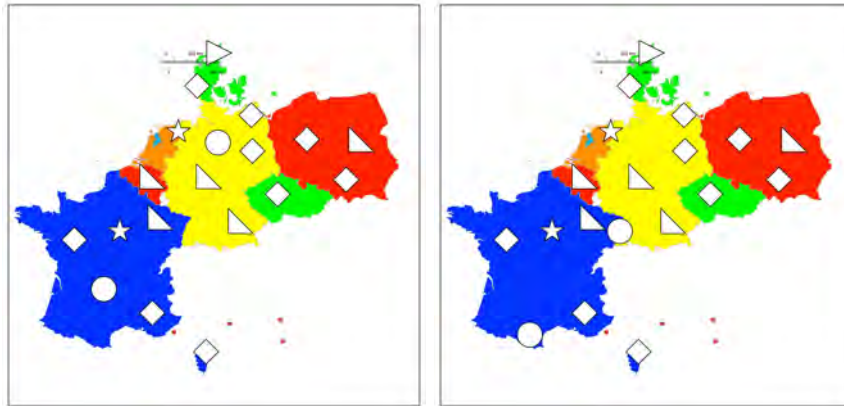
4:10 Graphical Proximity and Geographical Nearness



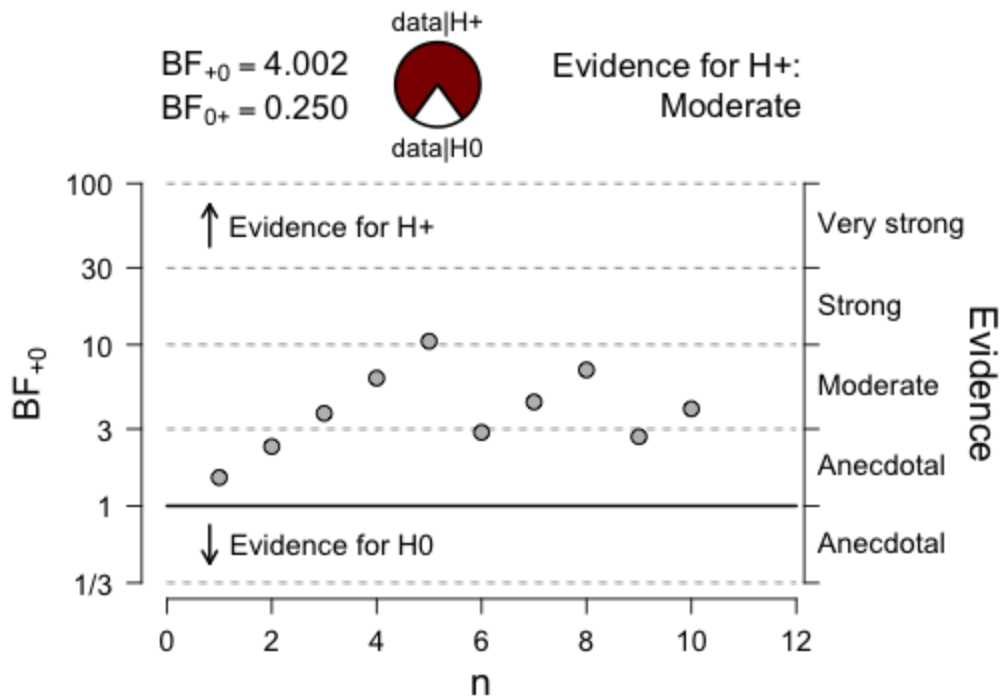
■ **Figure 6** Image pair used for the third comparison, which include complex arbitrarily placed graphic elements.



■ **Figure 7** Bayesian sequential analysis chart for figure pair three. The result offers strong support for the alternative hypothesis, increasing in reliability until the image with more distant elements was selected by participant.



■ **Figure 8** Image pair used for the fifth comparison, which include complex arbitrarily placed graphic elements.



■ **Figure 9** Bayesian sequential analysis for figure pair five. The result offers moderate support for the alternative hypothesis, the Bayes Factor increasing and decreasing in reliability as each response is calculated.

4:12 Graphical Proximity and Geographical Nearness

■ **Table 1** Response time (in seconds) statistics for all participants from JASP.

	Q 1	Q 2	Q 3	Q 4	Q 5	Q 6	Q 7	Q 8
Mean	22.90	44.30	30.52	26.80	23.38	13.70	11.38	10.78
Standard Deviation	11.28	56.08	26.20	21.37	9.15	6.08	5.97	3.76
Minimum	9.49	9.19	10.68	8.65	14.44	6.50	6.43	5.60
Maximum	40.68	183.3	99.14	73.62	40.61	24.13	23.88	16.81

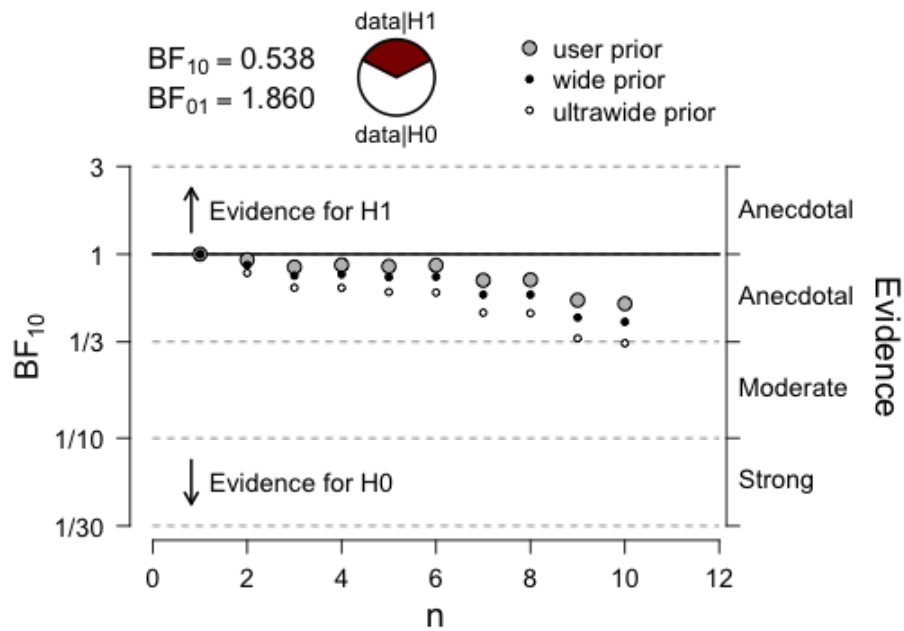
4.2 Impacts of training and experience

The Bayesian binomial analysis helps assess the reliability of the interpretations and the presentation in the sequential analysis charts documents the strength of Bayesian analysis in exploring data. With this small data set the effects are not especially pronounced, but in future work with much larger data sets Bayesian analysis will be of great assistance. The significantly reduced response times among all participants suggest that immediate learning of the affordances available in the study instruments is of considerable impact.

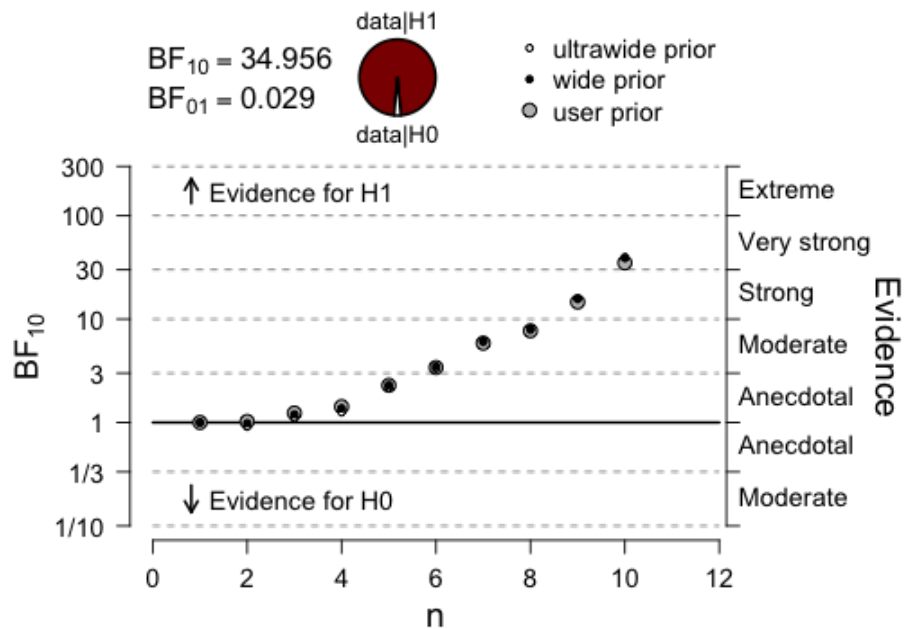
4.3 Bayesian correlation

As a final set of Bayesian analysis, the Bayesian binomial analysis presented above suggests that some image pairs were more complicated to interpret. Did they require more time to analyze? Did training or experience impact the response times? Did immediate learning of the affordances influence the response times despite increasing graphic and geographic complexity in the image pairs? The results of the statistical analysis are inconclusive. The standard deviations in table 1 above suggest that lack of clear associations between participant response times and image pair complexity. The Bayes Factors from the Bayes Paired T-Tests comparing response times from figure pairs with an imperfect identification of the image with the closest pair of elements varied in direction and strength.

As pointed out at several points in this paper, but worth repeating and emphasizing here, the sample size of 10 is a great limitation to considering the relevance of the study's results. The insights from the explorative analysis, however, clearly point to the need for improvements in the methods and instruments beyond increases to the sample size, which will be discussed in the next and final section of the paper.

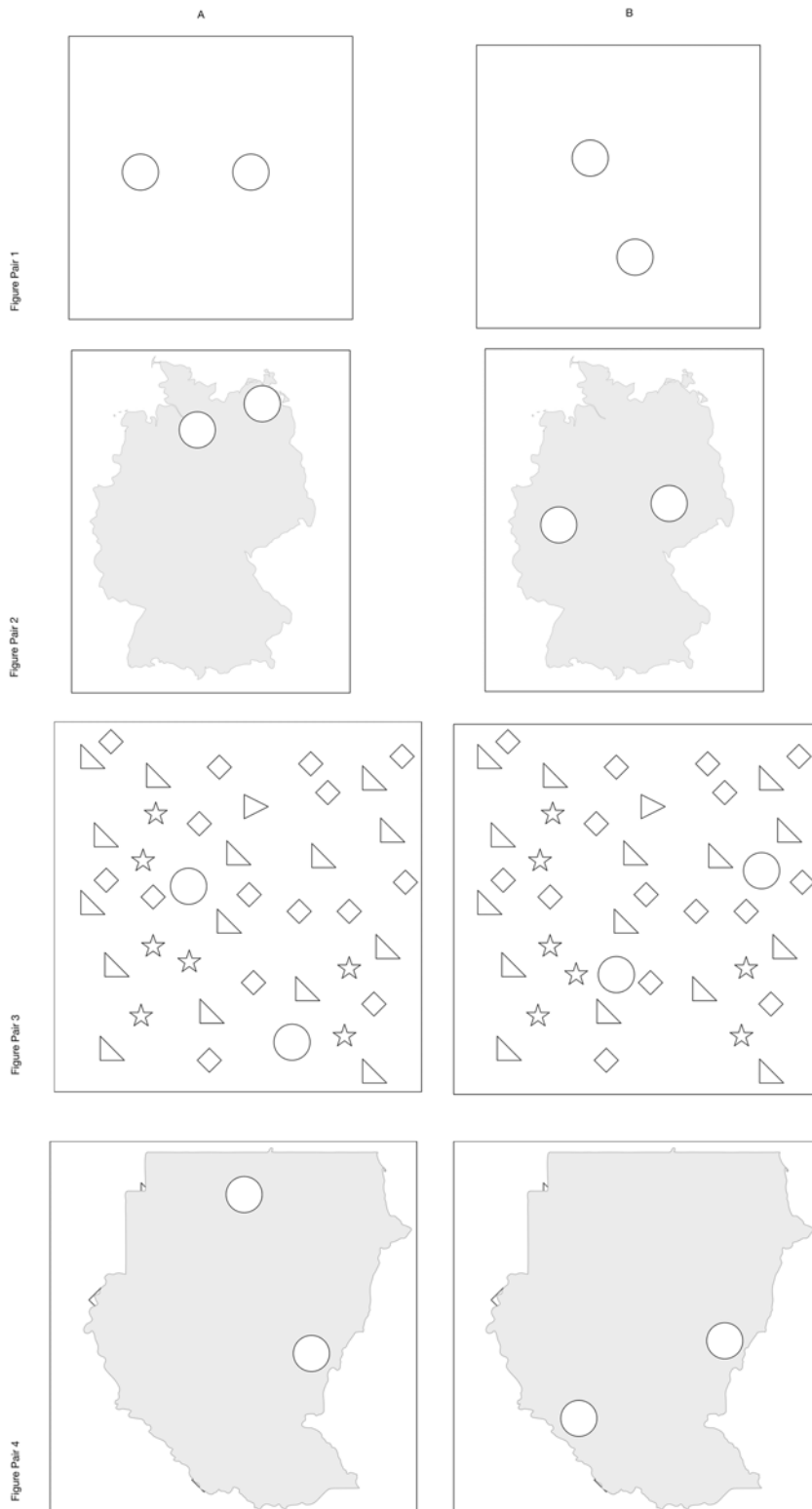


■ **Figure 10** Sequential analysis of Bayes Factors chart comparing response times from image pair three and image pair five.

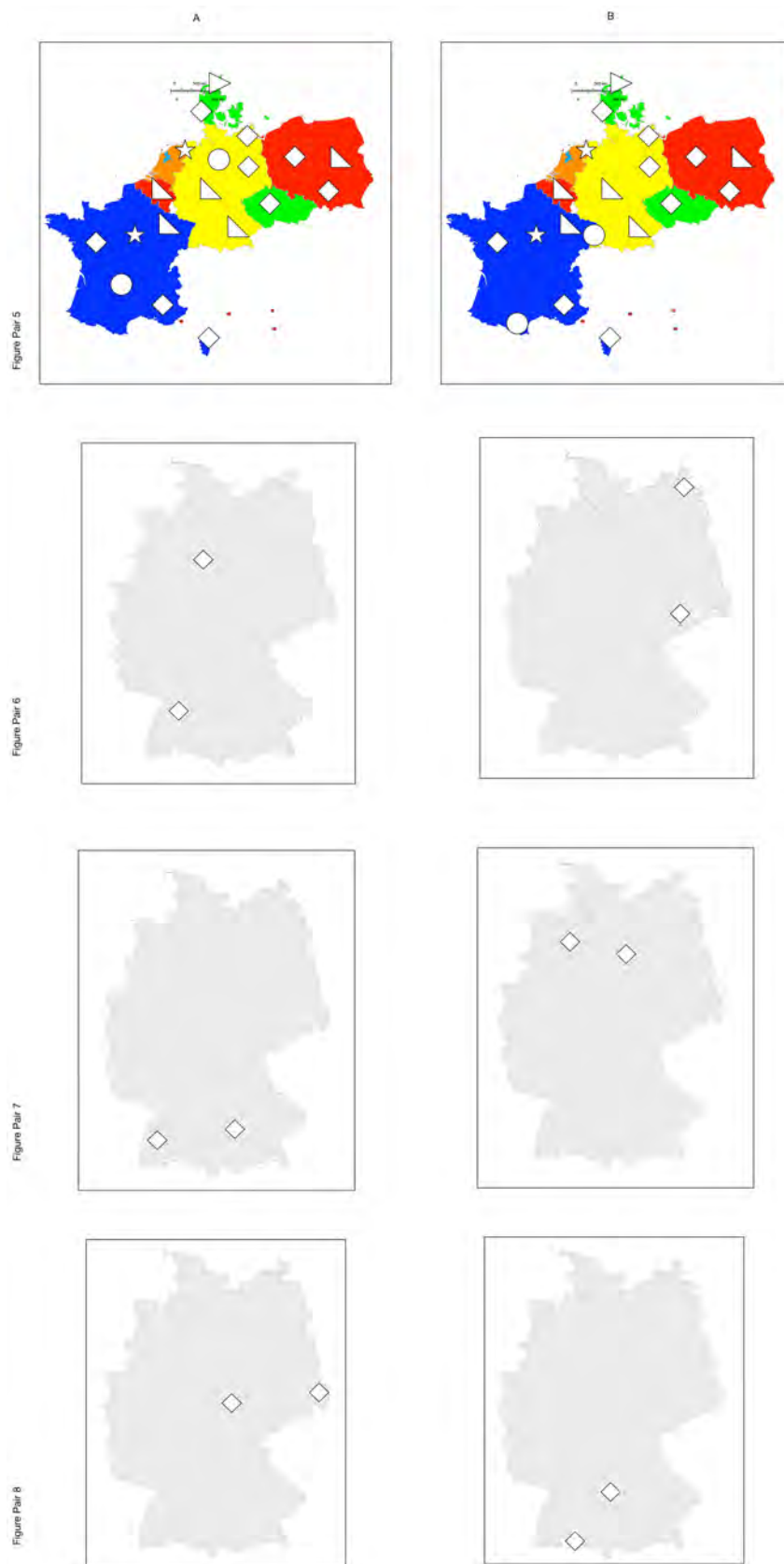


■ **Figure 11** Sequential analysis of Bayes Factors chart comparing response times from image pair five and image pair seven.

4:14 Graphical Proximity and Geographical Nearness



■ **Figure 12** Figure pairs 1 - 4 used in the online survey (not to scale).



■ **Figure 13** Figure pairs 5 - 8 used in the online survey (not to scale).

5 Conclusions and future research

The results of this explorative study suggest that an understanding of geographic nearness from graphical proximity, and thus Tobler's first law of geography, involves more than intuition in work with geovisualizations. The theoretical framework used in this research seems robust for this research which involves cognitive and social dimensions. The methods and instrument need further development for continued research to better account for preattentive grouping and acquisition of affordance while taking the study. The reduced response times identified from the pilot study data can be explained by the acquisition and application of affordances arising from learning how to make sense of the survey image pairs efficiently. Whether this interpretation is true remains to be verified through future research.

It remains essential to gain a better understanding of the how people come to understand graphical proximity and geographical nearness in geovisualizations. The starting point is already there. The over 60 year old, but then prescient and still relevant, insight from Herbert Bayer to refer to the special and often exceptional capabilities of graphics in geographic communication by separating with a hyphen geo from graphic (geo-graphic) emphasizes how graphic elements and their composition involve different visual perceptions than reading text or even environmental perception. The theoretical framework resting on Gestalt concepts and ecological psychology shown as the foundation for exploratory research presented in this paper suggests geographical nearness is more than an intuitive deduction from graphical proximity. Instead, the two modes of perception and comprehension intertwine in still to be understood ways. The pilot study results suggest in agreement with Gibson that the human mind learns very rapidly how to apply preattentive affordances to visual tasks. The evidence collected for this study does not make clear to what degree training in spatial thinking may be a significant factor in this learning and to what degree experience is a factor. Future research should explore the degree to which developing and applying orthogonal and absolute coordinate reference systems may influence the mind's transformations of graphical proximity to geographical nearness.

The development of these affordances may profit from considering research by Ann Treisman and colleagues on patterns and pre-attentive perception or tunable mental images [26, 21]. It seems possible that both training and experience lead to enhanced mental faculties that tune the post-attentive process of visual perception using acquired patterns.

While open for continued study and refining the theoretical framework, these exploratory results point that geo-graphic nearness understood from graphics is separate from the geographical nearness concepts intrinsic to way-finding and environmental perception. A conjecture about this difference to consider in future research seems straightforward: A geo-graphic visualization provides fundamentally different affordances to the point of producing biases and even distorting our environmental knowledge. In contrast, geographic concepts seem probably based on stronger cognitive concepts. Learning seems essential to the capacity, and hence social and cultural factors become relevant. A known place geographically close to us, in, for example, the sense of measurable distance, maybe still be nearer yet in our understanding of graphic proximity, e.g., topological maps of urban transportation including Harry Beck's famous map of the London Underground network. This difference is relevant in many daily and emergency situations. The results from continued research into these differences can lead to a better understanding of the socio-cognitive mechanisms how people understand proximity/nearness. Future research should consider the relevance of these mechanisms to help address accessibility issues for specific requirements and needs, e.g., people with disabilities or seniors. Another aspect for future research is consideration of

temporal aspects in the use of graphic geovisualizations including a more realistic study of the use of maps, e.g., the consideration of travel modes to gain more insight into map reading activities pursued in relation to the reader's goals and the functional support of a map. Ware's tunable action maps seem here to be a useful reference to consider also how naive geographical concepts contradict common mapped presentations, e.g., the experience that Reno is further west than Los Angeles.

The tentative findings are not conclusive and require improvements to the experimental design that accounts for the tentative findings and methodological issues the preliminary work raised. Improvements to the survey instruments used in this exploratory research need to address several points. First, is a research design that allows the learning of the survey instrument's affordance to be measured. Second, the difference between types of image pairs needs to be analyzed and accounted for in the statistical analysis. Third, the number of participants needs to be significantly increased.

To summarize the study and its relevance to GIScience, people in the survey reported here learned how graphical proximity corresponds to geo-graphical nearness, establishing specific cognitive mechanisms in the context of improvements to the methodology regardless of the defined task. While the graphic format and media are relevant, an acculturated, acquired sense of distance in coordinate system representations, arising or enforced by graphical presentations without geographic knowledge could lead to a misleading or even a false sense of actual topographic distance between objects. Training and experience remain factors with the fast learning of affordances to control for in future research. In developing this research, the distinction graphical and geographical, which becomes in some cases problematic due to information-based functional approaches to geospatial comprehension [14] offers a good foundation for continuing this research and refining it empirically to understand contributing factors and the specifics of visual biases that impact Tobler's first law of geography.

References

- 1 Jacques Bertin. *Semiology of Graphics: Diagrams, networks, maps*. University of Wisconsin Press, Madison, WI, 1983.
- 2 Geert-Jan Boudewijnse. Gestalt theory and bauhaus-a correspondence. *Gestalt Theory*, 34(1):81–98, 2012.
- 3 Jerry Brotton. *A history of the world in twelve maps*. Penguin UK, 2012.
- 4 Helen Couclelis and Jon Gottsegen. What maps mean to people: Denotation, connotation, and geographic visualization in land-use debates. In S. Hirtle and A. U. Frank, editors, *Spatial Information Theory. A Theoretical Basis for GIS*, pages 151–162. Springer Verlag, Berlin, 1997.
- 5 Matthieu M de Wit, John van der Kamp, and Rob Withagen. Visual illusions and direct perception: Elaborating on gibson's insights. *New Ideas in Psychology*, 36:1–9, 2015.
- 6 Allen Downey. *Think Bayes*. O'Reilly, Sebastopol, CA, 2013.
- 7 Paul Feyerabend. *Against Method*. Verso, London, 1997.
- 8 James J Gibson. *The perception of the visual world*. Houghton Mifflin, 1950.
- 9 James J Gibson. *The ecological approach to visual perception: classic edition*. Psychology Press, 2014.
- 10 Francis Harvey. *A Primer of GIS. Fundamental Geographic and Cartographic Concepts*. Guilford, New York, 2016.
- 11 ICA, editor. *Cartographic futures on a digital earth*, Ottawa, Canada, 1999. ICA,.
- 12 Daniel Kahneman. *Thinking, Fast and Slow*. Penguin, New York City, 2011.

- 13 Menno-Jan Kraak. The web, maps, and society. In Bruce Gittings, editor, *Integrated Information Infrastructures with GI Technology. Innovations in GIS*, pages 67–78. Taylor and Francis, London, 1999.
- 14 Werner Kuhn. Core concepts of spatial information for transdisciplinary research. *International Journal of Geographical Information Science*, 26(12):2267–2276, 2012.
- 15 Dan Montello and Scott Freundschuh. Cognition of geographic information. In Robert McMaster and Lynn Usery, editors, *A research agenda for geographic information science*, pages 61–91. CRC Press, Boca Raton, FL, 2005.
- 16 Donald A. Norman. *Things That Make Us Smart: Defending Human Attributes in the Age of the Machine*. Addison-Wesley Publishing Company, 1993.
- 17 Sugihara T. T., Qiu F and von der Heydt R. Figure-ground mechanisms provide structure for selective attention. *Nature Neuroscience*, 10:1492–1499, 2007.
- 18 W. Tobler. Preface. In Francis Harvey, editor, *Are there fundamental principles in Geographic Information Science?* CreateSpace/Amazon Kindle, Seattle, 2012.
- 19 W. R. Tobler. A transformational view of cartography. *The American Cartographer*, 6(2):101–106, 1979.
- 20 Waldo Tobler. On the first law of geography: A reply. *Annals of the Association of American Geographers*, 94(2):304–310, 2004.
- 21 Anne Treisman. Preattentive processing in vision. In *Human and machine vision II*, pages 313–334. Elsevier, 1986.
- 22 Anne Treisman and Stephen Gormican. Feature analysis in early vision: evidence from search asymmetries. *Psychological review*, 95(1):15, 1988.
- 23 Barbara Tversky. Distortions in memory for maps. *Cognitive psychology*, 13(3):407–433, 1981.
- 24 Amos Twersky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131, 1974.
- 25 Johan Wagemans, James H Elder, Michael Kubovy, Stephen E Palmer, Mary A Peterson, Manish Singh, and Rüdiger von der Heydt. A century of gestalt psychology in visual perception i. perceptual grouping and figure-ground organization. *Psychol Bull Psychological bulletin*, 138(6):1172–1217, 2012.
- 26 Colin Ware. *Visual thinking: For design*. Morgan Kaufmann, New York, 2010.
- 27 Gerald Westheimer. Gestalt theory reconfigured: Max wertheimer’s anticipation of recent developments in visual neuroscience. *Perception*, 28(1):5–15, 1999.

An Empirical Study on the Names of Points of Interest and Their Changes with Geographic Distance

Yingjie Hu

GSDA Lab, Department of Geography, University of Tennessee, Knoxville, USA
yhu21@utk.edu

Krzysztof Janowicz

STKO Lab, Department of Geography, University of California, Santa Barbara, USA
jano@ucsb.edu

Abstract

While Points Of Interest (POIs), such as restaurants, hotels, and barber shops, are part of urban areas irrespective of their specific locations, the names of these POIs often reveal valuable information related to local culture, landmarks, influential families, figures, events, and so on. Place names have long been studied by geographers, e.g., to understand their origins and relations to family names. However, there is a lack of large-scale empirical studies that examine the *localness* of place names and their changes with geographic distance. In addition to enhancing our understanding of the coherence of geographic regions, such empirical studies are also significant for geographic information retrieval where they can inform computational models and improve the accuracy of place name disambiguation. In this work, we conduct an empirical study based on 112,071 POIs in seven US metropolitan areas extracted from an open Yelp dataset. We propose to adopt term frequency and inverse document frequency in geographic contexts to identify local terms used in POI names and to analyze their usages across different POI types. Our results show an uneven usage of local terms across POI types, which is highly consistent among different geographic regions. We also examine the decaying effect of POI name similarity with the increase of distance among POIs. While our analysis focuses on urban POI names, the presented methods can be generalized to other place types as well, such as mountain peaks and streets.

2012 ACM Subject Classification Information systems → Language models

Keywords and phrases Place names, points of interest, geographic information retrieval, semantic similarity, geospatial semantics

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.5

1 Introduction

People name the environment that surrounds them to communicate about it. Almost every aspect of geographic space that can be described and depicted can be named. It has been suggested that place names, or toponyms, play a key role in stabilizing the otherwise unwieldy space into more manageable textual inscriptions [38, 25, 42]. From a perspective of *space* and *place* [45], the creation of a place name signifies the important moment when people explicitly integrate human experience with space.

Place names, made available via digital gazetteers, power GIS, geographic information retrieval (GIR), and modern search engines and recommender systems more broadly [20, 13, 47]. After all, people communicate using place names not coordinates. Interestingly, and in difference to human geography, most GIR research simply uses place names as identifiers instead of examining how those names were formed and how similar they are to nearby



© Yingjie Hu and Krzysztof Janowicz;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 5; pp. 5:1–5:15

Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

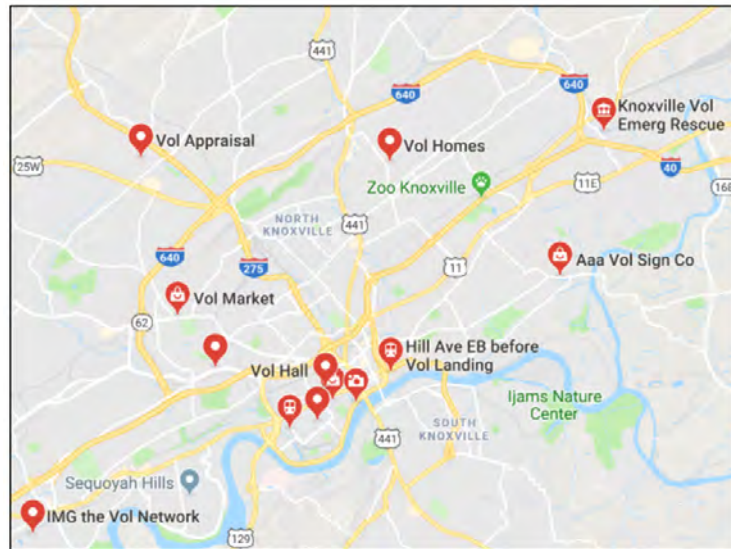
names. This is understandable since we are often interested in questions such as *What are the best Italian restaurants within 10 minutes driving?* instead of the specific names of these restaurants or what they reveal about the history of a region, such as immigration trends.

Place names have long been studied in human geography with a traditional focus on etymology and place taxonomies [52, 40]. For example, the place name *Las Vegas* means *The Meadows* in Spanish and points to the former abundance of wild grasses and desert springs, both of which were crucial information for travelers and led to the descriptive place name. While such studies provide in-depth explanation of place names, they are often limited to case-by-case examinations with qualitative descriptions. This could include studies focusing on specific regions, names, places types, and so forth.

In contrast, this work is based on more than 110,000 place names of different types distributed across seven metropolitan areas within the US. Our focus is on uncovering term usage patterns and their relations with geographic locations, e.g., as modeled by a decaying influence or local names with increasing distance. There are several reasons for performing such a large-scale, data-driven study. First, place names reveal many social and cultural characteristics, and can help us understand various aspects of urban areas. Previous research in human geography has considered place names, such as street names, as *city-text* embedded in the cityscape [6, 7]. A systematic examination on these city-texts, can help expand our knowledge of the studied regions. Second, large-scale empirical research examining place names can aid in discovering common principles in place naming and relations to environments. This can be distinguished from case-by-case place name studies in which the discovered knowledge often cannot be generalized to other names or geographic areas. Third, such studies can facilitate the development of computational models for places. We can integrate the discovered common principles, socio-cultural characteristics, and local terms into computational models, e.g., via an implemented knowledge base, to better support tasks such as place name disambiguation [4, 27, 37, 17]. This last point is a key strength of this work. Our results can act as a quantitative foundation for the localness of identifiers *per place*, which will enable future research to push the envelop on place name disambiguation. In fact, our previous *Things and Strings* place disambiguation method [22] has demonstrated the usefulness and need for combining geographic and linguistic information.

The names of Points Of Interest (POIs), such as restaurants, hotels, grocery stores, and auto repairs, are examined in this work. These POI names are from an open dataset released by Yelp, a company that provides search services for local businesses. POIs play important roles in supporting many aspects of our daily life [33, 36, 51]. One reason we select POI names for this study is that these names reflect more of the diverse views of the general public, since the business owners can decide on names themselves. This can be differentiated from other place names, such as street names, which often result from political and administrative decisions [7, 1, 41]. In addition, the names of POIs often contain local information, such as city or state names, natural or man-made geographic features, vernacular names, local families (e.g., a family-owned business), language patterns, local cultural differences, and others. Figure 1 shows an example of searching for the word “Vol” in the city of Knoxville, Tennessee, USA using Google Maps. It returns many places which use this term as part of their names, as “Vol” is the local nickname of the popular football team “Volunteer”. The use of American sports team names in toponyms was also noted in previous human geography research [8]. In GIR and place name disambiguation, understanding the link between “Vol” and the city of Knoxville can help locate related place names more accurately.

More specifically, we aim to answer the following questions in this work: 1) what are the local terms that are used in POIs in different geographic areas? 2) how are these local terms used in different types of POIs, such as restaurants, hotels, and barber shops? and 3) how



■ **Figure 1** An example of POIs in Knoxville, TN, USA that use “Vol” as part of their names.

do POI names change with geographic distance? **The contributions of this paper are as follows:**

- We propose adopting the technique of term frequency and inverse document frequency in geographic contexts to identify local terms used in POIs in different metropolitan areas.
- We find an uneven usage of local terms in the names of POIs across POI types, and such an uneven usage is highly consistent across the seven studied metropolitan areas.
- We test two types of models, count-based vector and word2vec, for understanding and capturing the distance decay effect of the similarity of POI names.

The remainder of this paper is structured as follows. Section 2 reviews related work on place names and toponym disambiguation. Section 3 describes the dataset used in this study and an exploratory data analysis. Section 4 presents methods and experiments for identifying local terms from POI names, examining their usages across POI types, and modeling the distance decay effect of POI name similarity. Section 5 summarizes this work and discusses future directions.

2 Related Work

Place names have attracted the interest of many researchers in geography. For decades, geographers have been collecting and categorizing place names, studying their origins, and understanding their meanings [50, 52, 35]. It has been argued that the act of assigning a name to *space* plays a key role in producing the social construct of *place* [40]. As suggested by Carter [10], place names transform space into knowledge that can be read. The social, cultural, and political implications of place names have been widely studied [5, 6]. Examples include the renaming of streets after the establishment of a new regime to memorize new stories [30, 41], the use of street names to challenge racism [2, 3], and assigning more marketable names to local businesses and hospitals [39, 24].

Digital gazetteers provide systematic organizations of place names (N), place types (T), and spatial footprints (F) [16, 13]. As valuable knowledge bases, gazetteers provide important functions for various applications by connecting the three core components. The key functions of a gazetteer include lookup ($N \rightarrow F$), type-lookup ($N \rightarrow T$), and reverse-lookup ($F \times T$)

→ N) [19]. The first case, for example, corresponds to a query for the spatial footprint of the place name *CMS Auto Care*, the second to the place type, and the third to the place names given the spatial footprint and a place type (e.g., *Automotive*). Research was conducted to enrich gazetteers with (vague) place names and their fuzzy spatial footprints. Jones et al. [21], for instance, used a search engine to harvest geographic entities (e.g., hotels) related to vague place names (e.g., “Mid-Wales”), and utilized the locations of these harvested entities to construct vague boundaries. Flickr photos present a natural link between textual tags and locations, and have been used in many studies on identifying the boundaries of vague places and regions [15, 26, 18, 28]. Twaroch and Jones [46] developed a Web-based platform, called “People’s Place Names”, which invites local people to contribute vernacular place names.

In geographic information retrieval [20], place names are frequently discussed in the context of place name disambiguation. Since different place names can refer to the same place instance and the same place name can refer to different place instances, it is challenging to determine which place instance was referred to by a name in text, e.g., the abstract of a news article [4, 27]. Gazetteers have been used in many ways for supporting place name disambiguation. Based on the related places in a gazetteer (e.g., higher-level administrative units), researchers developed methods, such as co-occurrence models [37] and conceptual density [9], to disambiguate place names. Based on the spatial footprints of place instances, researchers designed heuristics for place name disambiguation, e.g., place names mentioned in the same document generally share the same geographic context [29, 43]. The process of recognizing and resolving place names from texts is called *geoparsing* [12, 23, 14, 49]. Place names are also examined in studies on toponym matching and geo-data conflation [44].

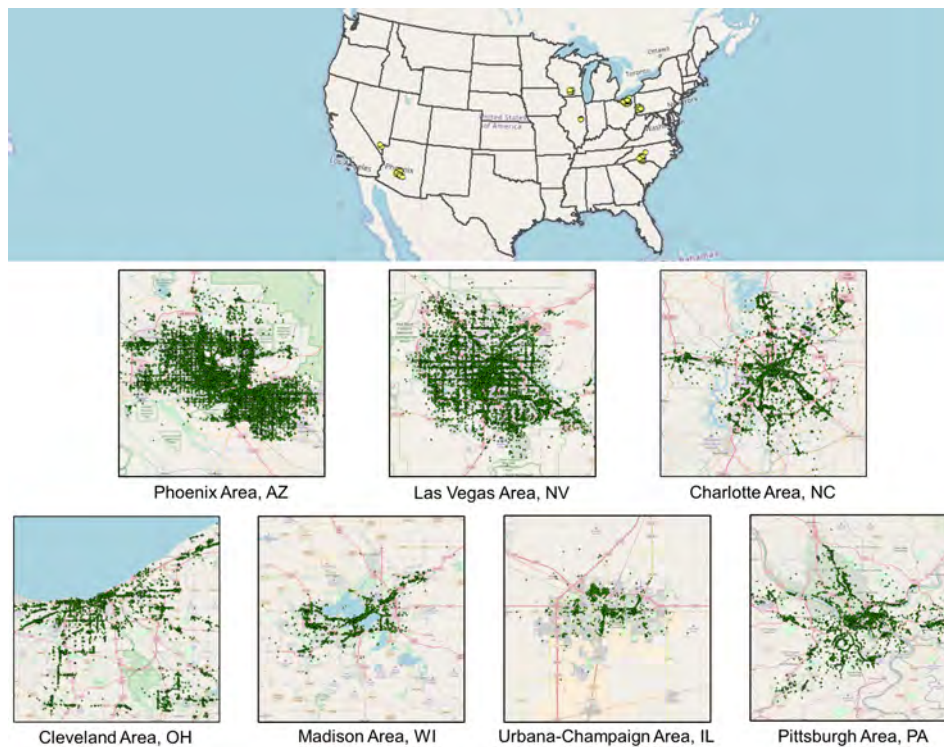
Few existing studies, however, have empirically examined the term usage of place names and their relations with geographic locations based on large datasets. Longley, Cheshire, and colleagues [31, 11] investigated the geospatial distributions of surnames based on the data from the Electoral Register for Great Britain and delineated surname regions. Their study is related to our work, since family names are included in the names of some local business. We perform an empirical study based on a large number of POI names in different US metropolitan areas. Compared with the existing literature, this work is unique in that it examines the local terms in POI names, explores the term usage patterns, and analyzes the relations of POI names to geographic locations as well as their decay in this relationship over distance.

3 Dataset

We first describe the data used in this empirical study, which is an open POI dataset from Yelp (<https://www.yelp.com/dataset>). The original dataset contains POIs from 11 metropolitan areas in four countries: the US, Canada, the UK, and Germany. Considering the language differences in POI names (e.g., German and English) and the barrier effects of country borders, we focus on the seven metropolitan areas within the US, which contain 112,071 POIs. Each POI data record has the POI name, city name, state name, latitude-longitude coordinates, and other information, such as the number of reviews and average rating. Figure 2 shows the general locations of the seven metropolitan areas and the geographic distributions of the POIs in each of these areas.

We start by performing an exploratory analysis on the term usage frequency in the POI names. It has been found that Zipf’s law exists in the usage of terms in natural language texts [32], namely the frequency of a term is proportional to the inverse of its frequency rank among all terms (Equation 1).

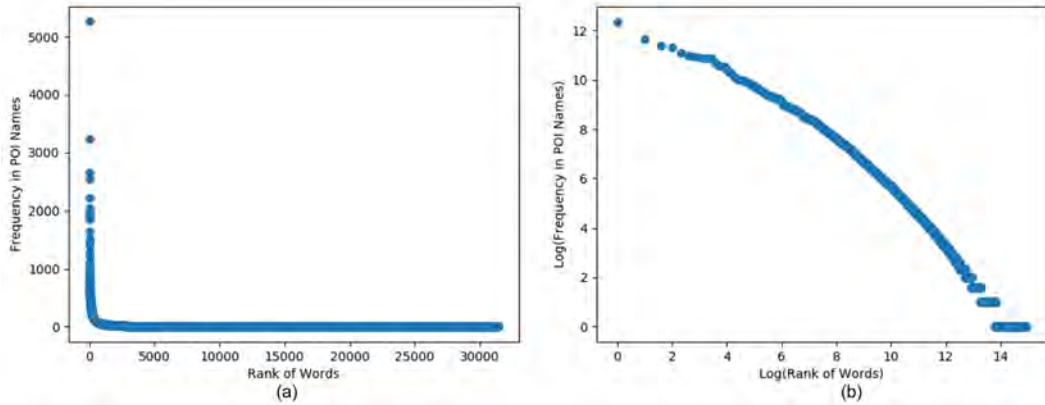
$$f \propto \frac{1}{r} \tag{1}$$



■ **Figure 2** The seven US metropolitan areas and their POIs used for this study.

where f is the frequency of a term and r is the rank of the term among all terms based on frequency. According to Zipf's law, a small number of terms are used highly frequently while most others are used only occasionally. The names of POIs are different from natural language texts in that they are typically not complete sentences but phrases. In this situation, does Zipf's law still hold in POI names?

To answer this question, we develop a Python script which reads through the names of the POIs in the seven metropolitan areas, counts the frequencies of all terms contained in each name, and ranks the terms based on their frequencies. We then use the ranks as the horizontal coordinates and term frequencies as the vertical coordinates, and the result is shown in Figure 3(a). As can be seen, there is a highly skewed distribution of term frequency with a long tail, which suggests that a small number of terms are used much more frequently than most other terms. In fact, Figure 3(a) shows almost a right angle fall-off since the term frequency decreases rapidly with a small increase of the rank. The log-log plot of the frequencies and ranks is shown in Figure 3(b), and we see almost a straight line. To quantitatively measure the match of term usage in POI names to Zipf's law, we fit a linear regression model with $\log f = A + b \log r$, and obtained an R-squared value of 0.962. Based on this exploratory analysis, we conclude that the term usage in POI names also follow Zipf's law, even though POI names are usually not complete sentences. The top 10 most frequent terms in POI names in this Yelp dataset are: *the*, *and*, *of*, *center*, *pizza*, *grill*, *spa*, *bar*, *auto*, *restaurant*. These most frequent terms reflect the inherent characteristics of POI names and POI types. It is worth noting that the most frequent terms in POI names may change across countries, depending on the corresponding cultures and lifestyles.



■ **Figure 3** Term frequencies and their ranks in POI names: (a) original values; (b) log-log plot.

4 Data Analysis

In this section, we perform in-depth analyses on POI names. We organize this section into three subsections based on the three core components of gazetteers [16]. Thus, the first subsection focuses on *place names*, and aims to identify the local-specific terms used in these POI names. The second subsection looks into the interaction between POI names and *place types*, and examines the usage of local terms in different POI types. Finally, the third subsection analyzes the change of POI names with geographic distance based on the *spatial footprints* of the POIs.

4.1 Identifying local terms from POI names

In this first analysis, we attempt to answer the question: *what are the local terms used in the names of POIs in a geographic area?* While not every POI name contains local specific terms, some names are influenced by local factors, such as the “Vol” example discussed in the Introduction. We consider local terms as those frequently used in a local geographic area but less likely to be used in other areas. Identifying these local terms can help enhance computational models for place name disambiguation. We make use of the technique, term frequency and inverse document frequency (TF-IDF), a method commonly used in information retrieval, and adapt it to the context of geography. Equation 2 shows the adapted version of TF-IDF.

$$w_{ij} = tf_{ij} \times \log \frac{|G|}{|G_j|} \quad (2)$$

where w_{ij} is the weight of a term j in geographic area i , tf_{ij} is the frequency of term j in area i , $|G|$ is the total number of geographic areas in a study (which is seven in our case), and $|G_j|$ is the number of geographic areas that contain the term j . TF-IDF will highlight the terms that are frequently used in a local area, while reducing the weights of those commonly exist in POI names everywhere. In fact, the weights of the terms that occur in all seven metropolitan areas will become zero based on Equation 2.

Before applying the adapted TF-IDF to the POI names, we perform several data pre-processing steps. All POI names are converted to lowercase, and punctuations in POI names are removed. We did not remove typical stop words, such as “the” and “of”, since the term frequencies in POI names are not the same as other natural language texts, as shown in the



■ **Figure 4** Local terms identified based on the POI names in the seven US metropolitan areas.

exploratory analysis. Thus, typical stop words may not be stop words in the names of POIs. We also performed one special step for this analysis by counting the exact same POI names only once within a metropolitan area. The rationale behind this step is that term frequency can be increased in two situations: 1) one term is used by many different POIs (e.g., the term “Vol” is used in the names of many POIs); and 2) one word is used by the same POI business which simply shows up many times in a metropolitan area (e.g., “walmart”). We would prefer to keep the terms in the first situation, since those are endorsed by many different POIs and are more likely to be valid local terms than those in the second situation. After removing these repeating POI names, we group the names that belong to the same metropolitan areas using the bag-of-words model. We then use the adapted TF-IDF to identify local terms. Figure 4 shows the top 30 local terms identified for each of the seven metropolitan areas.

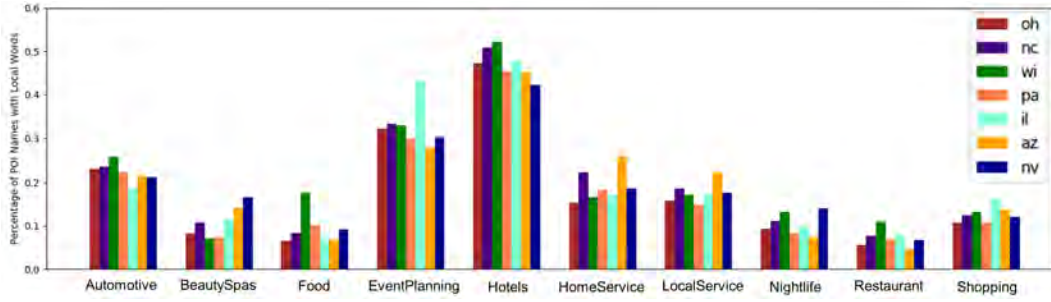
We can group the identified local terms into the following categories:

- **City names:** This is the most common type. POI names in all seven metropolitan areas contain city names, such as *scottsdale*, *las vegas*, *charlotte*, and *cleveland*.
- **State names:** This is similar to city names. State names, such as *arizona* and *wisconsin*, are used in POI names. There are also name abbreviations, such as *az* and *wi*.
- **Natural features:** Examples include *desert* and *canyon* in Phoenix and Las Vegas areas, *prairie* in Madison and Urbana-Champaign areas, and *rivers* in Pittsburgh area.
- **Sports teams:** Examples include *badger* in Wisconsin and *illini* in Illinois.
- **Family names:** A notable example is *zimbrick* in Madison, Wisconsin, which is a regional car dealer started by *John Zimbrick* (<http://www.zimbrickbuickgmceast.com/Zimbrick-History>).
- **Local cultures:** Terms such as *sin* and *casino* are observed in the POI names in Las Vegas, while the term *steel* is observed in the POI names in Pittsburgh area.

4.2 Examining local term usage in different POI types

The first analysis identified the local terms used in POI names in each geographic area. However, do POIs in different types have similar probabilities in using local terms as part of their names? In addition, are there regional differences in using local terms for names among POI types? In this second analysis, we aim to answer these questions.

In order to examine the interaction between POI names and POI types, we need to first divide the dataset based on POI types. Yelp has grouped their POIs into 23 root categories



■ **Figure 5** The percentages of POI names that contain local terms across POI types and different metropolitan areas.

which include *Restaurants*, *Shopping*, *Food*, *Hotels & Travel*, and other categories. We make use of these Yelp POI categories, and the POIs in each metropolitan area are divided into subsets based on their categories. Yelp allows one POI to belong to multiple categories (e.g., one POI can be both *Restaurants* and *Nightlife*), and therefore the same POI is put into more than one subset when multiple categories exist. Not all metropolitan areas contain POIs in all 23 categories. In addition, one metropolitan area may contain only a small number of POIs in a certain category, which can cause a biased result if those POIs are directly used for analysis. Thus, we only examine the POI types which are shared by all seven metropolitan areas and have at least one hundred POI instances in each area. Based on these criteria, we are left with ten categories, which are *Automotive*, *Beauty & Spas*, *Food*, *Event Planning & Services*, *Hotels & Travel*, *Home Services*, *Local Services*, *Nightlife*, *Restaurants*, and *Shopping*. The TF-IDF weights from the first analysis are then re-used, and we extract the top 100 terms that have the highest TF-IDF weights in each metropolitan area and use them as the local terms. The percentage of POI names in each POI type that contain local terms is calculated using Equation 3:

$$Pr_{ij} = |LP_{ij}|/|P_{ij}| \quad (3)$$

where $|LP_{ij}|$ is the number of POI names that contain any of the local terms in metropolitan area i in POI type j , $|P_{ij}|$ is the total number of POI names in metropolitan area i in POI type j , and Pr_{ij} is the calculated percentage. The result is shown in Figure 5.

Two things can be observed in Figure 5. First, there is an uneven usage of local terms across POI types. Overall, it seems people (business owners) are more likely to include local terms in the names of hotels, event planning services, and automotive shops. In contrast, local terms are less likely to be used in the names of restaurants, shopping places, and beauty spas. This is understandable since we frequently see hotels (especially hotel chains) include city names as part of their names to indicate locations, such as *holiday inn charlotte center city*. Meanwhile, restaurant names may focus on describing food and cuisine styles to attract customers. Second, the uneven usage of local terms is highly consistent across the seven metropolitan areas. This result suggests that the identified local term usage patterns are not specific to a particular region but can be generalized to other geographic areas.

To quantify the similarity and difference of local term usage in different POI types across geographic regions, we employ Jensen-Shannon divergence (JSD), which measures the similarity between two probability distributions. Equation 4 and 5 show the calculation of Jensen-Shannon divergence, where $KLD(P||Q)$ is the Kullback–Leibler divergence. The output of JSD is in $[0, 1]$, with 0 indicating that the two distributions are highly similar and

1 suggesting that the two distributions are largely different.

$$JSD(P||Q) = \frac{1}{2}KLD(P||M) + \frac{1}{2}KLD(Q||M) \quad (4)$$

$$KLD(P||Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)} \quad (5)$$

JSD requires the input probabilities to sum to 1. To satisfy this criterion, we normalize the initial percentage values using Equation 6:

$$NPr_i = \frac{Pr_i}{\sum_j Pr_j} \quad (6)$$

We then iterate through the seven metropolitan areas and calculate the pair-wise JSD, and finally calculate the average JSD value (there are in total 21 values). The obtained average JSD is 0.007, suggesting that the local term usage in different POI types are highly similar across geographic regions. The findings in this subsection can help us select suitable POI types in future for building computational models. For example, in the task of place name disambiguation, we may choose to focus on the POI names of certain types, such as *Hotels* and *Automotive* rather than *Restaurant* and *BeautySpas*, to extract more local terms which can then be associated with the related place names.

4.3 Analyzing POI name change with geographic distance

In this third analysis, we examine the change of POI names with geographic distance. Many phenomena follow Tobler's First Law and show a distance decay effect. Do POI names, which reflect many underlying social and cultural processes, also show such an effect? Here, we look into the *collective similarity* of POI names between metropolitan areas, namely how the POI names in one area are overall similar or dissimilar to the POI names in another area. For instance, we may expect the Phoenix metropolitan area to have more similar POI names compared with the Las Vegas metropolitan area than with the Cleveland metropolitan area.

One major challenge for this analysis is how to measure the *collective similarity* of POI names between metropolitan areas. We propose two approaches to achieve this goal. The first and a straightforward approach is to group POI names in the same metropolitan area into a bag of words. This is similar to the TF-IDF approach discussed in our first analysis. However, we use only term frequency here, since TF-IDF artificially exaggerates the importance of local terms. While such an exaggeration is desired for local term extraction, it distorts the true frequencies of terms in POI names and therefore is not used in this analysis. We also do not remove the repeating POIs as we did in the first analysis. In short, we try to keep the POI names and term frequencies as they are in the real world in order to objectively model their change with geographic distance. The terms used in the POI names in each metropolitan area are combined together into a vector. We will refer to this approach as *count-based vector*. To formally define this approach, let r_1 and r_2 represent two geographic regions, and each region contains a set of POIs. We derive the vector for a geographic region by counting the frequencies of terms in POI names. A common vocabulary V is constructed based on all the terms of the POI names in a dataset. Thus, the names of POIs in the two regions, r_1 and r_2 , can be collectively represented as two vectors:

$$\langle w_{11}, w_{12}, \dots, w_{1|V}| \rangle \quad (7)$$

$$\langle w_{21}, w_{22}, \dots, w_{2|V}| \rangle \quad (8)$$

where $|V|$ represents the size of the vocabulary, and w_{ij} represents the count of term j used in the POI names in geographic region i .

While the count-based vector approach is straightforward, it does not capture the semantic similarity between terms. For example, the terms *kiku* and *sakana* are both used for the names of sushi restaurants in the dataset. The count-based vector will treat the two terms as completely different with a similarity of zero. However, the fact that these two terms both co-occur with *sushi* suggests there exists certain degree of similarity between them. *Word2vec* [34] is a model that has been found to effectively capture the semantic similarity between terms. It is a neural network model which learns *embeddings* (low dimension vectors) for terms. In this work, we use the word2vec model to learn embeddings for metropolitan areas based on POI names. The embeddings are learned by predicting the terms used in POI names based on a given region (e.g., what terms are likely to be used for POI names if the region is *Phoenix, AZ*). The embeddings are condensed vectors, and the POI names in r_1 and r_2 can be represented as the two vectors below:

$$\langle u_{11}, u_{12}, \dots, u_{1|d}| \rangle \quad (9)$$

$$\langle u_{21}, u_{22}, \dots, u_{2|d}| \rangle \quad (10)$$

where d is the dimensionality of the embeddings, which can be decided empirically. In this analysis, we set $d = 300$ following the recommendation from the literature [34]. u_{ij} is a weight value learned from the POI dataset. The word2vec model aims to minimize the objective function in Equation 11:

$$J = -\log\sigma(\mathbf{w}_o^T \mathbf{r}) - \sum_{k=1}^K \log\sigma(-\mathbf{w}_k^T \mathbf{r}) \quad (11)$$

where \mathbf{r} is the embedding of one geographic region, \mathbf{w}_o is the embedding of a term that is used for the POI names in region \mathbf{r} , while \mathbf{w}_k is the embedding of a term not used in region \mathbf{r} (which serves as negative samples). σ is a sigmoid function: $\sigma(x) = \frac{1}{1+e^{-x}}$.

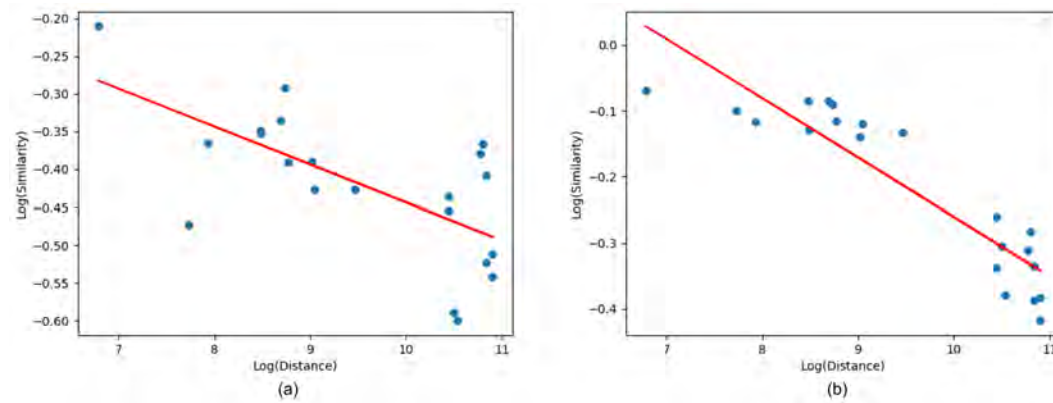
With different geographic regions represented as vectors in the same dimension, cosine similarity can be employed to measure the similarity of two vectors (Equation 12). $s(r_1, r_2)$ is then used as the collective similarity between regions r_1 and r_2 .

$$s(r_1, r_2) = \frac{\sum_{j=1}^d w_{1j} w_{2j}}{\sqrt{\sum_{j=1}^d w_{1j}^2} \sqrt{\sum_{j=1}^d w_{2j}^2}} \quad (12)$$

We apply both the count-based approach and word2vec to the Yelp POI dataset to derive vectors for the seven metropolitan areas. The center point of each metropolitan area is derived by averaging the location coordinates of the POIs in that area. We then employ Vincenty's formulae [48], which is based on the assumption of an oblate spheroid, to calculate the distance between two metropolitan areas. We then perform both Pearson's and Spearman's correlation to examine the relation between the collective similarity of POI names and the geographic distance of the corresponding metropolitan areas. Table 1 shows the correlation results. Overall, the collective similarity of POI names negatively and significantly correlates with geographic distance based on the four correlation coefficients in Table 1, which suggests that POI names indeed *gradually* become less similar with the increase of geographic distance. We emphasize *gradually* here because either no change or abrupt change can lead to no correlation between POI name similarity and geographic distance. It is often natural to assume that place names at different locations are of course different, but our experiment result suggests that place names are not randomly different

■ **Table 1** Pearson and Spearman correlation coefficients between the collective similarity of POI names and geographic distance.

	Count-based vector	word2vec
Pearson	-0.612 (p<0.01)	-0.963 (p<0.001)
Spearman	-0.626 (p<0.01)	-0.917 (p<0.001)



■ **Figure 6** Fitting the collective similarity of POI names with geographic distance: (a) count-based vector; (b) word2vec.

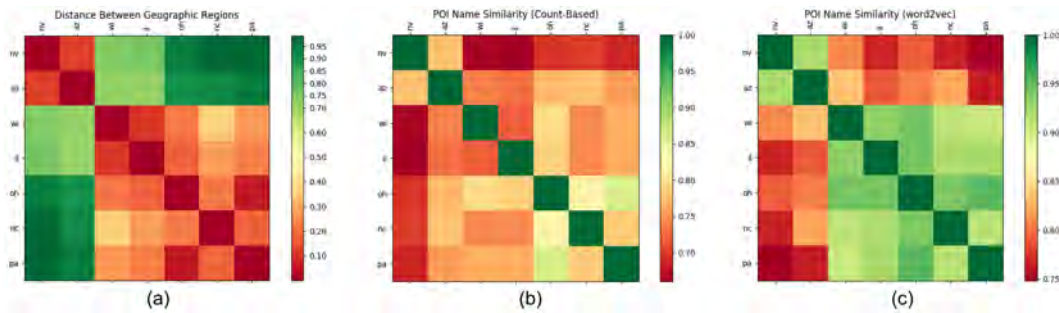
but follows a distance decay pattern. The statistical significance of the result is especially exciting given the fact that we have only 21 data points (21 region pairs from the seven metropolitan areas) for this correlation analysis. Such a result suggests that there is indeed a clear negative relation between POI name similarity and distance. In addition, it seems that word2vec better captures the POI name changes with geographic distance, as demonstrated by the higher correlation coefficients and stronger significances.

To further quantify the distance decay effect, we use a model $s = A * \frac{1}{d^\beta}$ to fit our data. We first transform it into its logarithmic form:

$$\log s = A + \beta * \log d \tag{13}$$

where s is the collective similarity of POI names between two metropolitan areas, A is a constant, β is the slope, and d is the geographic distance between them. We fit a linear regression model based on the logged values. Figure 6 shows the result. In the count-based vector approach, we obtained an R-squared value 0.434 and a slope of -0.050 . Using word2vec, we obtained a R-squared value 0.828 and a slope of -0.090 . More credibility can be given to the result from word2vec since it better captures the semantic similarity between terms in POI names. A slope of -0.090 indicates there is a clear distance decay effect with the increase of geographic distance. Besides, it is interesting to see how the data points clearly fall in two groups in Figure 6(b), which is consistent with their geographic distributions shown in Figure 2 (a group of city pairs has closer geographic distances, while the other group of city pairs has farther geographic distances). It would be interesting to examine the POI names in more metropolitan areas to see if their POI names also follow the general trend along the red line in Figure 6(b).

To further examine the result difference between the count-based vector and word2vec, Figure 7 shows the matrices of the geographic distances and the collective similarities obtained using the two approaches. It can be seen that the similarity pattern obtained using word2vec



■ **Figure 7** (a) The geographic distances between the seven metropolitan areas; (b) collective similarities based on count-based vector; (c) collective similarities based on word2vec.

in sub figure (c) is closer to the distance pattern in sub figure (a) compared with the pattern from the count-based vector in sub figure (b). This result is consistent with the distance decay pattern observed in Figure 6.

5 Conclusions and future work

Place names are texts given by people to natural or man-made geographic features. The act of assigning a name to space signifies the important moment of space and human experience integration, and further enhances the social construct of *place*. Place names, as *city-text*, reveal a considerable amount of information about the culture, lifestyle, community, and many other aspects of a city. While place names have long intrigued geographers, existing research often focuses on case-by-case qualitative descriptions related to the etymology or taxonomy of place names, or only considers place names as identifiers without analyzing their term usage and their relations with geographic distances.

This paper presents an empirical study on place names and their change with geographic distance. This study is based on an open dataset from Yelp, and examines more than 110,000 POIs, such as restaurants, hotels, and local services, in seven metropolitan areas in the United States. We perform an exploratory analysis on the frequencies of terms used in POI names, and find the term usage follows Zipf's law. We further conduct three analyses focusing on *place names*, *place types*, and *spatial footprints* respectively. We adapt the technique of term frequency and inverse document frequency in geographic context to identify local terms, and examine the term usage in the POI names in different types of POIs. We find an uneven usage of local terms across POI types (e.g., auto repairs are more likely to use local terms than restaurants), and such a usage pattern is highly consistent across different geographic regions. Finally, we test two approaches, count-based vector and word2vec, to model the collective similarity of POI names in different regions, and find a distance decay effect in the collective similarity of POI names.

This work is only a first step towards quantitatively and systematically examining place names and their relations with geographic distances. A number of topics can be explored in the near future. First, all the analyses are conducted based on the seven metropolitan areas available in the Yelp dataset. While a large number of POI names are examined, it would be interesting to apply the analyses to more metropolitan areas in other regions (e.g., north west and mid-south) as well as within local regions to further test the findings from this work. Second, we have so far used whole terms for the analyses, and it would be interesting to examine the parts or chunks of a term for measuring the collective similarity of place

names. For example, the place names, *Wawwatosa* in Wisconsin, *Wawatasso* in Minnesota, and *Wahwahtaysee* in Michigan, share similar chunks, and may have higher similarity values when a chunk-based approach is used. Third, future research can be conducted on how to integrate the information extracted from place names with existing computational models for tasks such as place name disambiguation. While Wikipedia articles and other datasets have been frequently used for training place-based models, there are situations when we have only short Wikipedia descriptions or no description for places. Local information extracted from place names can serve as additional resources to improve existing models.

References

- 1 Derek H Alderman. A street fit for a King: Naming places and commemoration in the American South. *The Professional Geographer*, 52(4):672–684, 2000.
- 2 Derek H Alderman. Street names as memorial arenas: The reputational politics of commemorating Martin Luther King in a Georgia county. *Historical Geography*, 30:99–120, 2002.
- 3 Derek H Alderman. Place, naming and the interpretation of cultural landscapes. *Heritage and Identity*, edited by Brian Graham and Peter Howard, pages 195–213, 2016.
- 4 Einat Amitay, Nadav Har’El, Ron Sivan, and Aya Soffer. Web-a-where: geotagging web content. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 273–280. ACM, 2004.
- 5 Maoz Azaryahu. Street names and political identity: the case of East Berlin. *Journal of Contemporary History*, 21(4):581–604, 1986.
- 6 Maoz Azaryahu. Renaming the past: Changes in "city text" in Germany and Austria, 1945-1947. *History and Memory*, 2(2):32–53, 1990.
- 7 Maoz Azaryahu. The power of commemorative street names. *Environment and Planning D: Society and Space*, 14(3):311–330, 1996.
- 8 Daniel L Baggio. *The dawn of a new Iraq: the story Americans almost missed*. US Army War College, 2006.
- 9 Davide Buscaldi and Paulo Rosso. A conceptual density-based approach for the disambiguation of toponyms. *International Journal of Geographical Information Science*, 22(3):301–313, 2008.
- 10 Paul Carter and Lawrie McKenzie. *The road to Botany Bay: an essay in spatial history*. Faber & Faber London, 1987.
- 11 James A Cheshire and Paul A Longley. Identifying spatial concentrations of surnames. *International Journal of Geographical Information Science*, 26(2):309–325, 2012.
- 12 Judith Gelernter and Nikolai Mushegian. Geo-parsing messages from microtext. *Transactions in GIS*, 15(6):753–773, 2011.
- 13 Michael F Goodchild and Linda L Hill. Introduction to digital gazetteer research. *International Journal of Geographical Information Science*, 22(10):1039–1044, 2008.
- 14 Milan Gritta, Mohammad Taher Pilehvar, Nut Limsopatham, and Nigel Collier. What’s missing in geographical parsing? *Language Resources and Evaluation*, pages 1–21, 2017.
- 15 Christian Grothe and Jochen Schaab. Automated footprint generation from geotags with kernel density estimation and support vector machines. *Spatial Cognition & Computation*, 9(3):195–211, 2009.
- 16 Linda L Hill. Core elements of digital gazetteers: placenames, categories, and footprints. In *International Conference on Theory and Practice of Digital Libraries*, pages 280–290. Springer, 2000.

- 17 Yingjie Hu, Krzysztof Janowicz, and Sathya Prasad. Improving Wikipedia-based place name disambiguation in short texts using structured data from DBpedia. In *Proceedings of the 8th workshop on geographic information retrieval*, pages 1–8. ACM, 2014.
- 18 Suradej Intagorn and Kristina Lerman. Learning boundaries of vague places from noisy annotations. In *Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems*, pages 425–428. ACM, 2011.
- 19 Krzysztof Janowicz and Carsten Keßler. The role of ontology in improving gazetteer interaction. *International Journal of Geographical Information Science*, 22(10):1129–1157, 2008.
- 20 Christopher B Jones and Ross S Purves. Geographical information retrieval. *International Journal of Geographical Information Science*, 22(3):219–228, 2008.
- 21 Christopher B Jones, Ross S Purves, Paul D Clough, and Hideo Joho. Modelling vague places with knowledge from the Web. *International Journal of Geographical Information Science*, 22(10):1045–1065, 2008.
- 22 Yiting Ju, Benjamin Adams, Krzysztof Janowicz, Yingjie Hu, Bo Yan, and Grant McKenzie. Things and strings: improving place name disambiguation from short texts by combining entity co-occurrence with topic modeling. In *European Knowledge Acquisition Workshop*, pages 353–367. Springer, 2016.
- 23 Morteza Karimzadeh, Wenyi Huang, Siddhartha Banerjee, Jan Oliver Wallgrün, Frank Hardisty, Scott Pezanowski, Prasenjit Mitra, and Alan M MacEachren. GeoTxt: a web API to leverage place references in text. In *Proceedings of the 7th workshop on geographic information retrieval*, pages 72–73. ACM, 2013.
- 24 Robin A Kearns and J Ross Barnett. To boldly go? Place, metaphor, and the marketing of Auckland’s Starship Hospital. *Environment and planning D: Society and space*, 17(2):201–226, 1999.
- 25 Robin A Kearns and Lawrence D Berg. Proclaiming place: Towards a geography of place name pronunciation. *Social & Cultural Geography*, 3(3):283–302, 2002.
- 26 Carsten Keßler, Patrick Maué, Jan Heuer, and Thomas Bartoschek. Bottom-up gazetteers: Learning from the implicit semantics of geotags. *GeoSpatial semantics*, pages 83–102, 2009.
- 27 Jochen L Leidner. *Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names*. Universal-Publishers, 2008.
- 28 Linna Li and Michael F Goodchild. Constructing places from spatial footprints. In *Proceedings of the 1st ACM SIGSPATIAL international workshop on crowdsourced and volunteered geographic information*, pages 15–21. ACM, 2012.
- 29 Michael D Lieberman, Hanan Samet, and Jagan Sankaranarayanan. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *2010 IEEE 26th International Conference on Data Engineering (ICDE)*, pages 201–212. IEEE, 2010.
- 30 Duncan Light. Street names in bucharest, 1990–1997: exploring the modern historical geographies of post-socialist change. *Journal of Historical Geography*, 30(1):154–172, 2004.
- 31 Paul A Longley, James A Cheshire, and Pablo Mateos. Creating a regional geography of Britain through the spatial analysis of surnames. *Geoforum*, 42(4):506–516, 2011.
- 32 Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.
- 33 Grant McKenzie, Krzysztof Janowicz, Song Gao, Jiue-An Yang, and Yingjie Hu. POI pulse: A multi-granular, semantic signature-based information observatory for the interactive visualization of big geosocial data. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 50(2):71–85, 2015.
- 34 Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.


- 35 Catherine Nash. Irish placenames: Post-colonial locations. *Transactions of the Institute of British Geographers*, 24(4):457–480, 1999.
- 36 Tessio Novack, Robin Peters, and Alexander Zipf. Graph-based strategies for matching points-of-interests from different vgi sources. In *AGILE 2017*, pages 1–6, 2017.
- 37 Simon Overell and Stefan Ruger. Using co-occurrence models for placename disambiguation. *International Journal of Geographical Information Science*, 22(3):265–287, 2008.
- 38 Kari Palonen. *Reading street names politically*. na, 1993.
- 39 Pauliina Raento and William A Douglass. The naming of gaming. *Names*, 49(1):1–35, 2001.
- 40 Reuben Rose-Redwood, Derek Alderman, and Maoz Azaryahu. Geographies of toponymic inscription: new directions in critical place-name studies. *Progress in Human Geography*, 34(4):453–470, 2010.
- 41 Reuben S Rose-Redwood. From number to name: symbolic capital, places of memory and the politics of street renaming in New York City. *Social & Cultural Geography*, 9(4):431–452, 2008.
- 42 Reuben S Rose-Redwood. "sixth avenue is now a memory": Regimes of spatial inscription and the performative limits of the official city-text. *Political Geography*, 27(8):875–894, 2008.
- 43 Joao Santos, Ivo Anastacio, and Bruno Martins. Using machine learning methods for disambiguating place references in textual documents. *GeoJournal*, 80(3):375–392, 2015.
- 44 Rui Santos, Patricia Murrieta-Flores, Pavel Calado, and Bruno Martins. Toponym matching through deep neural networks. *International Journal of Geographical Information Science*, 32(2):324–348, 2018.
- 45 Yi-Fu Tuan. *Space and place: The perspective of experience*. U of Minnesota Press, 1977.
- 46 Florian A Twaroch and Christopher B Jones. A web platform for the evaluation of vernacular place names in automatically constructed gazetteers. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*, page 14. ACM, 2010.
- 47 Maria Vasardani, Stephan Winter, and Kai-Florian Richter. Locating place names from place descriptions. *International Journal of Geographical Information Science*, 27(12):2509–2532, 2013.
- 48 Thaddeus Vincenty. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey review*, 23(176):88–93, 1975.
- 49 Jan Oliver Wallgrun, Morteza Karimzadeh, Alan M MacEachren, and Scott Pezanowski. GeoCorpora: building a corpus to test and train microblog geoparsers. *International Journal of Geographical Information Science*, 32(1):1–29, 2018.
- 50 John Kirtland Wright. The study of place names recent work and some possibilities. *Geographical Review*, 19(1):140–144, 1929.
- 51 Bo Yan, Krzysztof Janowicz, Gengchen Mai, and Song Gao. From ITDL to Place2Vec—Reasoning About Place Type Similarity and Relatedness by Learning Embeddings From Augmented Spatial Contexts. *Proceedings of 2017 ACM SIGSPATIAL Conference*, 17:7–10, 2017.
- 52 Wilbur Zelinsky. Along the frontiers of name geography. *The Professional Geographer*, 49(4):465–466, 1997.

Outlier Detection and Comparison of Origin-Destination Flows Using Data Depth

Myeong-Hun Jeong¹

Department of Civil Engineering, Chosun University, Gwangju, Republic of Korea


mhjeong@chosun.ac.kr

 <https://orcid.org/0000-0003-4850-8121>

Junjun Yin²

Social Science Research Institute; Institute for CyberScience, Penn State University, PA, USA

jjyin@psu.edu

 <https://orcid.org/0000-0002-4196-2439>

Shaowen Wang³

CyberGIS Center for Advanced Digital and Spatial Studies; Department of Geography and

Geographic Information Science, University of Illinois at Urbana-Champaign, IL, USA

shaowen@illinois.edu

Abstract

Advances in location-aware technology have resulted in massive trajectory data. Origin-destination (OD) trajectories provide rich information on urban flow and transport demand. This study describes a new method for detecting OD flows outliers and conducting hypothesis testing between two OD flow datasets in terms of the variations of spatial extent, that is, spread. The proposed method is based on data depth, which measures the centrality and outlyingness of a point with respect to a given dataset in \mathbb{R}^d . Based on the center-outward ordering property, the proposed method analyzes the underlying characteristics of OD flows, such as location, outlyingness, and spread. The ability of the method to detect OD anomalies is compared with that of the Mahalanobis distance approach, and an F-test is used to verify the difference in scale. Empirical evaluation has demonstrated that our method effectively identifies OD flows outliers in an interactive way. Furthermore, the method can provide new perspectives such as spatial extent by considering the overall structure of data when comparing two different OD flows in terms of scale.

2012 ACM Subject Classification Computing methodologies → Anomaly detection

Keywords and phrases Movement Analysis, Trajectory Data Mining, Data Depth, Outlier Detection

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.6

1 Introduction

With ubiquitous geolocation-aware sensors, knowledge discovery is greatly enhanced by extracting and mining interesting patterns from spatiotemporal big data in various domains. Massive movement data are collected to track people, animals, vehicles, and even natural

¹ This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2018R1C1B5043892).

² This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation grant number ACI-1548562

³ This work was supported by the U.S. National Science Foundation (grant numbers: 1047916 and 1443080)



© Myeong-Hun Jeong, Junjun Yin, and Shaowen Wang;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 6; pp. 6:1–6:14

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

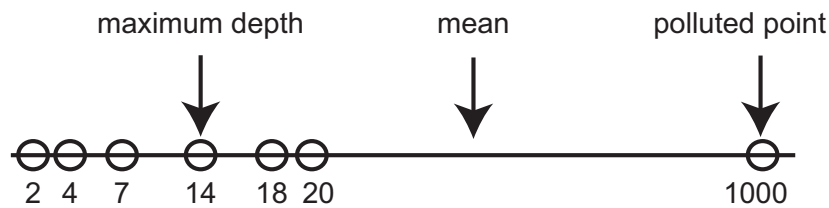
phenomena. Such data help us better model moving objects and reveal hidden patterns that are important to urban planning [17], understanding human mobility [30, 11], achieving the sustainability of urban systems [1, 3] and the environment [4], and improving public security and safety [2].

This paper a new method that identifies origination-destination (OD) flow anomalies and conducts hypothesis testing between two sets of different OD flows. In this study, the OD flow data represents a particular type of trajectory data, which records the origin and destination of each movement while ignoring the exact trajectory route [9]. The method was applied to OD flows derived from New York City taxi trip records, in which each record contains the origin and destination of each trip, without intermediate locations of the actual routes.

In recent years, researchers have investigated a variety of approaches to trajectory data mining. Most contemporary trajectory mining methods can be classified into four categories: clustering, classification, frequent/group pattern mining, and outlier detection [18, 33]. These methods can be used independently or together for trajectory mining applications. This study focuses on outlier detection of OD flows. Outlier detection aims to identify trajectories that do not follow the typical flows of trajectory that characterize the connectivity between regions [18]. Euclidean distance is employed by [7, 13] to find outlier patterns from trajectories. Studies by [20, 14] question the Euclidean distance approach because of the loss of local features and unavailability when external factors, such as topography, land cover or weather condition, affect the trajectories. In their research, [20, 14] addresses this by using robust distance measurements, e.g., Mahalanobis distance [20] and relative distance [14]. Instead of using distance or density, anomalous trajectories are detected by exploiting comparisons of the structural features of each trajectory segment [31] and an isolation tree of trajectories [32]. Most of these methods are related to trajectory data analysis, and thus, it is reasonable to extend the application of these approaches to the identification of OD flow anomalies. To overcome the sensitivity of Euclidean distance-based approaches to non-normal data distribution and the difficulty of selecting parameters for anomaly detection techniques based on distance or density, this study employs robust statistics, such as data depth, to detect OD flow outliers.

Flow mapping, a type of visual analytics, is a common approach to analyzing OD flow data. Visual representations of massive movement data facilitate comprehensive exploration of data, in turn enabling interpretation and understanding of complex flow trends. Aggregation and generalization of movement data are frequently utilized to resolve visual clutter [9, 29]. While visual analytics can help to extract inherent patterns from massive data, it is difficult to quantitatively compare two sets of different OD flows based on hypothesis testing. In other words, it is complicated to comprehend how two OD flows differ and, more importantly, the magnitude of the difference, using a test of statistical significance. Recently published articles employ multidimensional spatial scan statistics [8] and local Ripley's K-function [23] to identify clusters of flow data based on statistical significance testing. In a similar vein, this paper applies bivariate hypothesis testing methods based on data depth to understand the difference between two OD flow datasets in the context of different spatial extents.

It is worth noting that flow mapping approaches frequently suffer from the modifiable areal unit problem (MAUP). Essentially, MAUP reflects the influence of different aggregations determined by location on the identification and representation of coherent patterns. Kernel-based flow estimation and smoothing are used to overcome different spatial resolutions [9]. Instead of attempting to find the best areal unit by which to partition urban space and aggregate the OD flows, this study adopted the established traffic analysis zones of New



■ **Figure 1** Robustness of halfspace depth for the univariate case.

York City as a base unit. That said, the proposed method can be adapted to other areal units. In this study, New York City taxi trip data includes origins and destinations within traffic analysis zones, while ignoring the intermediate locations of the actual routes. Note that it is not necessary to reconstruct individual movements for flow estimation (see [5]).

In summary, this paper presents a new algorithm which conducts outlier detection as well as hypothesis testing on OD flow data. Our approach investigates the central regions of OD flows, based on data depth, to detect OD flow anomalies and conduct hypothesis testing between two different OD flow datasets. We believe that our method for analyzing taxi trip data has the potential to aid administrative authorities to better understand crowd patterns for improving urban planning activities such as determining transportation investments.

The remainder of this paper is organized as follows: Section 2 overviews how to detect OD flow outliers and conduct hypothesis testing between two different OD flow datasets using the concept of data depth. Experimental design and the evaluation of the proposed method are presented in Section 3. These results are discussed in Section 4. Section 5 concludes this paper with a summary and future work perspectives.

2 Methods

2.1 Data Depth

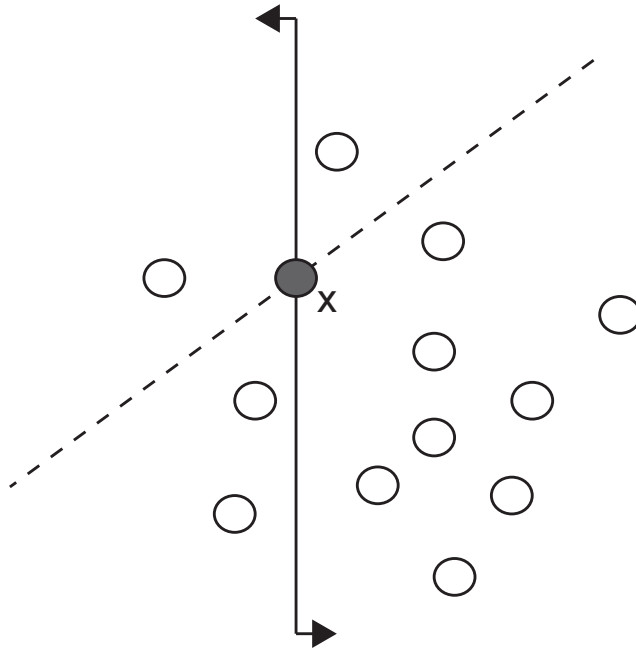
Data depth measures the centrality of a point with regard to a given dataset in \mathbb{R}^d . Originally developed by [24], the notion of data depth (i.e., halfspace depth) generalizes the univariate concept of ranking to multivariate data. Halfspace depth represents how deeply a point is located within a given dataset by ordering all points according to their degree of centrality.

Generally, the halfspace depth (HD) of point x in \mathbb{R}^d is defined as the minimum probability, P on \mathbb{R}^d , associated with any closed halfspace containing x [34].

$$HD(x; P) = \inf\{P(H) : H \text{ is a closed halfspace}, x \in H\}, x \in \mathbb{R}^d.$$

For the univariate case, all values less than or equal (greater than or equal) to x form a closed halfspace. All values less (greater) than x are an open halfspace. The smallest probability associated with two closed halfspaces developed by x is the halfspace depth of point x . In Figure 1, the probability of values less than or equal to 4 is $2/7$ and the probability of values greater than or equal to 4 is $6/7$. Thus, the halfspace depth of 4 is $2/7$, which is the minimum probability carried by any closed halfspace containing 4. Furthermore, as the sample median, 14 has the largest halfspace depth. Note that the polluted point inflates the standard error of the sample mean, thereby distorting the view of the data.

Similarly, the halfspace depth of x for the bivariate case is defined by the minimal number of data points in any closed halfspace, which is determined by a hyperplane through x [21]. In Figure 2, the solid line through x is rotated by 180° . The halfspace depth of x is determined



■ **Figure 2** Halfspace depth for the bivariate case.

by the smallest portion of data separated by such a hyperplane. For example, the halfspace depth of x is $3/13$, as determined by the dotted line. However, the halfspace depth of x determined by the solid line is $4/13$. Therefore, the halfspace depth of x is $3/13$, which is the minimal number of data points in any closed halfspace through x .

The property of halfspace depth is a center-outward ordering of points in \mathbb{R}^d and is affine invariant [19]. These features make halfspace depth a useful tool in nonparametric inference, which leads to various applications such as data classification and cluster analysis [12, 10]. There are multiple approaches to calculating data depth, including halfspace depth [21], projection depth [25], and simplicial depth [15]. While the computational complexity of the projection approach is $\mathcal{O}(n^2)$ (where n is the number of points), the computational complexity of simplicial depth is $\mathcal{O}(n^3)$. This can significantly increase computing time when n is large. Thus, this paper uses the more efficient method proposed by [21], in which the computational complexity for both approaches is $\mathcal{O}(n \log n)$.

2.2 OD Flow Outlier Detection Based on Data Depth

The center-outward ordering in data depth is closely related to the detection of outliers. The upper level sets of data depth in \mathbb{R}^2 form the central regions. The most central region can be regarded as a median. Conversely, the lower level sets of data depth, which coincide with larger distances from the center, can be regarded as outlyingness. This concept was utilized by [22, 28] to generate bag plots, which are analogous to one-dimensional box plots based on data depth. This paper uses the bag plot to identify the outliers of OD flows. Before explaining the method of outlier detection, we first introduce a basic definition of OD flow.

► **Definition 1.** Origin-destination (OD) flow. The OD flow $OD_i = (o_i, d_i, c_i, ts_i, te_i)$ is the number of trips (c_i) from the origin ID (o_i) to destination ID (d_i) of traffic analysis zones between the start time (ts_i) and the end time (te_i), where $ts_i < te_i$.

Based on this basic definition, Figure 3 depicts bag plots representing the OD flows of New York City taxi data collected on May 21, 2014 and July 1, 2014 respectively. We exploited taxi data on May 21, 2014 because the National September 11 Memorial Museum and Pavilion was opened to the public on this date. We also randomly selected another data set on July 1, 2014. In Figure 3a, the deepest depth of OD flows, depth median, is represented by a star symbol. This point is surrounded by a dark blue bag, which contains the half of OD flows. This region is regarded as a central region of OD flows. The OD flows in the bag are the dominant patterns. Magnifying the bag by a factor of three, relative to depth median, constructs a fence, as indicated by the light-blue area. The fence is comparable to the whiskers of a one-dimensional boxplot. The OD flows outside the fence, represented by red circles, are outliers. Every OD pair is represented by a point in Figure 3. The x-axis indicates the counts of forward OD flows (e.g., the number of OD flows from origin ID 2 to destination ID 10), and the y-axis indicates the counts of reverse OD flows (e.g., the number of OD flows from origin ID 10 to destination ID 2) in Figure 3a.

The bag plot presents the data using the following attributes: location is represented by the depth median; spread or the spatial extent of bag; correlation or the orientation of the bag; and skewness, as represented by the shape of the bag and the fence [22]. In Figure 3a, we observe that some forward OD flows have higher counts than their paired reverse OD flows. We also note the relatively linear correlation between forward OD flows and reverse OD flows and the skewness of forward (reverse) OD flows.

It is also possible to detect the outliers of OD flows of two different time stamps. In Figure 3c, we visualize the OD flows recorded on two different days. Comparing the two sets of OD flows not only indicates the central region of OD flows, it also distinguishes the significantly different OD flows.

The OD flows in high activity areas of a city are more likely to have large trip volumes. We use set operations to detect such outliers. We regard OD flows on July 1 as the control dataset (*control*); OD flows on May 21 as test dataset (*test*); and the combination of two OD flows as combination dataset (*combination*) in Figure 3. Then we can calculate the intersection of three outliers sets ($control \cap test \cap combination$), which are represented as rectangle symbols in Figure 3d.

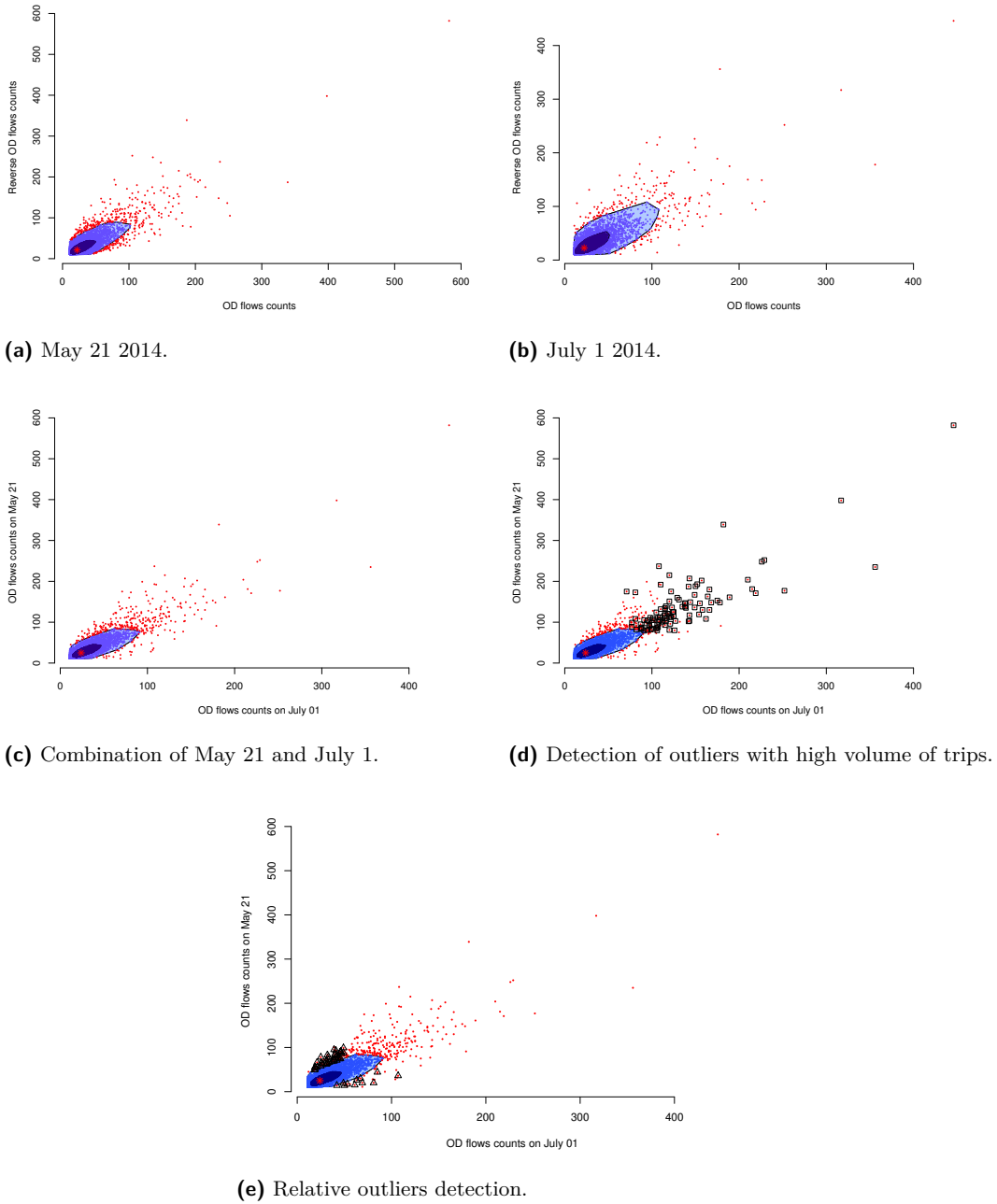
In addition, it is interesting to detect the outliers of OD flows which are typical patterns at time t_1 but atypical behaviors at time t_2 . We define the union of points in the bag, the central region, at time t_1 and t_2 . Then we calculate the intersection of two sets, the outliers of the combination set and the previous union set. These outliers are represented as triangle symbols in Figure 3e. These outliers are typical OD flows at time t_1 , located in the central regions in the bag plot. When we consider two OD flows together, they become unusual OD flows, some have more trips and some have fewer trips, relative to the control dataset. Thus, we can detect and treat outliers interactively based on data depth.

2.3 OD Flow Comparisons Based on Data Depth

Data depth can compare bivariate data from two independent groups. A t -test can be used to compare means from two independent groups. For example, the t -test reveals whether the means of two OD flows are different between two different temporal ranges. However, it is also worth examining how groups differ in terms of scale, which is also referred to as spread. Comparisons of central regions in data depth evaluate the marginal distribution, thereby considering the overall structure of the data [26].

Let X and Y be the random variables having distributions F and G for two independent groups. The quality index proposed by [16] is the probability that the depth of Y is greater than or equal to depth of X .

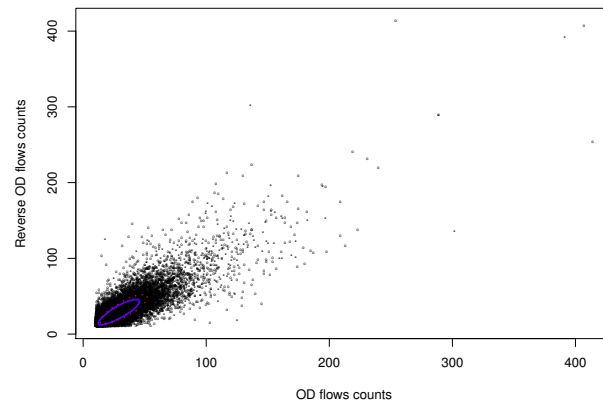
6:6 Outlier Detection and Comparison of Origin-Destination Flows using Data Depth



■ **Figure 3** Outliers detection of OD flows using a bag plot.

$$Q(F, G) = P[D(X; F) \leq D(Y; F)],$$

where P is the probability and $D(X; F)$ is the depth of randomly sampled observations according to distribution F . The range of Q , as presented by [16], is $[0, 1]$ and $Q(F, G) = 0.5$ if and only if $F = G$. If $Q < 0.5$ or if $Q > 0.5$, the scale increases or decreases from F to G . Therefore, it is possible to detect differences in scale using a bootstrap method.



■ **Figure 4** Central regions of two OD flows: \circ indicates the OD flows for Saturday, March 29 2014 and $*$ indicates the OD flows for a list of Saturdays; blue line presents the central region of the OD flows for the list of Saturdays and red dotted line presents the central region of the OD flows on March 29.

Let X_1, \dots, X_a be a random sample from F , and Y_1, \dots, Y_b be a random sample from G . The estimate of $Q(F, G)$ is calculated as shown below.

$$\hat{Q}(F, G) = \frac{1}{b} \sum_{i=1}^b R(Y_i; F_a),$$

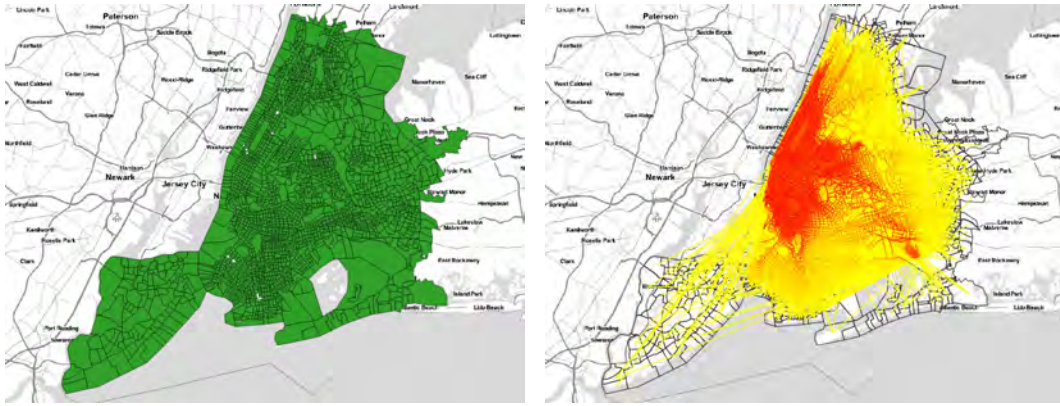
where $R(Y_i; F_a)$ indicates the proportion of X_j which has $D(X_j; F_a) \leq D(Y_i; F_a)$. Similarly, the estimate of $Q(G, F)$ can be defined as follows:

$$\hat{Q}(G, F) = \frac{1}{a} \sum_{i=1}^a R(X_i; G_b).$$

Bootstrap samples are obtained by resampling from the two groups (F and G). Under the null hypothesis ($H_0 : Q(F, G) = Q(G, F)$), the difference of the resulting bootstrap estimates is $Q^*(F, G) - Q^*(G, F)$. Thus, if the confidence interval of $Q(F, G) - Q(G, F)$ does not contain zero, we can reject the null hypothesis, H_0 [16, 26].

For ease of understanding, Figure 4 presents the central regions of two OD flows. One dataset is OD flows for Saturday, March 29, 2014, and the other dataset includes multiple Saturdays, those of March 1, 8, 15, 22, and April 5. At 552,064 taxi trips, the day of March 29 had the highest number of taxi trips for the year of 2014. The dataset for the other five Saturdays comprised 2,621,703 taxi trips. The bootstrap method reveals that the confidence interval is 0.0247 and 0.0596. This confidence interval does not include zero, thus rejecting the H_0 null hypothesis. This indicates that scale range is significantly changed between two OD flow datasets. Furthermore, the OD flows from the group of Saturdays are nested within the OD flows corresponding to March 29. This additional perspective was based on data depth comparisons.

The bootstrap method is a time consuming process. For this study, we generate 2,000 bootstrap samples. To improve the efficiency of the bootstrap computation, we distributed the work across multiple computing nodes and cores by implementing an embarrassingly parallel R code.



(a) 2,250 traffic analysis zones in New York City. (b) OD flows on July 1 2014.

■ **Figure 5** Experimental data: New York City taxi data.

3 Experiments

3.1 Data

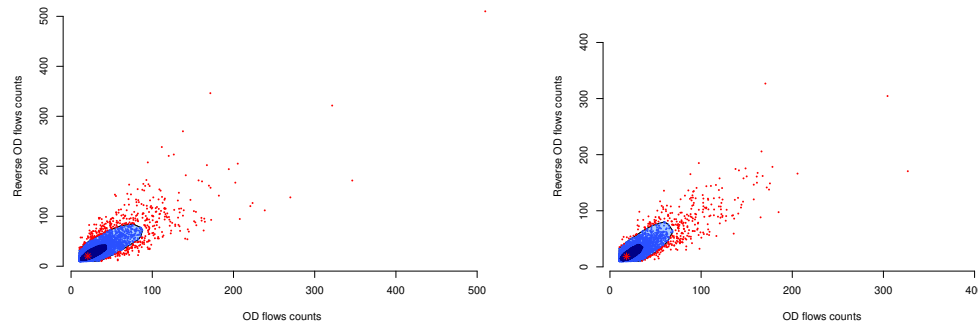
This study uses New York City taxi data collected in 2014 to evaluate the effectiveness of the proposed approach. Figure 5a presents traffic analysis zones in New York City which indicate the origin and the destination IDs of the OD flows. A traffic analysis zone (TAZ) is the most commonly adopted basic geographic unit in transportation planning models. The geographic areas of TAZ are delineated by transportation officials for tabulating traffic-related data. The size of TAZ varies because it accounts the underlying population in each zone, which consists of one or more census blocks, block groups, or census tracts. The shapes of the TAZs in this study are derived from the cartographic boundary shapefiles developed by the U.S. Census Bureau in conjunction with the 2010 census (<https://www2.census.gov/geo/tiger/TIGER2010/TAZ/2010/>). Considering the TAZs are particularly useful for journey-to-work and place-of-work statistics, we employed them as the basic units for accounting the taxi trips. Figure 5b shows OD flows on July 1. Red lines indicate the dominant OD flows.

As a case study, this paper examined OD flows recorded on weekdays and weekends in June 2014. The weekday dataset includes taxi trajectories collected on June 3, 10, 17, and 24, and represents 1,721,655 taxi trips. The weekend dataset includes taxi trajectories collected on June 8, 15, 22, and 29, and describes 1,593,480 trips.

3.2 Workflow

The performance of the proposed method was compared with alternative methods. Trajectory anomaly detection based on Mahalanobis distance [20] was used to evaluate the performance of outliers detection by the proposed method. The Mahalanobis distance is distinguished from Euclidean distance by its consideration of the correlations of the data, in this case, the two OD flow datasets. According to [20], the anomaly detection threshold can be defined as follows:

$$d_M(OD_{t_1}, \mu_{[t_0, t_1]}) \geq 3 \cdot \sqrt{\frac{1}{N} \sum_{t \in [t_0, t_1]} (OD_t - \mu_{[t_0, t_1]})^2}.$$



(a) Bag plot on weekdays.

(b) Bag plot on weekends.

■ **Figure 6** Outliers detection of OD flows: X-axis indicates forward OD flows counts and Y-axis indicates reverse OD flows counts.

where OD_{t_1} is the current OD flow, and $\mu_{[t_0, t_1]}$ is the median of all OD flows during $[t_0, t_1]$. In addition, we visualized the results in order to compare them and make the difference easier to understand. The difference of scale was evaluated using standard statistics, such as F-test, to compare the variance of two datasets.

For data cleaning process, this study used Hadoop with Pig. We developed a Hadoop program to resolve large data volume, which was composed of 173 million taxi trip records, remove trips with invalid OD coordinates, and assign each OD locations into the corresponding traffic analysis zone. To implement the OD flow outliers detection, this study used R. The computing environment used Amazon Web Service and the Bridges supercomputer at the Pittsburgh Supercomputing Center. This study only evaluated OD flows more than 10 trips, as the low trip number OD flows could have distorted the view of the data. All the code will be released as open source (the link to the code is available upon request).

3.3 Case study: weekdays vs weekends

3.3.1 Outlier Detection

The bag plots presented OD flow outliers on weekdays and weekends in Figures 6a and 6b, respectively. The outliers are detected by considering forward OD flows and reverse OD flows together.

To find the difference between two datasets, we considered two forward OD flows together with the bag plot. Then, we identified the outliers OD flows in Figure 7a. The outliers with rectangle symbols indicate OD flows with large volumes of taxi trips during weekdays and weekends. Figure 7b depicts these outliers superimposed on a map with red lines. The yellow lines represent the other OD flows, excluding the large volume OD flows on weekdays and weekends. This case clearly demonstrates that most OD flows occurred in three broad areas: within Manhattan, between the center of Manhattan and the two major airports (J.F.K International Airport and LaGuardia Airport), and between the two airports.

In addition, we investigated abnormal weekend OD flows that are typical weekday OD flows. These abnormal weekend OD flows exhibited substantial variance in number of taxi trips relative to their weekday counterparts. Figure 8a presents these OD flows outliers with triangle symbols. In Figure 8b, red lines indicate the substantial increases in weekend trip volumes. Conversely, blue lines indicate the decreases in trip volumes. Figure 8b reveals

6:10 **Outlier Detection and Comparison of Origin-Destination Flows using Data Depth**

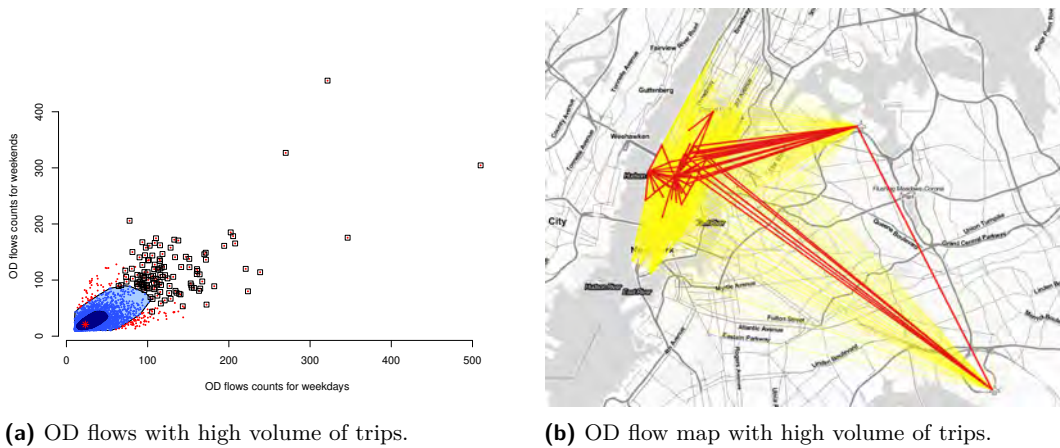


Figure 7 Outliers with high volume of trips on weekdays and weekends: Rectangles in Figure 7a coincide with red lines in Figure 7b.

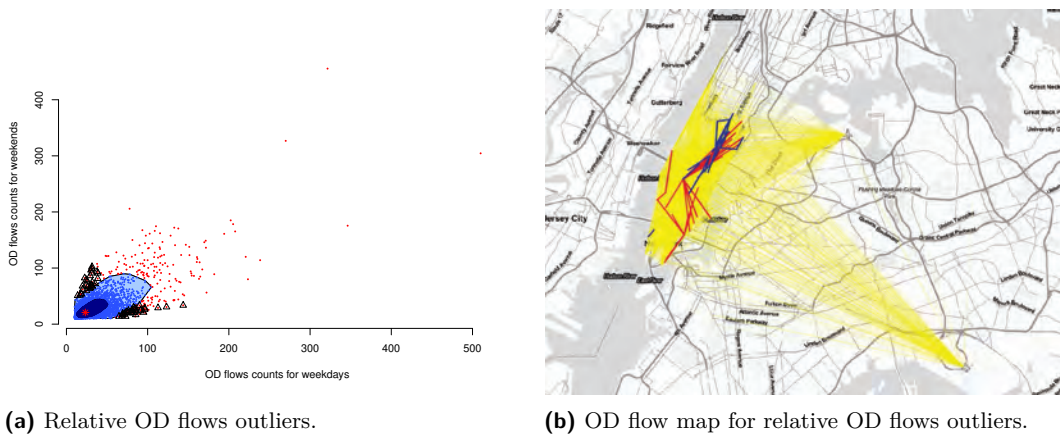
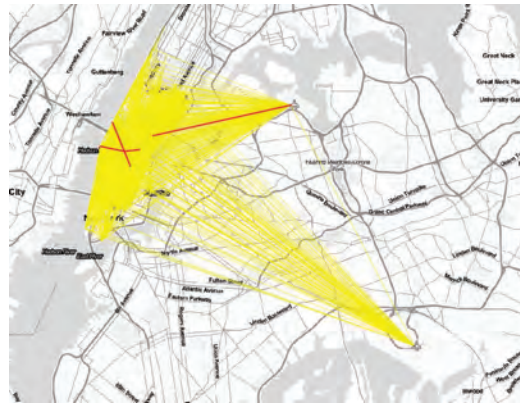


Figure 8 Relative OD flows outliers on weekdays and weekends: Triangles in Figure 8a coincide with red and blue lines in Figure 8b.

that OD flows between the center of Manhattan and the two airports or between the two airports were not significantly different during weekdays and weekends. However, we did observe some meaningful decrease in OD flows during the weekends in business district, as depicted by the blue lines in Figure 8b.

We also detected outlier OD flows using Mahalanobis distance. The results are presented in Figure 9. Far fewer outlier OD flows were detected using Mahalanobis distance than by our method. The Mahalanobis method only considers the forward OD flows of the two datasets. It identified OD flow outliers with high volume of trips because Mahalanobis distance considers the correlations between two OD flows. Thus, Mahalanobis distance is more likely to identify outliers when two OD flows have large trip volumes. In fact, the OD flows outliers from Mahalanobis distance are a subset of the outliers identified by our method, as depicted in Figure 7b. Furthermore, the Mahalanobis distance approach could not detect the outliers detected by our method in Figure 8 because the Mahalanobis distance approach cannot compare two flows to evaluate significant increases or decreases.



■ **Figure 9** Outlier OD flows on weekdays and weekends based on Mahalanobis distance.

3.3.2 Scale Comparisons

We further investigated how two OD flows differ. Our approach is sensitive to the difference in scale. Hypothesis testing of the differences between two central regions in Figure 10 inadvertently revealed that the confidence interval was -0.0277 and 0.0157 , which includes zero. Thus, it failed to reject the null hypothesis. The two central regions were similar in terms of the spread.

Interestingly, the standard statistic F-test was significant, $F(9530, 7637) = 1.1786$, $p \leq 0.05$. The variances of two groups were significantly different. The result of F-test directly opposed that of our method.

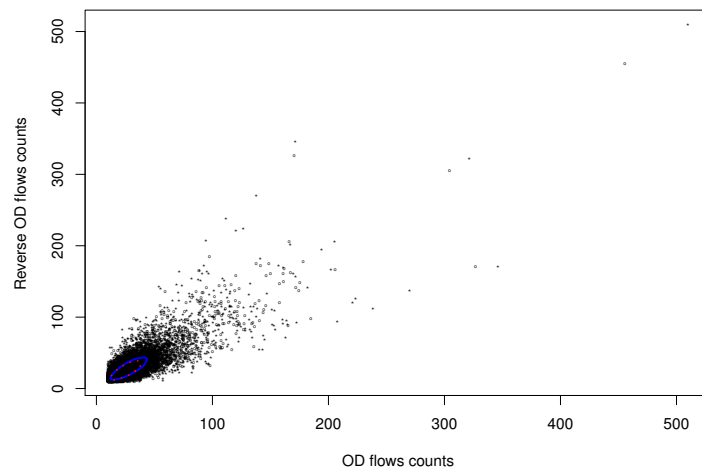
4 Discussion

The results demonstrate that the method effectively identifies outlier OD flows based on data depth. It is also feasible to detect outlier OD flows by querying with conditional clauses, such as which outlier OD flows always have high trip volumes during time t_1 and time t_2 .

As an alternative, the state-of-the-art Mahalanobis distance approach detected similar outlier OD flows. However, the number of outliers detected was different. This occurred because the proposed method's OD flows data had heavy tail distributions, which means many of the OD flows with a long distance from the depth median depicted in Figure 8a. Mahalanobis distance is known to be inadequate when the underlying data have heavy tail distributions [27]. Thus, the presence of outliers may mask the detection of other outliers in Mahalanobis distance approach. Furthermore, it can only detect OD flow outliers with high numbers of trips during time t_1 and time t_2 . It is difficult to detect OD flows outliers that have different properties, such as substantial differences in the number of trips when comparing between time t_1 and time t_2 .

In terms of the difference in spread, our method used a bootstrap technique to compare the central regions of data depth. This technique investigated the difference in scale as well as the structure of data. It can provide information about how deeply points from group 1, OD flows at t_1 , tend to be located within group 2, OD flows at t_2 . General statistics such as F-test only provide their difference in variation and do not further specify how groups differ.

Interestingly, the F-test results revealed a statistically significant difference in terms of variation of OD flows on weekdays and weekends. Our approach showed no statistically significant differences. This contrast may be caused by the sensitivity of F-test to non-normality [6], which increases the Type-I error rate. Conversely, data depth makes no assumptions about the distributions of the underlying dataset.



■ **Figure 10** OD flows comparisons based on data depth: \circ indicates the OD flows on weekdays and $*$ indicates the OD flows on weekends; blue line presents the central region of the OD flows for the weekdays and red dotted line presents the central region of the OD flows on weekends.

5 Conclusions and Future Work

This paper describes a new method for identifying outlier OD flows and the difference in scale between two different OD flows at t_1 and t_2 . The new method is based on the concept of data depth. Data depth is robust statistics, which is suitable to non-Gaussian distribution of the underlying datasets. Compared with standard statistics, our method enhances understanding of the differences and the magnitude of the differences between two OD flow datasets.

This study made no attempt to incorporate geographic contexts such as locational circumstances or surrounding environment in understanding OD flows. Ultimately, further research should focus on integrating the analysis of OD flows with appropriate geographic contexts. Such research will lead to desirable knowledge discovery and better understanding of movement dynamics.

References

- 1 Marina Alberti, John M Marzluff, Eric Shulenberger, Gordon Bradley, Clare Ryan, and Craig Zumbrunnen. Integrating humans into ecology: Opportunities and challenges for studying urban ecosystems. *AIBS Bulletin*, 53(12):1169–1179, 2003.
- 2 Maike Buchin, Somayeh Dodge, and Bettina Speckmann. Similarity of trajectories taking into account geographic context. *Journal of Spatial Information Science*, 2014(9):101–124, 2014.
- 3 Chao Chen, Daqing Zhang, Zhi-Hua Zhou, Nan Li, Tülin Atmaca, and Shijian Li. B-planner: Night bus route planning using large-scale taxi GPS traces. In *2013 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 225–233. IEEE, 2013.
- 4 Srinivas Devarakonda, Parveen Sevusu, Hongzhang Liu, Ruilin Liu, Liviu Iftode, and Badri Nath. Real-time air quality monitoring through mobile sensing in metropolitan areas. In

- Proc. 2nd ACM SIGKDD International Workshop on Urban Computing*, page 15. ACM, 2013.
- 5 Matt Duckham, Marc van Kreveld, Ross Purves, Bettina Speckmann, Yaguang Tao, Kevin Verbeek, and Jo Wood. Modeling checkpoint-based movement with the earth mover's distance. In *International Conference on Geographic Information Science*, pages 225–239. Springer, 2016.
 - 6 Andy Field, Jeremy Miles, and Zoë Field. *Discovering statistics using R*. Sage, London, UK, 2012.
 - 7 Vitor Cunha Fontes, Lucas Andre de Alencar, Chiara Renso, and Vania Bogorny. Discovering trajectory outliers between regions of interest. In *Proc. XIV GeoInfo*, pages 49–60, 2013.
 - 8 Yizhao Gao, Ting Li, Shaowen Wang, Myeong-Hun Jeong, and Kiumars Soltani. A multidimensional spatial scan statistics approach to movement pattern comparison. *International Journal of Geographical Information Science*, 0(0):1–22, 2018.
 - 9 Diansheng Guo and Xi Zhu. Origin-destination flow data smoothing and mapping. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2043–2052, 2014.
 - 10 Myeong-Hun Jeong, Yaping Cai, Clair J Sullivan, and Shaowen Wang. Data depth based clustering analysis. In *Proc. 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 29. ACM, 2016.
 - 11 Mei-Po Kwan. Space-time and integral measures of individual accessibility: A comparative analysis using a point-based framework. *Geographical Analysis*, 30(3):191–216, 1998.
 - 12 Tatjana Lange, Karl Mosler, and Pavlo Mozharovskyi. Fast nonparametric classification based on data depth. *Statistical Papers*, 55(1):49–69, 2014.
 - 13 Jae-Gil Lee, Jiawei Han, and Xiaolei Li. Trajectory outlier detection: A partition-and-detect framework. In *IEEE 24th International Conference on Data Engineering*, pages 140–149. IEEE, 2008.
 - 14 Liangxu Liu, Shaojie Qiao, Yongping Zhang, and JinSong Hu. An efficient outlying trajectories mining approach based on relative distance. *International Journal of Geographical Information Science*, 26(10):1789–1810, 2012.
 - 15 Regina Y Liu. On a notion of data depth based on random simplices. *The Annals of Statistics*, pages 405–414, 1990.
 - 16 Regina Y Liu and Kesar Singh. A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association*, 88(421):252–260, 1993.
 - 17 Jean Damascène Mazimpaka and Sabine Timpf. Exploring the potential of combining taxi GPS and flickr data for discovering functional regions. In *AGILE 2015*, pages 3–18. Springer, 2015.
 - 18 Jean Damascène Mazimpaka and Sabine Timpf. Trajectory data mining: A review of methods and applications. *Journal of Spatial Information Science*, 2016(13):61–99, 2016.
 - 19 Karl Mosler. *Robustness and Complex Data Structures*, chapter Depth Statistics, pages 17–34. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.
 - 20 Bei Pan, Yu Zheng, David Wilkie, and Cyrus Shahabi. Crowd sensing of traffic anomalies based on human mobility and social media. In *Proc. 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 344–353. ACM, 2013.
 - 21 Peter J Rousseeuw and Ida Ruts. Algorithm AS 307: Bivariate location depth. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 45(4):516–526, 1996.
 - 22 Peter J Rousseeuw, Ida Ruts, and John W Tukey. The bagplot: A bivariate boxplot. *The American Statistician*, 53(4):382–387, 1999.
 - 23 Ran Tao and Jean-Claude Thill. Spatial cluster detection in spatial flow data. *Geographical Analysis*, 48(4):355–372, 2016.

- 24 John W Tukey. Mathematics and the picturing of data. In *Proc. International Congress of Mathematicians*, volume 2, pages 523–531, 1975.
- 25 Rand R Wilcox. Approximating Tukey’s depth. *Communications in Statistics-Simulation and Computation*, 32(4):977–985, 2003.
- 26 Rand R Wilcox. Two-sample, bivariate hypothesis testing methods based on Tukey’s depth. *Multivariate Behavioral Research*, 38(2):225–246, 2003.
- 27 Rand R Wilcox. *Introduction to robust estimation and hypothesis testing*. Academic Press, 2012.
- 28 Hans Peter Wolf and Uni Bielefeld. aplpack: Another Plot PACKage: stem.leaf, bagplot, faces, spin3r, plotsummary, plothulls, and some slider functions, 2014. R package version 1.3.0. URL: <https://CRAN.R-project.org/package=aplpack>.
- 29 Junjun Yin, Yizhao Gao, Zhenhong Du, and Shaowen Wang. Exploring multi-scale spatiotemporal twitter user mobility patterns with a visual-analytics approach. *ISPRS International Journal of Geo-Information*, 5(10):187, 2016.
- 30 Junjun Yin, Aiman Soliman, Dandong Yin, and Shaowen Wang. Depicting urban boundaries from a mobility network of spatial interactions: A case study of great britain with geo-located twitter data. *International Journal of Geographical Information Science*, 31(7):1293–1313, 2017.
- 31 Guan Yuan, Shixiong Xia, Lei Zhang, Yong Zhou, and Cheng Ji. Trajectory outlier detection algorithm based on structural features. *Journal of Computational Information Systems*, 7(11):4137–4144, 2011.
- 32 Daqing Zhang, Nan Li, Zhi-Hua Zhou, Chao Chen, Lin Sun, and Shijian Li. iBAT: Detecting anomalous taxi trajectories from GPS traces. In *Proc. 13th International Conference on Ubiquitous Computing*, pages 99–108. ACM, 2011.
- 33 Yu Zheng. Trajectory data mining: An overview. *ACM Transactions on Intelligent Systems and Technology*, 6(3):29, 2015.
- 34 Yijun Zuo and Robert Serfling. General notions of statistical depth functions. *The Annals of Statistics*, 28:461–482, 2000.

Is Saliency Robust? A Heterogeneity Analysis of Survey Ratings

Markus Kattenbeck

University Regensburg, Information Science, 93040 Regensburg, Germany
markus.kattenbeck@ur.de

Eva Nuhn

University Augsburg, Geoinformatics Group, 86135 Augsburg, Germany
eva.nuhn@geo.uni-augsburg.de

Sabine Timpf

University Augsburg, Geoinformatics Group, 86135 Augsburg, Germany
sabine.timpf@geo.uni-augsburg.de

Abstract

Differing weights for saliency subdimensions (e.g. visual or structural saliency) have been suggested since the early days of saliency models in GIScience. Up until now, however, it remains unclear whether weights found in studies are robust across environments, objects and observers. In this study we examine the robustness of a survey-based saliency model. Based on ratings of $N_o = 720$ objects by $N_p = 250$ different participants collected in-situ in two different European cities (Regensburg and Augsburg) we conduct a heterogeneity analysis taking into account environment and sense of direction stratified by gender. We find, first, empirical evidence that our model is invariant across environments, i.e. the strength of the relationships between the subdimensions of saliency does not differ significantly. The structural model coefficients found can, hence, be used to calculate values for overall saliency across different environments. Second, we provide empirical evidence that invariance of our measurement model is partly not given with respect to both, gender and sense of direction. These compositional invariance problems are a strong indicator for personal aspects playing an important role.

2012 ACM Subject Classification Mathematics of computing → Multivariate statistics, Human-centered computing → Personal digital assistants, Human-centered computing → Empirical studies in ubiquitous and mobile computing

Keywords and phrases Saliency Model, Measurement Invariance, Heterogeneity Analysis, PLS Path Modeling, Structural Equation Models

Digital Object Identifier 10.4230/LIPICs.GIScience.2018.7

Acknowledgements We would like to thank the persons willing to participate in our experiments. Furthermore, we are grateful to Ludwig Kreuzpointner and David Elsweller for their valuable feedback on a draft version of this paper.

1 Introduction

Models of saliency have seen increased interest over the last two decades (see [39, 33, 9, 4, 8, 5, 37, 22, 34, 32, 18, 11, 30]). These models are important for several different reasons: they deepen the understanding of human perception and support the interpretation of spatial situations and subsequent decision making; they are applicable to provide route instructions enriched with salient objects for in- and outdoor environments, which is the preferred mode



© Markus Kattenbeck, Eva Nuhn, and Sabine Timpf;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 7; pp. 7:1–7:16

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

of route communication between humans (see e.g. [40, 43, 3, 26]). Finally, they may be used to design environments which are conducive to wayfinding and navigation.

Given their practical utility several different ways of estimating the saliency of objects have been proposed over the years (see e.g. [33, 4, 37, 34, 41, 30]). There is, however, general agreement that saliency is not inherent to objects but ascribed to them by an observer, where both, observer and observed, share the same environment (see [4]). Saliency (and each of its proposed subdimensions, e.g. visual saliency) itself is, in statistical terms, a latent variable, i.e. it cannot be directly observed, but must be measured using a combination of variables. Subdimensions may differ depending on the selected model of saliency (see section 2), e.g. in the model by Sorrows and Hirtle [39] the four subdimensions visual, cognitive, structural saliency and prototypicality were proposed. Using an extension of this saliency model Kattenbeck [19] proposes a set of measured variables for each of five subdimensions and analyses the impact these have on each other and how these can be used to calculate the overall saliency of objects.

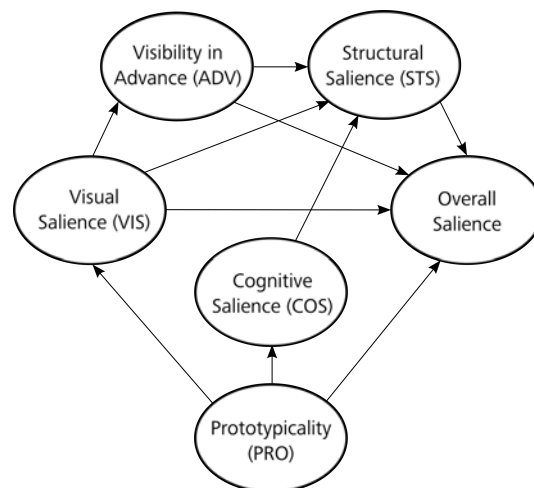
Survey-based methods are particularly useful with respect to this aspect because they allow to collect data in-situ. This study uses the survey developed in [18] to present an analysis of its measurement invariance. To this end, we collect a dataset of saliency ratings in Augsburg (Germany) and compare these ratings to those obtained in Regensburg, Germany (see [19]). The main goal of this paper is to assess measurement invariance with respect to environment, objects and observers of Kattenbeck's measurement model of saliency and to analyze the observed heterogeneity taking environment and sense of direction (stratified by gender) into account. The personal aspects were chosen for two reasons. First, there is evidence that differences between genders regarding the preferred mode of orientation exist (see [6] for an overview). Second, subdimensions of state of the art saliency models (see section 2) may be influenced by both, different levels (good vs. poor) and subdimensions of sense of direction (allocentric vs. egocentric vs. cardinal directions): for example, visual saliency might be more important for those with poorer orientation skills because visual dimensions do not require any knowledge of the structure of the space persons are navigating in.

2 Related Work

The interest in diverging degrees of saliency for different objects dates back to the 1960s [25, 1]. Subdimensions of saliency were, however, not distinguished before the turn of the century. Sorrows and Hirtle [39] distinguish four subdimensions influencing saliency:

1. visual saliency, which describes visual characteristics of an object (e.g. salient color, outstanding height),
2. cognitive saliency, which focuses on the meaning of a landmark (e.g. through cultural or historical importance),
3. structural saliency, which is important because of its location in the structure of the space and
4. prototypicality, which describes how typical an object is with respect to a category [36].

These subdimensions are not mutually exclusive. In contrast, a combination of all subdimensions contributes to the overall saliency ascribed to a single object. Many researchers use the classification by Sorrows and Hirtle [39] to develop their own models to assess the saliency of objects. Raubal and Winter [33] define independent characteristics of landmark saliency of objects based on visual attraction, semantic attraction and structural attraction. They do not consider prototypicality because extensive human subject testing would be required to derive useful results [33]. The aspect of prototypicality, however, plays an



■ **Figure 1** A graphical representation of the Structural Equation Model (i.e. its structural model part) presented in [19]. Table 4 provides the questions used as measured variables.

important role in the model presented in [8], where the usefulness of prototypes rather than particular object properties was used to determine cognitively salient landmarks.

Raubal and Winter's model [33] has been extended several times: Nothegger et al. [29] extend and test the model on façades of buildings. Their proof of concept based on real world data and human judgment shows that the model is a viable way to assess the saliency of landmarks. Winter [42] extends [33] by adding *advance visibility* as important factor for landmark saliency, i.e. a feature is more salient if it is identifiable earlier in a route than a feature that can only be spotted at the very last moment.

Klippel and Winter [23] complement landmark research with an approach to formalize structural saliency. They describe objects as structurally salient if “their location is cognitively or linguistically easy to conceptualize in route directions” [23, p. 347]. In their work they propose taxonomic considerations of point-like objects with respect to their position along a route.

A final extension to the original model stresses the importance of the observer. Caduff and Timpf [4] provide a strong argument that the saliency of landmarks is affected by the perspective of the observer, the surrounding environment and the objects contained therein. Saliency is contingent on the current navigational context [4], i.e. an object's saliency does not only depend on its individual attributes but also on its distinction with respect to attributes of objects nearby [33]. Saliency is, consequently, not an inherent property of an object but is assigned to an object by the observer.

Based on these developments, Kattenbeck [20, p. 2] provides the following definition:

Given a local environment an observer is in, (overall) saliency (OVSAL) is the degree to which an object, persistent enough to be used in route instructions, draws the average pedestrian observer's attention. This degree is evoked by:

1. visual features of the object (visual saliency - VIS),
2. the degree of prototypicality it shows (prototypicality - PRO),
3. how identifiable it is when approached (advance visibility - ADV),
4. the ease with which it may be integrated into a route description (structural saliency - STS) and
5. the degree as to which it can evoke prior knowledge (cognitive saliency - COS).

Overall saliency seems to be highly dependent on personal subdimensions (see also [32, 11, 30, 38]), since VIS, PRO, COG and ADV depend on either perception or cognition of the observer and only STS and, to a certain extent, ADV and VIS are influenced by the physical environment. Taking the definition above as basis, Kattenbeck [18] reported data collection based on a survey presented there (see table 4). The predictive capability of these ratings was shown in [18, 19, 20] by means of PLS-based Structural Equation Models and suggests highly intertwined subdimensions of saliency.

The goal of the present study is to follow up on these survey-based methods of saliency measurement. This means, we collect an additional dataset applying the method described in [19] in order to assess whether the model derived from the results presented there (see figure 1) shows invariance across different environments and user groups. We, therefore, use the same statistical method as was used in [19], i.e. we apply PLS-based estimations (see section 4 for a short introduction on this method) of structural equation models to the new dataset collected in Augsburg. We do this in order to gain a better understanding of the model of saliency, to determine if all necessary parameters have been included and to determine the robustness of the model.

3 Data Collection Method

In this study we analyze two different datasets of saliency ratings by individuals collected while walking predetermined routes under guidance of an experimenter. For the first city, Regensburg, which is a town in Southern Germany, the first author of this paper collected data throughout his PhD [19]. As the goal of this study is the analysis of measurement invariance, it was most important for the current study to gain a second dataset by collecting the data for Augsburg in exactly the same manner as described there [19]. The data collection method and the resulting dataset are detailed below. This data will be accessible via Data in Brief <https://www.journals.elsevier.com/data-in-brief> by the end of 2018.

3.1 City 1: Regensburg, Germany

The Regensburg dataset is built from $N_{rR} = 55$ routes with $N_{oR} = 362$ objects (on average, 6 objects per route), which were rated by $N_{pR} = 112$ participants (68 females, age range: 18-65 years, $\bar{x}_{ageR} = 25.46$ years). Experiments took 60 minutes on average (SD = 12 min, range: 38-113 min). The data was collected between November 2014 and February 2015 (see [19] for more details). The methods employed to find a sample of objects and conduct experiments were identical to those described for Augsburg below.

3.2 City 2: Augsburg, Germany

3.2.1 Selection of objects and routes

First, a sample of objects comparable to the one chosen for Regensburg had to be selected in Augsburg. In accordance to [19] it included salient as well as non-salient objects and, in addition, objects other than buildings (e.g. recycling bins, fountains or monuments) which can be referred to in route instructions. Therefore, geographical coordinates of 480 locations were generated randomly to gain a random sample of objects. The locations were inspected on-site. If an object or building was located at the coordinate, it was added to the sample. If neither a building nor any other object was located there the closest object in a randomly drawn direction was chosen. In case an object was not accessible (e.g. railways) they were excluded from the sample. Similarly, parked cars or other temporary objects were not added

to the sample. This resulted in a sample size of $N_{oA} = 352$ objects for Augsburg. The sampled objects were randomly combined into routes, such that the time required for a single experiment was expected to be no more than 60 minutes. We aimed for an average of 6 objects per route. Taking these preconditions into account, $N_{rA} = 59$ routes were derived for Augsburg. The walking direction of each route was chosen randomly and each route was assigned randomly to participants. As in [19] we aimed at two independent ratings for each object.

3.2.2 Procedure

Data acquisition for Augsburg took place as part of course work for a seminar. Students taking the class were carefully instructed such that they were able to carry out experiments on their own. Participants were acquired via verbal announcements in university lectures or directly by student experimenters. Two restrictions applied: First, participants had not taken part in a prior experiment on pedestrian navigation. Second, special care was taken to ensure that there was no relationship between participants and student experimenters to avoid biases. A custom designed Android application facilitated the data collection in [19] and this application was reused for our study in Augsburg. The experiments were conducted between July 2017 and December 2017.

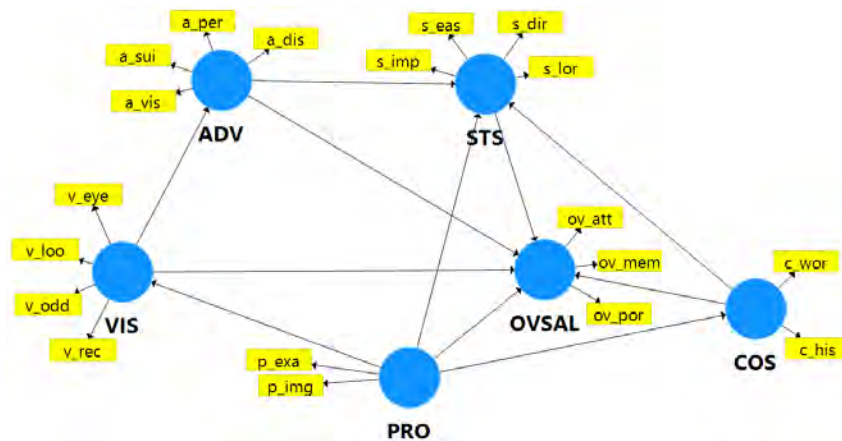
Each participant was guided on one of the routes by a student experimenter. Before walking the route, participants were asked to complete a demographic data questionnaire also comprising their personal interests. Participants completed, moreover, a German language self-report sense of direction survey [27]. On completion of these questionnaires, a picture of the first object to be rated was shown to the participants. Along the route, participants had to identify each of the objects on their own. Once the object had been identified they rated the object's salience by answering the questions presented in table 4. Having finished the survey, a picture of the upcoming object was displayed. Overall, $N_{pA} = 109$ (age range: 19-65 years, $\bar{x}_{ageA} = 25.97$ years, 38 females, 14% non-students) persons participated in Augsburg. The experiments took 51 min on average (SD = 13 min, range: 23-83 min). These values are comparable to those in Regensburg (see above). Unfortunately, due to issues with the mobile Internet connection the answers of 15 participants were lost. As a consequence, 90 objects were rated by only one person.

4 Statistical Analysis

Structural Equation Modeling is a multivariate statistical analysis technique that is used to analyze relationships between measured variables and latent constructs, i.e. between the five constructs describing salience and the measured variables to describe them (such as shape, age, length etc.). This section introduces PLS Path modeling as a statistical method and as an adequate means of assessing measurement invariance. This is an important property of a survey used to collect salience ratings: If given the survey measures the same construct across different environments, user groups etc. and weights do not need to be updated for different contexts.

4.1 PLS Path Modeling – A quick glance

In general, Structural Equation Models consist of two parts (see e.g. [12, p. 634f.]): The structural model part describes the relationships among latent variables (constructs), whereas the measurement model part establishes connections between each construct and the variables



■ **Figure 2** The model used for the analysis throughout this paper (see section 5.1 for the empirical reason to use reflective measurement for visual saliency).

used to measure its value (see figure 2). Constructs with outgoing arrows only are referred to as exogenous, whereas those with incoming arrows are known as endogenous variables. A set of measured variables (depicted as rectangles) is used to assess the value of each of these latent variables, as they cannot be observed directly. Measured variables are related to latent variables in one of two measurement modes [10]. *Reflective measurement* (indicated by arrows pointing to measured variables) assumes that the unknown value of the latent variable causes the observed values of the measured variables. In contrast, *formative measurement* causes (arrow heads point to the construct) are thought of as causing the latent variable's value (see [2]).

Two methods to estimate structural equation models exist. The covariance-based approach aims to maximize similarity between the model's and the empirical covariance matrix. It is, hence, based on the assumption of multivariate normality of the data. The variance-based approach, which is called PLS Path Modeling [44, 45] is, in contrast, not based on any distributional assumptions. It focuses, similar to other approaches involving regression, on prediction, i.e. it maximizes the amount of variance explained in the endogenous construct(s) [13, p. 140]. This predictive focus is particularly valuable in case of the analysis reported here, where *overall saliency* is the key target construct. When ratings of objects are collected in different environments it is particularly interesting to see, whether the impact that different latent variables have on each other is different. It is important to note that in traditional PLS Path Modeling (for a discussion of consistent PLS Path Modeling see [7]) error terms are not included on the latent variable level, i.e. latent variables are treated as composites regardless of the measurement model specification (see [15] for details).

The statistical analysis proposed here comprises two steps: First, the measurement invariance of the measurement model must be assessed. Second, the analysis of observed heterogeneity is performed taking city and sense of direction (the latter also stratified by gender) into account.

4.2 Assessing Measurement Model Invariance in PLS Path Modeling

Following the so-called MICOM-procedure suggested by Henseler et al. [16], measurement model invariance is tested based on three different criteria. Configural invariance is a necessary but insufficient condition for compositional invariance, which can be divided into partial and full measurement invariance, respectively. These three components are explained below.

4.2.1 Configural Invariance

Configural invariance can be achieved only in those cases where the same set of measured variables has been used for all groups and preprocessing steps and settings during the estimation process were identical (see [17, p. 142–143]). These preconditions were met in terms of data collection as the survey used to collect data comprised the same set of German language questions as presented in [19] (see table 4 for a translated version) and the measured variables were used to serve as proxies for the same set of latent variables. Moreover, SmartPLS software [35] was used for all comparisons. The weighting scheme (path), maximum number of iterations (300) and the stop criterion (10^{-10}) were kept equal across group comparisons. The configural invariance is thus given for all comparisons reported in this paper.

4.2.2 Compositional Invariance

Compositional variance can be divided into partial and full measurement invariance, both of which have an immediate effect on the type of comparisons which are feasible. Therefore, compositional invariance will be checked as a first step in each part of the analysis.

4.2.2.1 Partial Measurement Invariance

This criterion deals with latent variable score correlations (see [16] and [17, pp. 143–146]), which are assessed by means of a permutation test. First, the weights are found for each group. Second, latent variable scores are calculated for the whole dataset based on weights of each group separately. Pairwise correlations between the resulting latent variable scores are then established. Confidence intervals for correlations are found by permuting observations across groups and re-assessing the latent variable scores and correlations at least 1 000 times. This procedure provides statistical evidence whether the correlations of scores for the same composites differ significantly from one. Throughout the analysis presented below, 5, 000 permutations were used in all cases.

4.2.2.2 Full Measurement Invariance

If both, configural invariance and partial measurement invariance are given, full measurement model invariance can be achieved. It is given if and only if “the confidence intervals of differences in mean values and logarithms of variances between the construct scores of the first and second group include zero” [16, p. 416]. It is important to note, however, that full measurement invariance will not be discussed throughout this analysis because we focus on structural relationships between the latent variables.

5 Results

We use the results presented in [19] to base our analysis on the structural model depicted in figure 1, including all formative causes for *visual salience* (see table 4).

The results are reported in the following order: We, first, assess differences between the two cities. Based on these results, we, second, analyze structural model differences based on the three subfactors (allocentric orientation, ego-centric orientation, orientation using cardinal directions) proposed in [27]. A third step of the analysis will reveal whether an interaction between gender and sense of direction yields group differences.

■ **Table 1** Outer weights for both cities (standard PLS algorithm). Significant differences ($K = 5000$ permutations) are indicated by bold-faced column headers.

	age	area	intensity	tone	condition	height	length	location	material	motion	pattern	shape	signage	size	width
Augsburg	.073	-.018	.109	.256	.005	.085	-.058	.315	-.225	.094	.157	.267	.142	.017	.300
Regensburg	.240	.090	.266	.116	-.173	.010	-.010	.318	.017	.035	.097	.156	.194	.378	-.161

5.1 Comparing Two Cities

A permutation test revealed that compositional invariance was not given between the two cities: Correlations of cognitive saliency (COS), prototypicality (PRO) and visual saliency (VIS) differed significantly from one. With respect to COS ($cor = .947$, 90%-CI[.985]) the indicator *c_eas* turned out to have particularly adverse properties: Its outer loading in Augsburg is very small ($\lambda_{c2} = .105$). As a consequence, the indicator was removed from the model for the whole analysis, leaving COS as a 2 item construct. Furthermore, a closer look into VIS ($cor = .918$, 90%-CI[.940]) revealed significant differences in outer weights between both cities. While the Regensburg data suggests variable *size* to be most important (see [19]), this causal indicator is rendered insignificant for Augsburg. Table 1 shows the outer weights for both cities based on 5 000 permutations. Given these differences a redundancy analysis [14, p.121–122] was conducted to check whether formative measurement is statistically adequate for the Augsburg dataset. Based on the fact that the path coefficient did not meet the threshold of ($\beta = 0.80$) suggested in [14] we decided to use the reflective indicators to measure visual saliency.

With respect to prototypicality, a very slight ($cor = .996$, 90%-CI[.997]), yet significant difference in correlations from one was found. As no theoretical insights justify the deletion of the construct (see [17] for this kind of advice), we decided to keep this construct but did not take direct or total effects of this latent variable into account. It is, however, reported for completeness reasons. As a consequence analyses reported in the remainder of this paper will be based on the model shown in figure 2.

A reassessment of compositional invariance with *c_eas* being removed and reflectively measured visual saliency establishes partial compositional invariance. Thus, an analysis of structural relationships on pooled data is statistically feasible.

When Regensburg and Augsburg are compared, no significant differences are found for path coefficients nor total effects, i.e. the structural relationships are invariant across different environments of data acquisition. Pooled data from both cities can, thus, be used for the subsequent analyses reported in this paper.

5.2 Sense of direction

The pooled dataset was now used to compare good and poor orientation per one of the factors allocentric, egocentric or cardinal direction. The construct correlations in table 2 indicate that partial compositional invariance was established for all groups and differences in structural relationships can be assessed.

Based on the compositional invariance we uncovered the following significant differences, where groups of spatial abilities were found according to [28]. This means, good and poor groups were found based on raw values by age for the three subscales of sense of direction (allocentric, egocentric, cardinal directions) proposed in [27]. For example, a person aged 35 years having a raw score of 7 or less for factor cardinal direction strategy is assigned to the *poor* group, whereas persons with a raw score greater than 7 are assigned to the *good* group (see [28, pp. 805 and 809]).

■ **Table 2** The mean correlations between bad and good groups based on 5 000 permutations. Neither of these correlations differs significantly from zero (the smallest p-value found across groups and latent variables was $p = .132$), i.e. partial measurement invariance is established between groups and structural relationships can be assessed. Please note: PRO is given for the sake of completeness, only, yet not taken into account (see section 5.1).

	ADV	COS	OVSAL	PRO	STS	VIS
allocentric	1.000	.999	1.000	.999	1.000	1.000
egocentric	1.000	.999	1.000	.999	1.000	1.000
cardinal	1.000	.999	1.000	.999	1.000	1.000

good allocentric vs. poor allocentric orientation The direct effect $ADV \rightarrow STS$ ($\beta_g = .718$, $\beta_b = .769$, 90%-CI = $[-.047; .047]$) differs significantly between both groups, suggesting that poorly allocentric oriented person's rely more on visibility in advance when judging structural salience than good allocentric oriented persons do.

good egocentric vs. poor egocentric orientation Both groups differ with respect to the direct effect visual salience has on overall salience ($\beta_g = .561$, $\beta_b = .643$, 90%-CI = $[-.072; .070]$), i.e. visual aspects turn out to be more important for persons with poor egocentric orientation.

good cardinal vs. poor cardinal The direct effect $VIS \rightarrow OVSAL$ ($\beta_g = .573$, $\beta_b = .655$, 90%-CI = $[-.073; .072]$) differs between both groups as well as $COS \rightarrow OVSAL$ ($\beta_g = .039$, $\beta_b = -.041$, 90%-CI = $[-.060; .059]$) does. These figures, again, indicate that visual aspects are more important to poorly cardinally oriented persons and that cognitive salience might have a negative impact for this group.

Taken together, these results indicate slight yet important differences between these groups. There is, however, evidence in psychology suggesting that gender may be an important factor with respect to orientation preferences (see [6] for a review).

5.3 Sense of Direction Stratified by Gender

We assessed the influence that gender has, first, between and, second, within groups. The between comparison is used to shed light on whether gender is a sufficient explanation for the SoD-related differences found, while the within part examines gender-related differences. The sense of direction groups were, again, found according to [28] (see above, section 5.2). Compositional invariance for both types of comparisons is presented in table 3. It reveals that compositional invariance is not given for several group comparisons across sense of direction factors.

5.3.1 Between sense of direction groups within gender

allocentric In contrast to the other factors, three out of four group comparisons show compositional invariance. Comparing well oriented females to poorly oriented males does not yield significant results and well oriented males do not differ from poorly oriented males. In contrast, well allocentric oriented females differ from poorly oriented females. Visibility in advance has a stronger direct effect on structural salience in the poor group ($ADV \rightarrow STS$ ($\beta_{gf} = .720$, $\beta_{pf} = .821$, 90%-CI = $[-.065; .065]$)); this turns out to be the case for the impact visual salience has on overall salience ($VIS \rightarrow OVSAL$ ($\beta_{gf} = .554$, $\beta_{pf} = .689$, 90%-CI = $[-.104; .135]$)).

■ **Table 3** The construct correlations (4-digits, not rounded), where group comparisons showing compositional invariance are bold-faced. Correlations significantly ($\alpha = .1$ was applied to ensure conservative results) different from 1 are shown in italics. PRO is given for the sake of completeness, yet not taken into account (see section 5.1). The group sizes, i.e. the number of ratings in each group, are given in parentheses once per group for each factor. Level of SoD and gender are denoted as follows: *g-f* means *good oriented females*, *p-m* means *poor oriented males* etc.

Factor	level of SoD and gender	ADV	COS	OVSAL	PRO	STS	VIS
allocentric	<i>g-f</i> (389) vs. <i>g-m</i> (295)	<i>.9998</i>	.9985	.9999	<i>.9991</i>	.9997	<i>.9998</i>
	p-f (200) vs. p-m (309)	.9996	.9987	.9999	.9981	.9997	.9998
	<i>p-f</i> vs. <i>g-m</i>	<i>.9997</i>	.9991	.9999	.9958	.9996	.9998
	g-f vs. p-m	.9997	.9982	.9999	.9994	.9998	.9998
	g-f vs. p-f	.9996	.9971	.9999	.9992	.9997	.9998
	g-m vs. p-m	.9997	.9993	.9999	.9972	.9997	.9998
egocentric	<i>g-f</i> (275) vs. <i>g-m</i> (167)	<i>.9996</i>	.9960	.9999	.9984	<i>.9997</i>	.9997
	p-f (314) vs. p-m (437)	.9997	.9992	.9999	.9990	.9998	.9999
	p-f vs. g-m	.9996	.9977	.9999	.9981	.9996	.9997
	<i>g-f</i> vs. <i>p-m</i>	<i>.9997</i>	.9989	.9999	.9990	.9998	<i>.9998</i>
	<i>g-f</i> vs. <i>p-f</i>	<i>.9997</i>	.9976	.9999	.9993	.9998	<i>.9998</i>
	<i>g-m</i> vs. <i>p-m</i>	<i>.9997</i>	.9990	.9999	.9938	.9996	<i>.9998</i>
cardinal	g-f (247) vs. g-m (225)	.9997	.9983	.9999	.9987	.9996	.9997
	p-f (342) vs. p-m (379)	.9997	.9988	.9999	.9988	.9998	.9998
	<i>p-f</i> vs. <i>g-m</i>	<i>.9996</i>	.9983	.9999	.9987	.9997	.9998
	g-f vs. p-m	.9997	.9987	.9999	.9988	.9998	.9998
	<i>g-f</i> vs. <i>p-f</i>	<i>.9996</i>	.9974	.9999	.9993	.9998	.9998
	<i>g-m</i> vs. <i>p-m</i>	<i>.9997</i>	.9982	.9999	.9987	.9996	.9997

egocentric Given the previous finding regarding the non-stratified egocentric group, the results of the compositional invariance when comparing males and females between both groups was unexpectedly not given for three out of four possible comparisons due to correlational differences in visual saliency. Visual saliency has a higher total effect on overall saliency in poorly egocentric oriented females ($\beta_{gm} = .725$, $\beta_{pf} = .855$, 90%-CI = $[-.078; .079]$). Moreover, the direct ($\beta_{gm} = .064$, $\beta_{pf} = -.069$, 90%-CI = $[-.104; .103]$) and total ($\beta_{gm} = .071$, $\beta_{pf} = -.064$, 90%-CI = $[-.064; .065]$) effects of cognitive saliency on overall saliency are rendered significant. These are, however, in general very low.

cardinal Similar to the findings for the egocentric factor, only one group comparison is feasible out of four when orientation abilities based on cardinal directions are considered. The reason for this, however, is different: It is due to significant correlational differences found for construct visibility in advance. Based on this result, poorly visual saliency has a higher impact on overall saliency for poorly oriented males than is the case of good oriented females ($\beta_{gf} = .506$, $\beta_{pm} = .636$, 90%-CI = $[-.104; .104]$). Vice versa, advance visibility is more important for overall saliency in good cardinally oriented females than poorly oriented males (direct effect $ADV \rightarrow OVSAL$ ($\beta_{gf} = .196$, $\beta_{pm} = .072$, 90%-CI = $[-.120; .117]$) and the total effect $ADV \rightarrow OVSAL$ ($\beta_{gf} = .382$, $\beta_{pm} = .243$, 90%-CI = $[-.099; .096]$).

5.3.2 Within sense of direction groups but across gender

allocentric Poorly allocentric oriented females turn out to differ significantly from the male group with respect to two direct effects: Visibility in advance has a higher impact on structural salience for females ($\beta_{pf} = .821$, $\beta_{pm} = .726$, 90%-CI = $[-.064; .063]$). Furthermore, cognitive salience shows an adverse effect on structural salience in females ($\beta_{pf} = -.010$, $\beta_{pm} = .095$, 90%-CI = $[-.101; .102]$). This effect is very small, though. A group comparison by gender for the good group, however, is not feasible because compositional invariance is not given.

egocentric For poorly egocentric oriented females the direct effects $ADV \rightarrow STS$ ($\beta_{pf} = .766$, $\beta_{pm} = .693$, 90%-CI = $[-.057; .058]$), $COS \rightarrow OVSAL$ ($\beta_{pf} = -.069$, $\beta_{pm} = .009$, 90%-CI = $[-.070; .074]$), $COS \rightarrow STS$ ($\beta_{pf} = .029$, $\beta_{pm} = .147$, 90%-CI = $[-.082; .082]$), $VIS \rightarrow ADV$ ($\beta_{pf} = .639$, $\beta_{pm} = .705$, 90%-CI = $[-.064; .063]$) must be distinguished from poorly egocentric oriented males. These findings indicate that visual salience has a larger impact on advance visibility for males as well as cognitive has on structural salience. Similar to allocentric orientation visibility in advance shows a larger impact on structural salience for females than for males. Similar to allocentric orientation compositional invariance is not given for a good group comparison between gender.

cardinal Females showing a poor orientation based on cardinal directions differ from males with respect to the direct effect $VIS \rightarrow ADV$ ($\beta_{pf} = .645$, $\beta_{pm} = .710$, 90%-CI = $[-.061; .063]$): Visual salience has a higher impact on advance visibility for poorly oriented males than for females and vice versa for well-oriented females as compared to males ($\beta_{gf} = .684$, $\beta_{gm} = .586$, 90%-CI = $[-.096; .097]$).

6 Discussion

Our first goal is to assess measurement invariance; secondly, we are interested in differences between groups of environments and participants. As measurement invariance is a precondition of a heterogeneity analysis, we will discuss both aspects with respect to the different grouping variables.

6.1 Environment

The results suggest that the strength of the relationships (see figure 2) between the sub-dimensions of salience does not differ significantly. The coefficients found can, hence, be used to calculate values for overall salience across different environments. Having found no heterogeneity among different cities is, however, in contrast to those models stressing the importance of the environment (see e.g. [4, 11, 38]). Having said this, one must keep in mind that the data were collected in European cities of Roman descent with a similar layout, although the architectural differences between these two environments are substantial. These differences are reflected in the formative measurement model for visual salience (see section 5.1): In Regensburg the variable *size* has the strongest impact, but is rendered insignificant in Augsburg where *shape* is most important. This finding suggests that the differences between environments are most important at the level of individual formatively measured variables. The structural relationships based on reflective measurements, however, can be used to calculate overall salience scores across different environments and can, consequently, be used in mobile information systems.

6.2 Sense of direction and gender

Although measurement invariance was not established for a number of group comparisons with respect to these factors, we find evidence for the interaction between gender and orientation ability. The effect visual saliency has on overall saliency is particularly affected. The results suggest that a poorer orientation in females yields a larger importance of visual saliency than is the case for good oriented women. This indicates the importance of personal cognitive factors. Individual aspect may also play an important role regarding the impact of cognitive saliency. The coefficients found for cognitive saliency are, although significant, very small. They show, moreover, a sign change in the poor allocentric oriented group, indicating an adverse effect of cognitive on structural saliency in females. One has to keep in mind, though, that random measurement error may have an impact on these results because all but two indicators were removed for this construct, i.e. the lower bound for a suitable number of indicators according to reflective measurement theory is reached (see [21, pp. 178–179]).

We also find a gender-related effect in general. For example, we find evidence that visual cues have a larger impact on overall saliency for females than males – despite their equal level of sense of direction. This finding may be related to the general difference in orientation strategies (see [6]): The preference for egocentric orientation in females may invoke visual cues more. This difference in strategies may also be important to explain the effect visual saliency shows on visibility in advance (larger for females than males in the good cardinal group and vice versa for the poor cardinal group and the poor egocentric group) and visibility in advance has on structural saliency (larger for females in both, the poor egocentric and poor allocentric group). These results are generally in line with those by Picucci et al. [31], who report on gender differences based on spatial confidence and orientation strategies

These findings with respect to sense of direction and gender stress the importance of personal factors in saliency ratings. They reinforce the findings for indoor environments by Lawton et al. [24]: Individual and gender related differences seem to exist in outdoor environments, too. The importance of individual factors is fostered statistically by the generally large number of group comparisons which do not show partial measurement invariance. This statistical property indicates missing variables or constructs within the model which need to be studied in the future.

7 Conclusions and Future Work

The main goals of this paper are to assess invariance with respect to environment, objects and observers of Kattenbeck's measurement model of saliency. Based on this, we analyze the observed heterogeneity, taking environment and sense of direction (stratified by gender) into account. We are, therefore, interested in assessing whether the measurement model may be re-used in different contexts, i.e. whether it provides a robust way of collecting saliency ratings. The results indicate that the structural model is invariant across environment, i.e. the strength of the relationships between the subdimensions of saliency does not differ significantly. The coefficients found can, hence, be used to calculate values for overall saliency across different environments. We, moreover, provide empirical evidence that this is true with respect to both, gender and sense of direction. The degree of influence found for visual dimensions is, generally speaking, in line with what was to be expected: The impact of visual dimensions seems to be different for women and men. Mobile information systems should, thus, take these differences into account, when calculating route instructions. The compositional invariance problems (configural invariance is given for all comparisons reported) occurring throughout the analysis of personal factors can be regarded as an indicator for

the importance of personal factors beyond gender and sense of direction. Taken together, our results indicate that more studies on salience, especially on the impact of personal characteristics, are needed and models have to be adapted so that they can incorporate personal factors.

With respect to future work a next step will be to assess whether the found, often slight, differences have an impact on wayfinding performance in real world scenarios. This will also be examined with respect to the different salience yielded by different models, e.g. by a comparison of wayfinding performance when salience values are based on the Raubal and Winter model [33] vs. the survey-based ratings used in the current study. Furthermore, the need to empirically measure personal preferences has become obvious and will be examined in a future workshop. Thirdly, it will be interesting to learn more about differences in weights subdimensions of salience show on each other and on overall salience, when, e.g. urban and non-urban environments are compared or different languages and/or Non-European urban settings are contrasted.

References

- 1 Donald Appleyard. Why buildings are known: A predictive tool for architects and planners. *Environment and Behavior*, 1(2):131–156, 1969.
- 2 Kenneth A. Bollen. Evaluating Effect, Composite, and Causal Indicators in Structural Equation Models. *Management Information Systems Quarterly*, 35(2):359–372, 2011.
- 3 Gary Burnett. “Turn right at the Traffic Lights”: The Requirement for Landmarks in Vehicle Navigation Systems. *Journal of Navigation*, 53(3):499–510, 2000.
- 4 David Caduff and Sabine Timpf. On the assessment of landmark salience for human navigation. *Cognitive processing*, 9(4):249–267, 2008.
- 5 Edgar Chan, Oliver Baumann, Mark Bellgrove, and Jason Mattingley. From objects to landmarks: The function of visual location information in spatial navigation. *Frontiers in Psychology*, 3:304, 2012.
- 6 Emanuele Coluccia and Giorgia Louse. Gender differences in spatial orientation: A review. *Journal of Environmental Psychology*, 24(3):329–340, 2004.
- 7 Theo K. Dijkstra and Jörg Henseler. Consistent Partial Least Squares Path Modeling. *Management Information Systems Quarterly*, 39(2):297–316, 2015.
- 8 Matt Duckham, Stephan Winter, and Michelle Robinson. Including landmarks in routing instructions. *Journal of location based services*, 4(1):28–52, 2010.
- 9 Birgit Elias. Extracting landmarks with data mining methods. In W Kuhn, M Worboys, and S Timpf, editors, *Spatial Information Theory, Proceedings: Foundations of Geographic Information Science*, volume 2825 of *LNCS*, pages 375–389, 2003.
- 10 Claes G. Fornell and F. L. Bookstein. Two structural equation models: LISREL and PLS applied to consumer exit-voice theory. *Journal of Marketing Research*, 19:440–452, 1982.
- 11 Jana Götze and Johan Boye. Learning landmark salience models from users’ route instructions. *Journal of Location Based Services*, 10(1):47–63, 2016.
- 12 Joseph F. Hair, William C. Black, Barry J. Babin, and Rolph E. Anderson. *Multivariate Data Analysis. A Global Perspective*. Person Education, 7th edition, 2010.
- 13 Joseph F. Hair, Christian M. Ringle, and Marko Sarstedt. PLS-sem: Indeed a silver bullet. *Journal of Marketing Theory and Practice*, 19(2):139–151, 2011.
- 14 Joseph F. Hair Jr., G. Tomas M. Hult, Christian M. Ringle, and Marko Sarstedt. *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*. SAGE Publications, Los Angeles et al., 2014.
- 15 Jörg Henseler. On the convergence of the partial least squares path modeling algorithm. *Computational Statistics*, 25(1):107–120, 2010.

- 16 Jörg Henseler, Christian M. Ringle, and Marko Sarstedt. Testing Measurement Invariance of Composites Using PLS. *International Marketing Review*, 33(3):405–431, 2016.
- 17 F. Joseph F. Hair Jr., Joseph, G. Tomas M. Hult, Christian Ringle, and Marko Sarstedt. *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*. SAGE Publications, Los Angeles et al., 2 edition, 2018.
- 18 Markus Kattenbeck. Empirically measuring saliency of objects for use in pedestrian navigation. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 3:1–3:10, New York, NY, USA, 2015.
- 19 Markus Kattenbeck. *Empirically Measuring Saliency of Objects for Use in Pedestrian Navigation*. Dissertation, Chair for Information Science, University of Regensburg, 2016. URL: <http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:bvb:355-epub-341450>.
- 20 Markus Kattenbeck. How subdimensions of saliency influence each other. comparing models based on empirical data. In E. Clementini, M. Donnelly, M. Yuan, Ch. Kray, P. Fogliaroni, and A. Ballatore, editors, *13th International Conference on Spatial Information Theory (COSIT 2017)*, pages 10:1–10:13. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2017.
- 21 David A. Kenny. *Correlation and Causality*. John Wiley & Sons, 1979.
- 22 Pyry Kettunen, Katja Irvankoski, Christina M. Krause, and L. Tiina. Sarjakoski. Landmarks in nature to support wayfinding: the effects of seasons and experimental methods. *Cognitive Processing*, 14(3):245–253, Aug 2013.
- 23 Alexander Klippel and Stephan Winter. Structural saliency of landmarks for route directions. In Anthony G. Cohn and David M. Mark, editors, *Spatial Information Theory. COSIT 2005. LNCS vol 3693*, pages 347–362. Springer Berlin Heidelberg, 2005.
- 24 Carol A. Lawton, Stephanie I. Charleston, and Amy S. Zieles. Individual- and gender-related differences in indoor wayfinding. *Environment and Behavior*, 28(2):204–219, 1996.
- 25 Kevin Lynch. *The image of the city*. MIT press, 1960.
- 26 Andrew J. May and Tracy Ross. Presence and quality of navigational landmarks: Effect on driver performance and implications for design. *Human Factors*, 48(2):346–361, 2006.
- 27 Stefan Münzer and Christoph Hölscher. Entwicklung und Validierung eines Fragebogens zu räumlichen Strategien. *Diagnostica*, 57(3):111–125, 2011.
- 28 Stefan Münzer and Christoph Hölscher. Standardized norm data for three self-report scales on egocentric and allocentric environmental spatial strategies. *Data in Brief*, 8:803–811, 2016.
- 29 Clemens Nothegger, Stephan Winter, and Martin Raubal. Selection of salient features for route directions. *Spatial Cognition & Computation*, 4(2):113–136, 2004.
- 30 Eva Nuhn and Sabine Timpf. A multidimensional model for selecting personalised landmarks. *Journal of Location Based Services*, 11(3-4):1–28, 2017.
- 31 Luciana Picucci, Alessandro O. Caffò, and Andrea Bosco. Besides navigation accuracy: Gender differences in strategy selection and level of spatial confidence. *Journal of Environmental Psychology*, 31(4):430–438, 2011.
- 32 Teriitutea Quesnot and Stéphane Roche. Quantifying the significance of semantic landmarks in familiar and unfamiliar environments. In Sara Irina Fabrikant, Martin Raubal, Michaela Bertolotto, Clare Davies, Scott Freundsuh, and Scott Bell, editors, *Spatial Information Theory*, volume 9368 of *LNCS*, pages 468–489. Springer, 2015.
- 33 Martin Raubal and Stephan Winter. Enriching wayfinding instructions with local landmarks. In Max J. Egenhofer and David M. Mark, editors, *Geographic Information Science. GIScience 2002. LNCS vol 2478*, pages 243–259. Springer, 2002.
- 34 Kai-Florian Richter and Stephan Winter. *Landmarks. GIScience for Intelligent Services*. Springer International Publishing, 2014.
- 35 Christian M. Ringle, Sven Wende, and Jan-Michael Becker. SmartPLS 3, 2015. Retrieved from <http://www.smartpls.com>.

- 36 Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439, 1976.
- 37 Florian Röser, Antje Krumnack, Kai Hamburger, and Markus Knauff. A four factor model of landmark salience—a new approach. In *Proceedings of the 11th International Conference on Cognitive Modeling (ICCM)*, pages 82–87, 2012.
- 38 Ahmed Sameer and Braj Bhushan. Effect of landmark type on route memory in unfamiliar homogenous environment. *Psychological Studies*, 62(2):152–159, Jun 2017.
- 39 Molly E. Sorrows and Stephen C. Hirtle. The nature of landmarks for real and electronic spaces. In Christian Freksa and David M. Mark, editors, *Spatial Information Theory. Cognitive and Computational Foundations of Geographic Information Science. COSIT 1999. LNCS vol 1661*, pages 37–50. Springer, 1999.
- 40 Lynn A. Streeter and Dan Vitello. A profile of drivers’ map-reading abilities. *Human Factors*, pages 223–239, 1986.
- 41 Rul von Stülpnagel and Julia Frankenstein. Configurational salience of landmarks: an analysis of sketch maps using space syntax. *Cognitive Processing*, 16(1):437–441, Sep 2015.
- 42 Stephan Winter. Route Adaptive Selection of Salient Features. In Werner Kuhn, Michael Worboys, and Sabine Timpf, editors, *Spatial Information Theory. Foundations of Geographic Information Science*, LNCS, pages 349–361. Springer, Berlin / Heidelberg, 2003.
- 43 Kathryn Wochinger and Deborah Boehm-Davis. Navigational preference and driver acceptance of advanced traveler information systems. *Ergonomics and safety of intelligent driver interfaces*, pages 345–362, 1997.
- 44 Herman Ole Andreas Wold. Soft Modelling: The Basic Design and Some Extensions. In K. G. Jöreskog and H. Wold, editors, *Systems under indirect observation. Causality, structure, prediction, part II*, pages 1–54. North-Holland, Amsterdam, 1982.
- 45 Herman Ole Andreas Wold. Partial Least Squares Regression. In S. Kotz and N. L. Johnson, editors, *Encyclopedia of Statistical Sciences*, pages 581–591. John Wiley, New York, 1985.

A Appendix: Variables and Questions

■ **Table 4** Table 4 was taken literally from [18, p. 10]: “A description of constructs (LV) and measured variables (MV) used in this study. Column *ToM* indicates the type of measurement employed for the MV, where *R* denotes *reflective* and *F* means *formative* measurement, respectively. Please note: All questions were translated from German to English.” Please note: We used the German language questions presented in [19] to conduct experiments in Augsburg.

LV	Description	MV	Phrasing	ToM	
Saliency [OVSAL]	“The overall saliency of geographic features is defined as a three-valued vector, whereby the components capture perceptual, cognitive, and contextual aspects of geographic objects” [4, p. 264].	ov_att	To what extent does this object draw your attention?	R	
		ov_por	How suitable is this object to be used as a point of reference?	R	
		ov_mem	How memorable is this object?	R	
Proto- typicality [PRO]	“[...] that is, how typically they represent a category” [39, p. 43]	p_exa	To what extent is this object suitable as an example of objects belonging to the category you named?	R	
		p_img	To what extent does this object represent your impression of such objects?	R	
		p_sim	How often do you encounter similar objects?	R	
Visual Saliency [VIS]	“the features of contrast with surroundings, prominence of spatial location, and visual characteristics that make the landmark particularly memorable” [39, p. 45].	v_loo	To what extent does the appearance of this object draw your attention?	R	
		v_odd	How unusual is the appearance of this object?	R	
		v_eye	How eye-catching is this object?	R	
		v_rec	How recognizable is this object?	R	
			Please find below several visual attributes. For each of these please indicate the extent to which the named visual attribute contributes to an object’s saliency given its surroundings.		
		v_cin	intensity of color	F	
		v_mot	motion (e.g. flashing, flow)	F	
		v_col	tone	F	
		v_loc	location (e.g. raised, very close to street)	F	
		v_siz	size	F	
		v_sha	shape	F	
		v_con	condition (e.g. new, dirty, etc.)	F	
		v_sig	signs attached	F	
		v_hei	height	F	
		v_wid	width	F	
v_len	length	F			
v_are	area	F			
v_pat	pattern	F			
v_mat	material (as far as identifiable)	F			
v_age	To what extent is this object salient as a result of how old it looks?	F			
Structural Saliency [STS]	“Objects are called structurally salient if their location is cognitively or linguistically easy to conceptualize in route directions” [23, p. 347].	s_eas	How easy is it for you to refer to this object in a route description?	R	
		s_lor	How easy is it to describe this object’s location as part of the current route?	R	
		s_imp	To what extent is this object located at an important location within the current route?	R	
		s_dir	To what extent may this object be suitable to determine whether this is the appropriate route or a change in course is required?	R	
Advance Visibility [ADV]	The degree as to which an object at a potential decision point may be seen from the direction it is approached at (cf. [42]).	a_dis	To what extent can one easily refer to this object from afar?	R	
		a_vis	Given the current route, to what extent were you able to see this object from a distance?	R	
		a_per	To what extent is this object generally perceptible from afar?	R	
		a_sui	In the context of the current route to what extent is this object suitable to explain the route?	R	
Cognitive Saliency [COS]	“[...]the processing of information is based on prior knowledge, while intentions and strategies of the observer are in control of the allocation of attention. In our framework, we will use the term Cognitive Saliency to refer to the endogenous factors that influence saliency” [4, p. 255]	c_per	To what extent do you have personal memories concerned with this object?	R	
		c_his	To what extent does this object’s appearance suggest it to be historic?	R	
		c_wor	To what extent do you regard this object to be worthy of preservation?	R	
		c_cus	To what extent is the current use of the object obvious?	R	
		c_pus	To what extent is the former use of the object obvious?	R	
		c_eas	How easy is it for you to label this object?	R	

Labeling Points of Interest in Dynamic Maps using Disk Labels

Filip Krumpe

Department of Computer Science - University of Stuttgart, Germany

filip.krumpe@fmi.uni-stuttgart.de

Abstract

Dynamic maps which support panning, rotating and zooming are available on every smartphone today. To label geographic features on these maps such that the user is presented with a consistent map view even on map interaction is a challenge. We are presenting a map labeling scheme, which allows to label maps at an interactive speed. For any possible map rotation the computed labeling remains free of intersections between labels. It is not required to remove labels from the map view to ensure this. The labeling scheme supports map panning and continuous zooming. During zooming a label appears and disappears only once. When zooming out of the map a label disappears only if it may overlap an equally or more important label in an arbitrary map rotation. This guarantees that more important labels are preferred to less important labels on small scale maps. We are presenting some extensions to the labeling that could be used for more sophisticated labeling features such as area labels turning into point labels at smaller map scales.

The proposed labeling scheme relies on a preprocessing phase. In this phase for each label the map scale where it is removed from the map view is computed. During the phase of map presentation the precomputed label set must only be filtered, what can be done very fast. We are presenting some hints that allow to efficiently compute the labeling in the preprocessing phase. Using these a labeling of about 11 million labels can be computed in less than 20 minutes. We are also presenting a datastructure to efficiently filter the precomputed label set in the interaction phase.

2012 ACM Subject Classification Human-centered computing → Geographic visualization

Keywords and phrases Map labeling, dynamic maps, label consistency, real-time, sorting/searching

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.8

Acknowledgements I want to thank the OpenStreetMap project for the huge amount of geographic data they collected and provide to the public for free use. To have free access to a worldwide, coherent geographic data set was invaluable for this project. Further thank goes to a group of students at the university of Stuttgart for expanding the OpenLayers framework to use the proposed labeling scheme. This work would not have been possible without the open sources of the OpenLayers project. We also want to thank them for their work. Many thanks to the anonymous reviewers for their detailed comments and suggestions for improvements.

1 Introduction

If today someone wants to explore a place anywhere on earth, he uses map services like Google Maps, Here Maps or others. In addition to a classical map these services allow not only to view a static map but to interactively explore the map via zooming and panning of the displayed region. This interactive approach allows to show a rough overview as well as a detailed view of the interesting places on demand. In contrast to former static maps it is not possible to label interactive maps by hand unless zooming is restricted to a set of fixed zoom



© Filip Krumpe;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 8; pp. 8:1–8:14

Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



■ **Figure 1** Germany labeled in 3 different levels of detail at a coarse, medium and high level of detail (© OpenStreetMap contributors).

levels. But even in this case it's a lot of work to develop an appropriate labeling for the map on each of the fixed map scales. So people started working on algorithms to automatically perform the label selection and placement. In figure 1 you see a map of Germany in three distinct levels of detail from coarse (in the top third) to detailed in the bottom third, labeled with the proposed labeling scheme

In order to help the user to track the labeled features while continuously zooming or panning the map some best practices were presented (e.g. [9, 3]). An essential one is that two labels should not overlap each other to ensure readability of the presented map. Of course each label should be placed close to the feature it labels. During map interaction a well-known requirement is that a feature which is labeled on a specific scale should not vanish except for moving out of the viewing range. When it comes to zooming, the label of a less important feature, e.g. a street, should disappear before the label of a more important feature, e.g. the town the street is located in. All of these requirements need to be considered if a map is labeled, may it be by hand or automatically.

In navigation systems the panning and zooming of the map is often done by the system itself. So the most appropriate view is presented to the user, e.g. a car driver. If the driver needs to change roads, a detailed view is presented to him. He is presented with a coarse map view if he travels long distances on a highway. What is special to the navigation system setting is that the presented map often is not north oriented but oriented in the current direction of travel. This leads to a special demand for the map labeling. Imagine you are sitting in your car focused on traversing a road junction. When you looked at your

navigation system most recently you saw some nearby towns labeled with their names. After having passed the junction you look at the presented map again that has been rotated in the meantime. Unfortunately you are now presented with a completely different set of labels. This is because some of the former labels would intersect in the current orientation. Some labels, which were overlapping before, can now be presented without overlapping. These modifications in the presented labeling make it difficult for you to regain orientation again.

The scenario shows that a labeling in case of rotating maps especially in navigation systems need to fulfill an additional requirement. A label that is presented to the user at a specific rotation angle should not overlap with another label in any rotation angle. Of course the requirements we described above for interactive maps also need to be fulfilled in this particular use case.

In this paper we will present a new labeling scheme, which allows to compute labelings for arbitrary map scales at interactive speed. The consistency requirements, our labeling scheme guarantees to be fulfilled, are described in section 2. The scheme is based on a preprocessing phase and a filtering step during run-time. It is described below in section 3. In section 4 we describe some extensions to the model, which allow to add more sophisticated labeling features. We implemented the approach and extended the well-known OpenLayers web based map visualization framework. Some details and practical considerations are presented in section 5. Section 6 concludes the paper and sketches some further research topics.

Let us shortly create a common understanding of the basic terms we are using in the context of map interaction before describing some related work. A map may show a specific geographic region, for example Germany, Europe or the whole planet. If the map shows only the most important details, we are talking about a coarse or low level of detail. In contrast a fine grained map shows a lot of smaller features, i.e. a high level of detail (see figure 1 for an example). In practice the level of detail is interrelated to the map scale of the map visualization. So a map of a high scale (e.g. 1:1,000) shows a higher level of detail than a map of small scale, e.g. 1:1,000,000). Given an interactive map at a small scale, showing Germany for example, we may zoom into the map, i.e. increase the map scale while shrinking the view area, to get a more detailed view of our capital Berlin.

1.1 Related Work

Some of the first and well-known approaches to systematically describe general principles of static map labeling were done by Eduard Imhof in the 1980s [9]. He distinguishes three general types of features that can be labeled on a map: area, line and point features. For these features he describes best practices for the placement of associated labels to gain best visibility and readability of a map. His observations are the basis of the work at hand.

In the 1990s the problem to automatically label dynamic maps at an interactive speed arose. Kreveld et al. in 1997 published some models about how to select a suitable subset of settlements for interactive display [10]. In particular their approach of computing a ranking of the settlements that allows to efficiently obtain a solution using filtering can be found in the labeling scheme we are presenting here.

In 2006 Been et al. proposed some general consistency desiderata for dynamic map labelings [3]. Those criteria can be considered almost standard in the field of automatic labeling of dynamic maps. They also describe a framework to automatically derive map labelings fulfilling these criteria at interactive speed for map interactions like panning and zooming. Their approach can be considered a direct parent of the approach we are describing in the paper at hand. In the cited paper Been et al. also provide a nice overview of the research on this topic before 2006.

While Been at al. only consider zooming and panning interaction, Gemsa et al. [7, 8] are considering map rotation. Starting with a given map labeling they defined a model to efficiently derive a subset of labels that can be visible at a specific rotation angle without overlapping each other. Their approach covers map rotation only in particular they do not target the problem of presenting a consistent labeling at several levels of details.

From the algorithmic side there is some previous works of our working group targeting the efficient computation of the so called elimination sequences of growing disks [5, 2, 6]. Given a set of points $p_i \in \mathbb{R}^2$ (or \mathbb{R}^d) with associated radii r_i . Each of the points induces a disk (or a hyperball) with radius $r_i \cdot t$. Starting at $t = 0$, t increases continuously and the less prioritized disk is eliminated if two disks touch. The goal is to efficiently compute the elimination sequence and the elimination times for each of the disks. In these two papers a total order for the growing disks is assumed but the results (algorithms and complexity analysis) nicely transfer to other priority functions which only have constant evaluation time. Ahn et al. in 2017 introduced another algorithmic approach that further decreases the computational complexity in [1]. If no order on the disks is given, the problem of computing optimal elimination sequences is NP-hard [3]. Some used mixed-integer programming to compute optimal solutions for these cases [4, 15]. But because of the computational complexity of the problem, in practice they were able to compute solutions only for instances of a few hundred points.

1.2 Contribution

We propose a labeling scheme for dynamic maps, which allows panning, rotating and zooming interaction. It provides consistent labelings at an interactive speed using a computationally challenging preprocessing phase that allows to use efficient and fast filtering techniques during the interaction phase. The scheme is based on the described works of Kreveld et al. and Been at al. . We go beyond the work of Kreveld et al. by additionally taking into account some consistency requirements during zooming operations. Those criteria are inspired by the ones proposed by Been at al. but are exceeding their work by respecting some sort of hierarchy of the geographic objects. In contrast to the work of Been at al. we are also considering rotation as a possible map interaction.

Our results can be extended to some more sophisticated labeling scenarios like area or line labels that turn to point labels while zooming out of the map. This idea is based on an observation of Imhof in [9], namely that area and line labels turn into point labels on smaller map scales.

The labeling scheme incorporates a preprocessing step and allows to use simple filtering and spatial search during the interaction phase to compute a consistent labeling at an adequate speed. The preprocessing of the data can be done efficiently by a proposed algorithm with a running time of $\mathcal{O}(\Delta^2 n (\log n + \Delta^2))$ where Δ is the maximum ratio between two distinct label radii (see [2]). To get an appropriate subset of labels out of the precomputed label ranking we use a priority search tree in combination with a 2-dimensional kd-tree. The implementation is open source and can be found on GitHub [11].

In an associated student project, we extended the well-known OpenLayers web framework [13] to present the capability of our new approach in practice. The so called Tile Rotating Universal Map Projection Presentation is open source and available online [14].

2 Consistency requirements

Based on the consistency desiderata defined by Been et al. in [3] we define the following requirements for interactive maps which allow panning, rotating and zooming of the map view:

- (D1) *During monotonous zooming labels should not appear and disappear more than once.* This requirement covers the user expectation that during zooming an object is visible until it is no longer important enough (when zooming out). When zooming in it ensures that a label can vanish if the labeled object covers the whole view or the labeled point for example got replaced by an area or line feature label at larger scales.
- (D2) *Labels should not change position or size abruptly on map interaction.* This is because abrupt label changes during map interaction may distract the user and make it difficult to track the position of the labeled objects.
- (D3) *During panning and rotation labels are not allowed to appear or disappear except for moving in and out of the view area.* This implies that labels might be partly visible if the label disk is not fully contained in the view area. Especially in navigation systems the map rotation changes automatically, for example when it is linked to a driving direction. The requirement ensures that always the same set of labels is visible even if the map rotation changed since the last time the user looked at the map. This allows the user to keep orientation with a low cognitive load.
- (D4) *The label placement and selection is a function of scale and the view area. It does not depend on the interaction history.* Given this requirement, the map at a specific map setting looks like a static map labeling. So users might directly recognize the places if they look at the map.

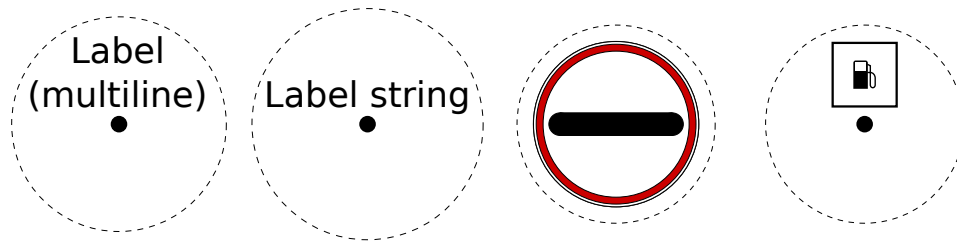
Based on the work of Kreveld et al. [10] we add another constraint targeting the fact that there is some inherent order of precedence for geographic features to be labeled.

- (D5) *During zooming a label disappears only if it is in conflict with an equally or more important label.* For example a megacity label is preferred to a label of a small rural settlement. So if those two labels are in conflict, the megacity label should be shown instead of the settlement label. Also on small map scales a street name is less important than the label of the city in which the street is located. During the label selection process these precedences need to be taken into account. Our label selection process additionally respects what Kreveld et al. called the *relative importance* of an object. For example a town might have a high relative importance if it is in the middle of nowhere compared to a city that is located directly beside a megacity with millions of inhabitants. The relative importance manifests itself in the fact that the label of the town is shown longer than the city label while zooming out of the map.

Now that we have defined the consistency requirements we want our labeling to fulfill, we will continue describing our model in the following section.

3 The Framework

Formally we are considering a set of point of interest locations $P = \{p_1, \dots, p_n\}$. For each point p_i we are considering its label l_i that may be a label string, an icon or the like. We also assume to have a priority function that decides for two points p and q if p is prioritized over q or vice versa or none of both. The problem we are faced with is the following: Given a



■ **Figure 2** Example of point of interest labels with label strings (left) centered above the labeled feature and icons centered at the labeled feature and centered above the labeled feature (right - [16]). The associated label disks are depicted by the surrounding dashed circle.

specific view area, scale and rotation angle select a subset of points such that a visualization of the corresponding labels fulfills the requirements as defined in section 2. In the following we will describe a labeling model at first and a label selection process that allows to efficiently retrieve such a point set. In our case efficiently means that the label selection process can be subdivided into two phases. In a first phase the label set is preprocessed such that in a second interaction phase the actual label selection reduces to a simple and fast filtering. This allows to efficiently query the data set for a consistent labeling during the interactive visualization phase.

3.1 The label model

In order to fulfill the consistency criteria as defined in section 2 we define a label disk for each of the points to be labeled. The label disk is centered at the corresponding point location and has a specific radius r depending on the label size, i.e. the label length, the font and font size or the icon and icon size. We require a point label to be completely contained within the corresponding label disk in each rotation angle but we do not care about its actual placement. In order to fulfill requirement **(D2)** and **(D4)** the label placement must be a function of scale and rotation angle. It must ensure that the label does not change its position and size abruptly during map interaction. Except for these restrictions, the concrete label placement within the label disk is unconstrained. A fairly simple example for such a placement function, which fits the idea of the model well, is depicted in figure 2. There you see a point label that is horizontally aligned and centered above the labeled feature. During rotation the label remains horizontally aligned and keeps its absolute size on zooming. Icons can be placed e.g. centered at their location or centered above the location like in case of the text label as described before. Of course many other placements are also possible.

Using these label disks, we define a consistent labeling to be a subset of the labels such that the corresponding label disks are non-overlapping. Because each label is completely contained within its corresponding label disk by definition, this ensures that none of the labels are overlapping in any rotation of the map view. So we ensured that during rotation none of the labels need to disappear to avoid label overlap. In figure 3 you can see a visualization of Germany labeled with our scheme in two different rotation angles.

To ensure consistency during panning we come back to a concept Been et al. called an “inverted sequence” in their approach in [3]. The intuitive label selection and placement method is to first select the subset of labels in the view area and placing the corresponding labels afterwards. As Been et al. argued in their paper it is hard to achieve interactive speed and consistency with this approach. What we suggest here is to first pick a consistent labeling globally. From this restricted label set we finally display the labels intersecting our



■ **Figure 3** A labeling of Germany in two different orientations (© OpenStreetMap contributors). The basic label set contains all human settlements extracted from the OpenStreetMap dataset [12].

view area. This selection process ensures that the only way labels appear or disappear on panning is by moving in and out of the view area.

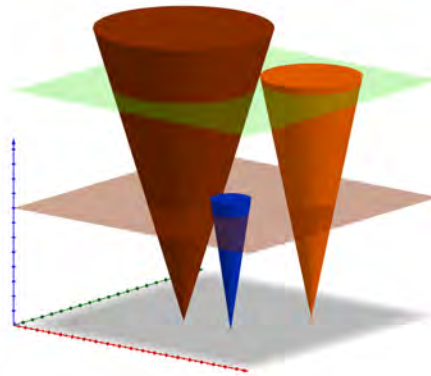
In summary until now our label model requirement **(D3)** is fulfilled, i.e. labels only appear and disappear by moving in and out of the view area. The requirements **(D2)** and (partially) **(D4)** are fulfilled by an appropriate label placement function. For the latter we did not yet define the dependency to the map scale but we are going to make up for it right now.

The map interaction we haven't considered yet is zooming. Zooming out of the map by decreasing the map scale naturally leads to decreasing the level of detail of the map, i.e. less details get visible and labeled on the map. To support this we define the label disk radii to be dependent on the map scale. Instead of the label radius r_i we define the disk radius of p_i to be $r_i \cdot \frac{1}{s}$ where s is the current map scale. You see that decreasing the map scale s enlarges the label disks so a consistent labeling contains less labels – the level of detail decreases. By using these scale dependent label disks, the selection of a consistent labeling gets a function of scale as required in **(D4)**.

The defined label model now allows to have a consistent map view for map interactions at arbitrary map scales. What we have not yet taken into account are the consistency criteria concerning the zooming process itself. As defined in the consistency requirements **(D1)** and **(D5)** for the zooming we have some requirements telling us that labels should not appear and disappear more than once during monotonous zooming. Additionally we require a label to be removed from the view only if it conflicts with a more or equally prioritized label. We will target this in the following section.

3.2 The label selection

In the previous section we defined a labeling model that ensured some requirements to be fulfilled when panning and rotating on an arbitrary map scale. We now want to focus on the process of selecting consistent labelings such that the remaining requirements are



■ **Figure 4** Label cones and two planes (brown, light green) that correspond to label selections at different map scales.

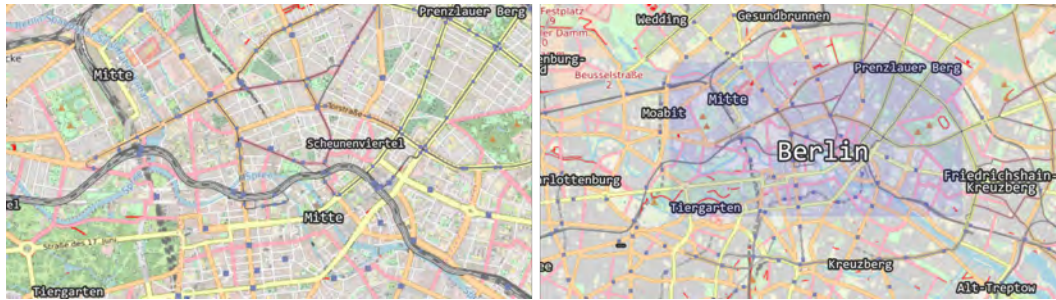
fulfilled. These are concerning the process of zooming in and out of the map, i.e. increasing or decreasing the map scale. There are two requirements left: **(D1)**: 'During monotonous zooming a label should not appear and disappear more than once' and **(D5)**: 'During zooming a label disappears only if it is occluded by an equally or more important point label'

For the sake of simplicity we only consider the process of zooming out of the map, i.e. decreasing the map scale. It is straightforward to transfer the following observations to the process of zooming in. Furthermore we will look at the label disks only and not rely on the actual label placement but assume that this is done in a suitable way as described in the previous section.

In order to find a proper consistent labeling for a target map scale S we use the following process: Starting with a sufficient large s we know that all the corresponding label disks are free of intersections. We continuously decrease s until two of the label disks touch. Now the priority function comes into play. If one of the corresponding points is prioritized over the other, we remove the less prioritized one. Otherwise we remove one of the two. We continue with the process until $s = S$, i.e. our target scale is reached. In the case that we are free to decide which of two equally important point labels to remove, the decision we make influences the further process and so the quality of the labeling. We will come back to this point in the conclusion and further research section.

The process immediately ensures requirement **(D5)** to be fulfilled as a label is removed only if its label disk is in conflict with a label disk of an equally or more important label. Requirement **(D1)** is also fulfilled by design of the label selection as a label never reenters the process after being removed once.

As you can see the label selection process always leads to the same label "elimination sequence" if we assume the label set to be unchanging and the breaking of the ties to happen deterministically. Each label can be assigned a specific map scale where it is removed from the label set during the process. This opens space for our promised precomputation phase. Because the label elimination sequence and the elimination scales for the labels do not change, we can compute them separately in advance. Having computed them for a set of labels we can derive a consistent labeling as follows. For a given map scale S we choose the subset of labels having an associated elimination scale smaller than S and restrict the subset to those labels intersecting the view area. This allows to retrieve a consistent labeling at an interactive speed.



■ **Figure 5** Labeling of a map using a map labeling with popup scales at a larger (left) and smaller map scale (right) where the “Berlin” label popped up (© OpenStreetMap contributors).

Looking at the process from a more abstract point of view we are presented with the following so called space-scale cube: The labels are located in a 2-dimensional plane for example when using the Mercator projection. Using $\frac{1}{s}$ as a third dimension, we see that each label is associated to a cone (see figure 4). The elimination scale of a label determines the height of the cone and the label cones do not intersect. In this view a labeling of a map on a specific scale S corresponds to the intersection of the label cones with the plane at the height $oh \frac{1}{s}$. In the referenced drawing you see two planes corresponding to two different map scales in brown and bright green. The intersection of a cone with the plane corresponds to the label disks of the label at the specific map scale.

4 Extending the model

The labeling model we developed in the previous section opens up opportunities to some extensions. In the following we will discuss some of these.

What we did not take into account yet is the point where a label occurs while zooming out of the map. In the basic labeling model we described before, all the labels are visible at the largest map scale. A straightforward approach to extend the labeling model is to introduce a “popup scale” for a label. It means that the label becomes visible at this specific map scale. At larger map scales the label does not exist and also does not occlude any of the existing labels. This concept for example allows to add a label of a city on coarser map views only such that the label does not occlude details of the city while being in a zoomed in map view. This extension does not violate the requirement (D1) as the label only appears once during a monotonous zooming operation. An example of a labeling of Berlin is depicted in figure 5.

Having in mind the concept of popup scales as described above, we introduce another extension. As Imhof pointed out in [9] line or area feature labels turn into point labels on smaller map scales. For example a church might be displayed as an area on larger map scales but on a coarser map view its area degenerates to a single point. Analogously the label needs to turn from an area label to a simple point feature label. In figure 6 you see a map in two different map scales where the area of Berlin is labeled as an area on the left hand side while it is labeled as a point object on a smaller map scale (right). The same can be applied to line segment labels. By using the concept of popup times we are able to include this fact into our point labeling scheme. Simply setting the popup time for the point label to the map scale where the area turns into a point in the visualization allows us to support this visualization feature.

Obviously the labeling scheme is not focused on maximizing the number of labels visible but on visibility and readability of the map with low cognitive load. For example for simple



■ **Figure 6** Labeling of an area using an area label (left) and a point label at smaller map scale (right) (© OpenStreetMap contributors).

horizontally aligned labels a lot of the available disk space remains unused. To further increase the number of labeled objects one can think of a multilayer approach as follows. Each label which is removed while zooming out is moved to a second label layer, which is overlaid by the main labeling layer. In this second layer labels might use another font type, color or opacity to make clear it is a background layer. Technically the second layer uses the same labeling model, i.e. none of the disks of labels in the second layer overlap. For each label the elimination scale in the first layer is the popup scale in the second layer. So for each of the layers the consistency requirements are fulfilled but the labels of the different layers might overlap in the visualization. In figure 7 you see an example of a labeling with two layers.

Now that we described the labeling scheme from theoretical point of view and discussed some extensions to the model, we will briefly have a look at the practical details of our work.

5 Computing labelings in practice

As described in the previous section the labeling computation subdivides into two phases. In a preprocessing phase an elimination order or label ranking is computed. The general process and some optimizations to do the precomputation more efficiently are discussed in the following section 5.1. When it comes to the visualization and interaction phase the precomputed elimination order needs to be filtered efficiently. Section 5.2 describes an approach to do this by using a combination of a priority search tree and a 2-dimensional kd-tree. To finally visualize the label set we extended the OpenLayers visualization framework. The changes we made are described in section 5.3.

5.1 Preprocessing

A basic algorithm to compute the elimination order is given in the following bottom up approach. For each label p_i we compute the next upcoming label disk collision, i.e. the collision at the largest scale s_i . For the collision at the largest scale we decide which label to remove and start over again with the shrunked label set until only one label is left. This basic algorithm has a computation complexity of $\mathcal{O}(n^3)$ as we need for each label to check the remaining labels for conflicts in a total of $n - 1$ iterations. The crucial observation is that



■ **Figure 7** Labeling of Stuttgart that uses a second label layer (filled gray font) to increase the number of labeled points. In this particular rotation the labels “Stuttgart” and “Stuttgart-Ost” in the first layer are occluding the label of “Umlandshöhe“ in the second layer (© OpenStreetMap contributors).

a label may collide with the most far away label if the disk radii are suitable chosen. This observation forces us to really check each other label when checking for the next collision.

In the following we describe some observations which allow to speed up the algorithm. This is a brief sketch of two of our previous results published in [5] and [2].

A first observation allowing us to reduce the computational complexity is that in each iteration we only need to enforce that the globally next collision needs to be correct. All other computed next collisions might not be correct but the overall sequence of eliminations nevertheless is computed correctly. We see that a collision may be computed twice from each of the two colliding labels. In fact it is sufficient if only the label with the larger radius correctly computes a collision. A second observation guiding to a more efficient algorithm affects two subsequent iterations. Consider an iteration at scale s and a label disk which cover less than half the distance to the nearest neighbor of the associated label. We can argue that we do not need to search for the next collision of this labels until its label disk covers half the distance to the corresponding nearest neighbor. In the meantime some labels might be removed from the label set shrinking the set of labels we need to check for possible conflicts. These two observations allow us to conclude the following: When searching for a next collision of a label we only have to consider a subset of the labels. The maximum size of this particular subset depends on $\Delta = \frac{r_{max}}{r_{min}}$, i.e. the ratio between the largest and the smallest of all radii r in the instance.

Furthermore we observe that many predicted collisions remain the same in two subsequent iterations. So we only need to compute them once. We proved that it is sufficient to maintain the computed collisions in a priority queue. When removing a label we only need to recompute the next collision for a subset of direct neighbors of this label.

The described improvements heavily depend on efficient implementations of two spatial operations: “Finding the nearest neighbor in a point set for a query point” and “Finding all

■ **Table 1** Peak memory consumption (space) and execution times (time) for the precomputation of real-world data sets from the OpenStreetMap project [12].

Dataset	#items [10^3]	time [mm:ss]	space [MB]
Germany	1,308	2 : 05	1,715
Europe	6,468	12 : 23	7,947
Planet	11,006	19 : 33	13,852

points within a distance of d around a query point”.

We implemented the preprocessing algorithm with its various optimizations for point sets with a total order. This allows us to decide the priority function in constant time. In our implementation we used Delaunay triangulations to implement the spatial queries. Our implementation allowed us to compute label rankings for millions of labels in a reasonable time (see table 1). These fast computation times enabled us to recompute labelings for quickly changing point sets, which for example are containing live traffic information.

5.2 Label Selection

In the interaction phase we are provided with a set of labels and the corresponding elimination scales, i.e. a location, a scale value and the label information. For a given query consisting of a range given in maximum and minimum latitude and longitude coordinate and a map scale we need to report all label points located in the region with an elimination scale smaller than the requested scale. To efficiently answer such queries we use an approach that combines a priority search tree with a 2-dimensional kd-tree. The data structure is a 2-dimensional search tree of the label positions. The tree fulfills the min heap property on the associated elimination map scales.

The root node of the tree contains the label with minimum elimination scale. The remaining labels are split into two equally sized subsets according to their *longitude* coordinate. For each of the subsets a subtree is constructed rooted at the label with the minimum elimination scale and the remaining labels are split according to their *latitude* coordinate. The left subtree contains all the labels with smaller coordinate and the right subtree all the labels with larger coordinate. In the third layer the labels are again split according to their *longitude* coordinate and so forth. To each of the tree nodes we also append the value of the coordinate value, which separates the labels in the two subtrees.

Now at query time we traverse the data structure starting at the root node. If the current node has an elimination scale which is less or equal to the requested map scale we need to further explore the subtree. If the current node is contained within the query rectangle the label is to be reported. We need to further explore the left subtree only if the split value of the node is larger than the corresponding minimum value of the requested range. If the maximum of the corresponding query dimension is larger than the split value, we need to explore the right subtree.

5.3 Visualization

To evaluate the labeling result and to provide them to the public there was a student project associated with the research project. An extension to the well-known OpenLayers framework for web based map visualization [13] was developed. Screenshots of the testing instance can be seen in the figures 1 and 3. We created a REST based web service that allows to query a

precomputed label set for the active labels in the given display setting, i.e. displayed map range and the current map scale. On the client side the modified OpenLayers framework uses map tiles without point of interest labels and adds a layer displaying our label set on top of it. The labels are placed centered at the label position and remain horizontally aligned when the map is rotated by the user. One of the harder parts in implementing the framework extensions was to present the continuous zooming capability of our approach. Therefore we needed to modify some parts of the original framework. First we needed to interpolate the map scale between the fix zoom levels, which are provided by the framework itself, in order to visualize the correct subset of labels. Second we had to adopt the framework such that during the zooming between the levels a refresh of the label layer was triggered.

We also implemented a caching mechanism for the labeling data that allowed to reduce the number of required server requests. This mechanism shows an important property of our labeling scheme. When requesting the labeling for a specific map setting, we in fact enlarge the requested query range and increase the requested map scale. The provided labeling now contains additional labels that do not need to be displayed in the current setting. But the required labeling is a subset of the provided one and we can filter out the additional labels by simply skipping elements with a larger elimination scale or those which are not contained within the displayed region. The crucial point is that provided label set allows us to directly handle small zooming or panning interactions locally without the need to immediately request new data from the server.

6 Conclusion

We introduced a new model for labeling interactive maps, which allows panning, zooming as well as rotating the map. The labeling ensures some consistency criteria to be fulfilled. During continuous zooming a label appears and disappears at most once. If a label disappears while zooming out, it is because there exists a map rotation where it overlaps a label of equal or higher importance. When panning or rotating a map a label only appears or disappears if moving out of the view area.

The labeling model depends on a precomputation phase that allows to efficiently derive labelings for arbitrary map scales in an interaction phase using filtering. The precomputation can be done in less than 20 minutes for labelings with around 10 million labels on a standard desktop computer if a total order on the labels is given. This allows to quickly recompute labelings for example to add traffic information labels.

In an interactive visualization phase the labeling computation is a filtering step only. An algorithm to efficiently do that even for large datasets was introduced. The resulting label set has some nice property namely that a labeling for smaller map scales is a subset of the current label set. So labelings for smaller map scales can be derived by filtering the current label set. This idea was also sketched in the paper at hand.

Further research should target the computation of labelings for point sets which do not have a total order but some kind of hierarchy levels the points of interest are belonging to. A top level might for example contain all the megacities. The next levels of decreasing priority might contain other city, town, street and point of interest labels and so forth. In such instances heuristics might be used to solve conflicts between labels of the same hierarchy level. The outcome of such heuristics could be compared with optimal results, e.g. computed by a linear program solver.

Regarding the quality of the computed labeling it is of interest to compare the computed labeling with a maximum subset of labels whose associated disks are free of intersections at

this particular scale. This would be suitable to show the influence of the zooming consistency requirement to the quality of the labeling, i.e. the number of labels.

Further development of our visualization might contain the labeling of area and line features that turn into point labels on smaller map scales. An initial approach how to do that was sketched in the section about extensions to the label model.


References

- 1 Hee-Kap Ahn, Sang Won Bae, Jongmin Choi, Matias Korman, Wolfgang Mulzer, Eunjin Oh, Ji-won Park, André van Renssen, and Antoine Vigneron. Faster Algorithms for Growing Prioritized Disks and Rectangles. *arXiv:1704.07580 [cs]*, 2017. arXiv: 1704.07580. URL: <http://arxiv.org/abs/1704.07580>.
- 2 D. Bahrtdt, M. Becher, S. Funke, F. Krumpke, A. Nusser, M. Seybold, and S. Storandt. Growing balls in \mathbb{R}^d . In *2017 Proceedings of the Nineteenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, Proceedings, pages 247–258. Society for Industrial and Applied Mathematics, 2017. DOI: 10.1137/1.9781611974768.20. URL: <http://epubs.siam.org/doi/abs/10.1137/1.9781611974768.20>.
- 3 K. Been, E. Daiches, and C. Yap. Dynamic Map Labeling. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):773–780, 2006. doi:10.1109/TVCG.2006.136.
- 4 Ken Been, Martin Nöllenburg, Sheung-Hung Poon, and Alexander Wolff. Optimizing active ranges for consistent dynamic map labeling. *Computational Geometry*, 43(3):312–328, apr 2010. doi:10.1016/j.comgeo.2009.03.006.
- 5 Stefan Funke, Filip Krumpke, and Sabine Storandt. Crushing Disks Efficiently. In *Combinatorial Algorithms*, Lecture Notes in Computer Science, pages 43–54. Springer, Cham, 2016. doi:10.1007/978-3-319-44543-4_4.
- 6 Stefan Funke and Sabine Storandt. Parametrized runtimes for ball tournaments. In *Proc. 33rd European Workshop Comput. Geom. (EWCG)*, pages 221–224, 2017.
- 7 Andreas Gemsa, Martin Nöllenburg, and Ignaz Rutter. Consistent Labeling of Rotating Maps. *arXiv:1104.5634 [cs]*, apr 2011. arXiv: 1104.5634. URL: <http://arxiv.org/abs/1104.5634>.
- 8 Andreas Gemsa, Martin Nöllenburg, and Ignaz Rutter. Evaluation of Labeling Strategies for Rotating Maps. *J. Exp. Algorithmics*, 21:1.4:1–1.4:21, apr 2016. doi:10.1145/2851493.
- 9 Eduard Imhof. Positioning Names on Maps. *The American Cartographer*, 2(2):128–144, jan 1975. doi:10.1559/152304075784313304.
- 10 Marc Van Kreveld, Rene Van Oostrum, and Jack Snoeyink. Efficient settlement selection for interactive display. In *In Proc. Auto-Carto 13: ACSM/ASPRS Annual Convention Technical Papers*, pages 287–296, 1997.
- 11 Filip Krumpke. Runtime datastructure project, 2017. [Online; accessed 10-February-2018]. URL: https://github.com/krumpefp/runtime_datastructure.
- 12 OpenStreetMap. Openstreetmap project, 2017. [Online; accessed 4-February-2018]. URL: <https://www.openstreetmap.org>.
- 13 OpenLayers Project. Openlayers project page, 2017. [Online; accessed 10-February-2018]. URL: <http://openlayers.org/>.
- 14 TRUMP Project. Tile rotating universal map projection project, 2017. [Online; accessed 10-February-2018]. URL: <https://trump-fmi.github.io/#/>.
- 15 Nadine Schwartges, Dennis Allerkamp, Jan-Henrik Haunert, and Alexander Wolff. Optimizing Active Ranges for Point Selection in Dynamic Maps. In *Proceedings of the 16th ICA Generalisation Workshop (ICA'13)*, 2013.
- 16 OpenStreetMap Wiki. Map icons/proposed icons — openstreetmap wiki, 2017. [Online; accessed 4-February-2018]. URL: http://wiki.openstreetmap.org/w/index.php?title=Map_Icons/Proposed_Icons&oldid=1463667.

Improving Discovery of Open Civic Data

Sara Lafia

Department of Geography, University of California, Santa Barbara, USA
slafia@geog.ucsb.edu


 <https://orcid.org/0000-0002-5896-7295>

Andrew Turner

Esri DC, Office of Research and Development, Arlington, VA, USA
ATurner@esri.com

Werner Kuhn

Department of Geography, University of California, Santa Barbara, USA
werner@ucsb.edu

 <https://orcid.org/0000-0002-4491-0132>

Abstract

We describe a method and system design for improved data discovery in an integrated network of open geospatial data that supports collaborative policy development between governments and local constituents. Metadata about civic data (such as thematic categories, user-generated tags, geo-references, or attribute schemata) primarily rely on technical vocabularies that reflect scientific or organizational hierarchies. By contrast, public consumers of data often search for information using colloquial terminology that does not align with official metadata vocabularies. For example, citizens searching for data about bicycle collisions in an area are unlikely to use the search terms with which organizations like Departments of Transportation describe relevant data. Users may also search with broad terms, such as “traffic safety”, and will then not discover data tagged with narrower official terms, such as “vehicular crash”. This mismatch raises the question of how to bridge the users’ ways of talking and searching with the language of technical metadata. In similar situations, it has been beneficial to augment official metadata with semantic annotations that expand the discoverability and relevance recommendations of data, supporting more inclusive access. Adopting this strategy, we develop a method for automated semantic annotation, which aggregates similar thematic and geographic information. A novelty of our approach is the development and application of a crosscutting base vocabulary that supports the description of geospatial themes. The resulting annotation method is integrated into a novel open access collaboration platform (Esri’s ArcGIS Hub) that supports public dissemination of civic data and is in use by thousands of government agencies. Our semantic annotation method improves data discovery for users across organizational repositories and has the potential to facilitate the coordination of community and organizational work, improving the transparency and efficacy of government policies.

2012 ACM Subject Classification Information systems → Digital libraries and archives

Keywords and phrases data discovery, metadata, query expansion, interoperability

Digital Object Identifier 10.4230/LIPICs.GIScience.2018.9

Supplement Material <https://github.com/saralafia/esri-hub>

Acknowledgements The work presented in this paper was part of a research internship of the first author at Esri’s Research and Development Office. Additional contributions by Pranav Kulkarni, Daniel Fenton, and Alexander Harris of Esri Research and Development are gratefully acknowledged. The work was supported by Esri and by UCSB’s Center for Spatial Studies.



© Sara Lafia, Andrew Turner, and Werner Kuhn;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 9; pp. 9:1–9:15

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

In recent decades, great strides have been made to encourage data creators and providers to make the findings of their research or results of their activities publicly accessible. Researchers receiving grant funding now face mandates to preserve and expose data resulting from their research [9]. Parallels can be drawn between the mounting movement surrounding open access in academia and similar movements well underway in the civic arena surrounding shared municipal data; all levels of government, from Federal agencies to city governments, have started exposing data [14]. Open data, also known as open Public Sector Information, contribute to citizens' rights to public access of government information. Open data policies at various levels of government have stimulated and guided the publication of both spatial and non-spatial government data [15]. The resulting creative downstream use of civic datasets is staggering, ranging from mobilization of grassroots citizen initiatives to uptake by private application developers [7]. By making civic data about a range of topics, from departmental budgets to bicycle collisions, consumable through APIs, governments such as the City of Los Angeles¹ have become better connected to their citizenry.

However, simply making data accessible online does not guarantee their discoverability [1]. The likelihood of discovering thematically relevant geospatial data is still quite low; this is due to two key geospatial issues. The first issue is that data produced by co-located and adjacent governments are often described differently. Thus, discovering spatial data about bicycle collisions provided by neighboring governments, such as Arlington and Fairfax, VA along with state data, for example, is not trivial. This is because data, such as bicycle collision statistics, are described in a heterogeneous way by neighboring municipalities and by various levels of government. A second issue is that civic data are not described using terms that public consumers use. Governments may collect and provide traffic collision statistics, while consumers may want to assess community safety for cyclists.

It is unrealistic to imagine all providers of civic data conforming to a single metadata standard or providing suites of additional colloquial keywords to resolve these issues. In fact, the multiplicity of inward-looking open data policies at various levels of government make this untenable [15]. Instead, we ask how semantic mappings can bridge the gap between terms used in peoples' daily lives and terms from technical governmental metadata, thus improving the recall and precision of open civic data. Our approach bridges data provider and data user terms by developing a crosscutting base vocabulary that expands core geospatial themes and can be used to better describe civic data. We demonstrate the value of our approach by applying the vocabulary to automatically annotate data on a novel open access platform.

The contributions of this work are as follows:

- A **system** for harvesting provider-contributed data descriptions
- A **base vocabulary** of core geospatial themes mapping provider to consumer descriptions
- A **protocol** for semantically annotating data with core geospatial themes for consumers

The remainder of this paper is organized as follows. Section 2 provides background on the studied open data platform. Section 3 surveys challenges of and approaches focused on improving data discovery. Section 4 discusses the method developed to enrich tags during metadata harvesting. Section 5 describes the resulting implementation. Section 6 discusses the results of the work and presents a research outlook.

¹ <http://geohub.lacity.org/>

2 Background

In order to validate the method design and evaluate results, this work integrates semantic annotation into an open access collaboration platform, Esri's ArcGIS Hub². This platform exposes organizational data via ArcGIS, which is a geospatial data management, visualization, and analytics system used by governments, industry, academia, and other organizations to support planning and operations. ArcGIS integrates desktop software with cloud-hosted tools and data services for distributed information access that can be shared privately or with the public. Using ArcGIS Online, members of organizations and the public can create, edit, and share maps and other data. This global system organizes a content-rich catalog of information across a breadth of scientific themes and operational domains.

ArcGIS Hub is a new open access platform that supports and organizes civic engagement and direct collaboration between governments and their constituents. ArcGIS Hub extends the ArcGIS Online system with new capabilities for open data sharing, configurable metadata catalogs, integration with regional and national metadata registries such as Data.gov³, and analysis tools for the public to visualize and share perspectives on data relationships.

Governments and other enterprises can use ArcGIS Hub to create custom websites for open data sharing that allow the public to easily search, access and download data. ArcGIS Hub's primary audience are the general public: people and groups outside of the organizations sharing the data. While ArcGIS Hub integrates with proprietary software, it also serves as a standalone platform that enables anonymous, public access to datasets from any other platform or data provider; it is not necessary to have any authentication credentials in order to discover or use open access data shared through ArcGIS Hub.

As of early 2018, over 100,000 datasets had been made available through ArcGIS Hub by more than 5,000 governments, academic institutions, and other organizations. These datasets are discoverable by search term, specified by user keyword, and by area of interest, which can be specified by map interface. The current state of search in ArcGIS Hub is based on keyword matching, which matches user queries against dataset titles, descriptions, and tags. A limitation of this type of search however, is that it fails to capture broader or related contexts of the query, only returning content that has a title, description, or tags containing the input term. For example, a search for "bicycle" would not return related content, such as "pedestrian", or broader content, such as "transport".

Civic data providers are primarily focused on making their data available and secondarily focused on making their data discoverable to public consumers, often only providing descriptions or tags when required and often using domain-specific terms. This creates semantic and schematic barriers to data discovery, resulting in a gulf between terms that users and terms that providers use to describe and search for the same data. Resulting challenges to discoverability and current approaches to address them are the focus of the next section.

3 Challenges and Approaches

Data shared through public repositories satisfy basic accessibility requirements, but are often siloed and difficult to discover. A recent report from the Open Research Data Task Force [13] found that the two main challenges to using open data are: 1) finding data to use and 2) (re)using them. While this is especially true of academic data scattered across diverse

² <http://hub.arcgis.com/>

³ <https://www.data.gov/>

domain repositories, it is also true of civic data. The current silos for civic data are not simply organizational, but semantic and schematic, rooted in the technical vocabularies used to categorize and structure data [4]. The main challenges to reusing civic data are the domain-specific terms used to describe data and their attribute schemata [13].

Innovations from the arenas of academia, government, and industry demonstrate contrasting, yet complementary, approaches to addressing discoverability challenges [9]; advances in each arena also inform this work. Recent innovations in discoverability have resulted from the implementation of linked data technologies, which allow for data to be self-describing [2]. The uptake of linked data technologies has resulted in an ever-expanding graph of shared knowledge⁴, replete with reusable ontologies from many domains. Linked data technologies address key semantic and schematic challenges, aiding in many arenas such as in the discovery of scientific data for reuse and discovery across integrated civic data streams [1, 11].

3.1 Semantic Challenges

The first challenge to civic data discovery is semantic. Semantic heterogeneity is understood to result from differing mental models of phenomena as well as from differences in naming conventions; naming heterogeneity can be overcome with term mappings using thesauri, but cognitive heterogeneity is understood to be a more difficult problem to solve in the absence of a minimum set of common definitions [5]. Our work focuses on overcoming heterogeneous naming of semantically similar content, resulting from divergent metadata standards.

The rigor and quality of data classification and tagging schemes can vary greatly by data provider. In the case of highly curated data, such as Federal data layers shared through Esri's Living Atlas of the World,⁵ tags for each dataset have high agreement and control, grouping the data into one of several predefined themes: demographics, transportation, landscape, oceans. . . Similarly, data conforming to the ISO 19115 metadata standard⁶ adhere to a highly controlled vocabulary describing what the contents are about by keyword: agriculture, biota, economy, health. . . However, as of early 2018, only 66,000 (about 8 percent) of the 760,000 items in the ArcGIS Hub catalog had formal metadata.

Metadata files in ArcGIS Hub are also not indexed for search; instead, keyword search in Esri's ArcGIS Hub is based on search by regular expression against the titles, descriptions, and tags of content. Organizations contributing data supply their own tags and descriptions, which results in varying levels of quality. Relatively few tags are based on a controlled vocabulary and descriptions of data have varying levels of completeness. This results in a situation where search for "bicycle collisions" returns results for Washington D.C. where data have been assigned the tags of "transportation" and "collision", but not for the neighboring city of Alexandria, VA where the data have been tagged with "transit" and "accident".

3.2 Schematic Challenges

The second challenge to civic data discovery is schematic. Schematic heterogeneity is understood to result from variations of conceptual schemata within or across disciplines; it can be overcome by schema integration [5].

Governmental organizations such as law enforcement agencies that report traffic accidents, including bicycle collisions, adhere to such integrated specifications, in this case the Model

⁴ <http://lod-cloud.net/>

⁵ <https://livingatlas.arcgis.com/en>

⁶ <https://www2.usgs.gov/science/about/thesaurus-full.php>

Minimum Uniform Crash Criteria (MMUCC)⁷ developed by the National Highway Traffic Safety Administration (NHTSA). This data model provides a reporting schema; local agencies can adapt it as needed, but it defines a minimum set of uniform fields that can be identified across municipal crash datasets. These criteria specify attribute names (i.e. “County Name”), definitions, and expected data types (i.e. “GLC Code”). Another well-adopted data model developed with interoperability in mind is the Local Government Information Model⁸. Similarly, it defines feature datasets (i.e. “Facilities Streets”), feature classes (i.e. “street lane width”), and attribute fields (i.e. “lane width, type: small integer”).

Where common data models are used, it is possible to easily reuse, and even combine, datasets. However, the majority of data discoverable through Esri’s ArcGIS Hub do not conform to any common data models. Attribute fields are defined ad-hoc and are also not indexed for search unless specified separately as tags.

3.3 Linked Data Approaches

The need for improved access to civic data parallels that for academic data. Just as research groups, or even academic domains, publish and reuse data according to different standards across various repositories, governmental agencies and municipalities also adhere to a variety of standards with varying levels of quality. The rise of Internet of Things (IoT) technology, which is enabling the evolution of “smart cities”, has also created new sets of challenges related to the volume, velocity, and variety of civic data streams. The challenges that have made heterogeneous civic data difficult to integrate and harmonize in the past have been successfully met by semantic annotation of data streams, which enables their alignment [3].

Rather than semantically annotating civic data after the fact, some governments have adopted linked open data principles as a standard for data sharing; “smart cities” such as London⁹ and Dublin¹⁰ have launched campaigns to expose operational city service data streams in an open, consumable format [7]. Esri Ireland for example now serves national geospatial information as linked data, consumable through an API [6]. In a linked data framework, it is not only easier for both humans and machines to consume civic data, but it is also easier to combine data from multiple sources, for example across levels of government.

One reason for this is that semantically annotated data can be dereferenced, resolving issues of uncertainty concerning attribute values or terms. For example, the United Nations Sustainable Development Goals ontology resolves terminological ambiguity while tracking progress toward shared goals on a multinational scale [11]. The outcomes of such successful linked data approaches motivated us to develop a similar method for semantically annotating civic data in order to improve user search. This method is the focus of the next section.

4 Methods

In order to improve the discoverability of civic data, we have developed and implemented a base vocabulary and a semantic annotation system. Semantic annotation augments official metadata with relevant tags supplied by a vocabulary, thus expanding the relevance recommendations of data. The method taken to develop and implement an automated semantic annotation system is summarized in the following steps:

⁷ <https://www.nhtsa.gov/mmucc>

⁸ <http://solutions.arcgis.com/local-government/help/local-government-information-model/>

⁹ <http://connected-data.london/>

¹⁰ <http://smartdublin.ie/>



■ **Figure 1** ArcGIS Hub categories reflect existing themes assigned to datasets manually as tags.

1. Formalize base vocabulary for core geospatial themes
2. Extend vocabulary by reusing existing concept hierarchies
3. Augment existing metadata with extended tag hierarchies
4. Evaluate system performance for search

4.1 Formalizing the Base Vocabulary

A key contribution of this work is the development and formalization of a compact base vocabulary that maps prototypical themes of government departments to aspects of users' lives. This vocabulary addresses two geospatial problems: 1) it makes data shared by governments that are co-located or adjacent discoverable; and 2) it makes descriptions of the phenomena that data are about semantically relevant to public users. The base vocabulary categories shown in Figure 1 were developed in collaboration with civic stakeholders, municipal staff, research organizations, and Esri's Local Government Team¹¹. The vocabulary holistically organizes data and tools, allowing them to be referenced.

While these categories reflect typical organizational structures of civic government, they also capture core geospatial themes that communities want to track and measure. These categories are currently used as search facets for data in ArcGIS Hub. While they may structurally reflect issues that communities prioritize, they may not reflect the terms that community members may use when searching for this data. They also may not reflect the terms that a given organization uses to describe its data.

In order to formalize ArcGIS Hub Categories, we began by building a thesaurus of concepts modeled in Protégé¹², an open source ontology editing software. We opted for a pragmatic adoption of the Simple Knowledge Organization System (SKOS) to model these concepts for a number of reasons: SKOS supports flexible modeling of hierarchical relationships; it is

¹¹<http://www.esri.com/software/arcgis/arcgis-for-local-government>

¹²<https://protege.stanford.edu/>

used widely across numerous domains; and it is often used in term expansion activities¹³. For these reasons, we were able to reuse authoritative and dereferenceable concepts already published to the Semantic Web by organizations also using SKOS.

Some data available through ArcGIS Hub, such as layers exposed through Esri's Living Atlas of the World, have already been classified and tagged with ArcGIS Hub Categories. These include broader categories like "healthy" and narrower categories like "disease".

However, user-specified terms are not reflected in the ArcGIS Hub Categories. Analysis of the ArcGIS Hub query log revealed that users of Esri's ArcGIS Hub tend to search for data using terms that relate to their own colloquial conceptualizations of theme and geography. In a sample of 470,796 queries performed in 2015, only 12,257 (or 2.6 percent) used any form of the predefined categorical Hub keywords, (i.e. "healthy", "transportation", ...). This means that the majority of themes present in user searches likely take another form. This could mean that users are searching with synonyms of these keywords (i.e. "well-being"), or narrower concepts (i.e. "bicycle"), which would not yield results. Similarly, in the same sample of queries, only 64,353 (or 27.3 percent) use geographic references, like coordinates, addresses, place types, or zip codes in their searches. Similarly, geographic concepts that reflect place hierarchy (i.e. "Ronald Reagan National Airport is in Arlington County, VA") or proximity (i.e. "Reagan Airport is next to East Potomac Park") are not reflected in results.

4.2 Extending the Base Vocabulary

We imported existing concepts matching the Hub category tags from Library of Congress Subject Headings (LCSH)¹⁴, Princeton WordNet 3.1¹⁵, and the USGS Thesaurus¹⁶. Reusing these three vocabularies to describe civic data is novel, as they have been developed and traditionally used to describe library resources and scientific data. These vocabularies provide sufficient terminological coverage for extending the Hub categories shown in Figure 2.

LCSH are a controlled and well-defined set of terms used for resource classification. In addition to providing a stable identifier, LCSH concepts also adhere to a SKOS scheme and provide broader, narrower, and related concepts for each term. For example, "agriculture" in LCSH has useful variants "farming" and "husbandry", narrower terms like "agronomy", and related terms like "food supply" and "land use, rural". LCSH is designed to be used as a thesaurus; its subject headings provide bibliographic access to related subject matter.

Similarly, WordNet terms are also available in a SKOS scheme and are consumable as RDF, a linked data model. WordNet is a lexical database that combines the capabilities of a dictionary and a thesaurus for the English language. Concepts matching Esri Hub categories were retrieved from WordNet synsets, which are sets of synonyms with translations. For example, the synset for "agriculture" in WordNet includes "husbandry" and "farming" along with multilingual translations for each. Designed to support cognitive science applications, WordNet is suitable for information retrieval, text classification, and translation tasks [10].

A final source of Hub concept extension comes from the United States Geological Survey (USGS) Thesaurus, which is currently under development. As such, it provides identifiers without dereferencing; despite this, it is a rich source of authoritative scientific definitions and related terms in a SKOS scheme. For instance, it provides examples of the term agriculture used in the topics of "farming" and "horticulture". The USGS Thesaurus is designed to aid public interpretation of science web resources and topics.

¹³<https://www.w3.org/TR/skos-ucr/>

¹⁴<http://id.loc.gov>

¹⁵<http://wordnet-rdf.princeton.edu>

¹⁶<https://www2.usgs.gov/science/about/thesaurus-full.php>



■ **Figure 2** Extension of ArcGIS Hub terms to related categories in existing vocabularies.

Other sources were experimented with but ultimately were not implemented. Schema.org¹⁷ was considered for thematic and geographic expansion, but was rejected as its top-level concepts are too broad, while narrowing too quickly. Geonames and DBpedia were also investigated, but have not yet been implemented; concepts from these sources may be included in the near future, as both are rich sources of colloquial place-types and themes found in users’ daily lives. It will be possible to extend the base vocabulary following the method developed in this work as other candidate vocabularies are considered.

To further expand ArcGIS Hub terms, we undertook additional mappings from existing categories to community standards, including INSPIRE¹⁸, FGDC¹⁹, and ISO 19115 data specifications. INSPIRE provides 34 spatial data themes, which specify common data models and code lists. INSPIRE themes aim to support the creation of a European Union spatial data infrastructure. These themes include “hydrography”, “transport networks”, and “protected sites”. Similarly, the National Geospatial Data Asset (NGDA) provides a set of 16 themes with appointed lead agencies and the aim of supporting data interoperability. These themes include “climate and weather”, “land use-land cover”, and “soils”. Finally, ISO 19115 provides a set of 19 themes, including terms like “biota”, “health”, and “oceans”. Each of these community standards function as a controlled vocabulary for describing spatial data resources in their respective metadata contexts; their terms overlap to varying extents.

Pragmatically, we were interested in areas of term overlap, as mapping these standardized community terms to the expanded set of ArcGIS Hub terms establishes semantic links between thematically related resources. Various agencies conform to these standards when describing their data. Federal agencies, such as the U.S. Geological Survey, use NGDA themes to describe resources shared through ArcGIS Hub. The FGDC for example maintains a keyword thesaurus with these terms and points to it as a best-practices resource for publishing

¹⁷ <http://schema.org/docs/schemas.html>

¹⁸ <https://inspire.ec.europa.eu/data-specifications/2892>

¹⁹ <https://www.fgdc.gov/what-we-do/manage-federal-geospatial-resources/a-16-portfolio-management/themes>

The figure is split into two panels. The left panel shows a SPARQL query interface with a query template. The right panel shows the 'Term Expander' interface with the term 'agriculture' and its expanded JSON results.

SPARQL query template (left):

```

2 PREFIX hub: <http://www.esri-hub.com/vocab/>
3 PREFIX owl: <http://www.w3.org/2002/07/owl#>
4 PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
5 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
6 PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
7 PREFIX ontology: <http://wordnet-rdf.princeton.edu/ontology#>
8 PREFIX loc: <http://ld.loc.gov/authorities/subjects/>
9
10 SELECT ?label ?synonym
11 WHERE {
12   ?term skos:prefLabel ?label .
13   ?term skos:altLabel ?synonym .
14   FILTER contains( lowercase(?label), "agriculture" )
15 }

```

Term Expander (right):

Term: agriculture

```

{
  "synonyms": [
    "Farming",
    "Husbandry",
    "agribusiness",
    "agricultura",
    "agroindustria",
    "factory farm"
  ],
  "translation": [
    "الزراعة الحضرية",
    "الزراعة",
    "الزراعة",
    "agricultura",
    "landbrug",
    "laborantza",
    "nekezaritza",
    "الزراعة",
    "maanviljelyst",
    "maanviljelyst",
    "maatloos",
    "agriculture",
    "cucuk tanam",
    "pengebanan",
    "penternakan",
  ]
}

```

■ **Figure 3** SPARQL query template (left) and expanded terms (right) for term “agriculture”.

datasets to open data clearinghouses; it states that “the more robust your theme keyword list, the more likely it can be located by others (and yourself)”. While this is true in principle, describing data with controlled keywords alone will not make data readily discoverable for public consumers of data who often search for data using colloquial terminology.

In order to augment official metadata, the controlled vocabularies for INSPIRE, NGDA, and ISO 19115 were incorporated into the expanded ArcGIS Hub terms. We designated mappings between related terms from each controlled vocabulary in Protégé using the SKOS predicate *related*. Thus, a term like “transportation” has: *related* terms from INSPIRE (“Transport networks”), NGDA (“Transportation”), ISO 19115 (“Transportation”); *broader* and *narrower* terms from LCSH and USGS Thesaurus (“public transit”); and *synonyms* and *translations* from WordNet (“ES - transporte”). Each of these tags becomes a triple statement pointing to externally defined resources.

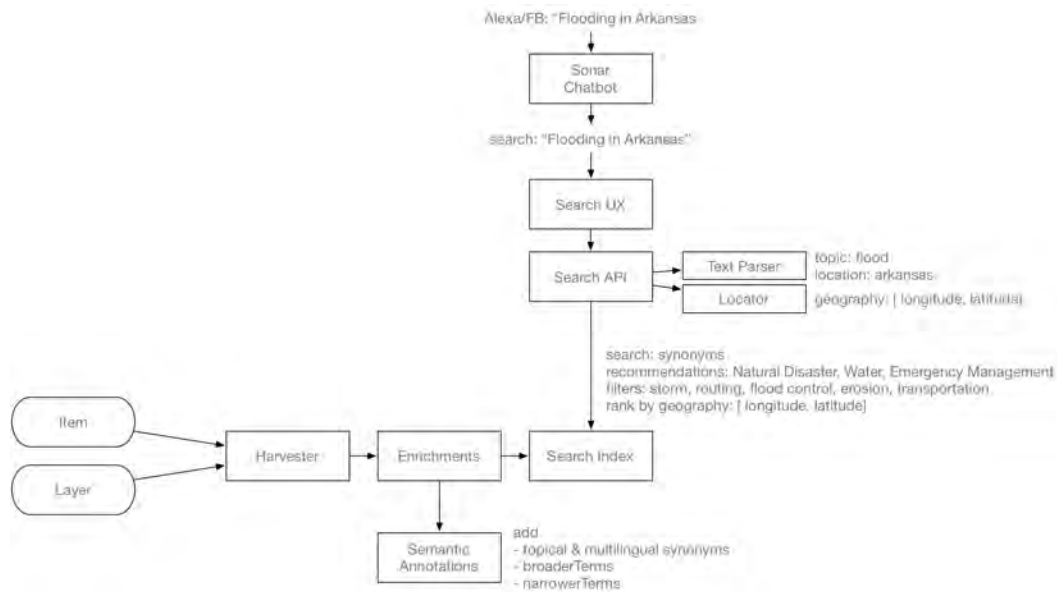
4.3 Augmenting Existing Metadata

We exported the base vocabulary from Protégé as triple statements in Terse RDF Triple Language (Turtle)²⁰ syntax and imported them into a Fuseki²¹ triplestore, set up as a public endpoint. The vocabulary is stored as a graph that can be queried using SPARQL syntax, which allows for queries across multiple endpoints. Figure 3 shows an example of a query template in Fuseki returning query results in JSON to be integrated as auxiliary metadata.

ArcGIS Hub includes a search index of aggregated dataset records from all data providers. When organizations like governments indicate their data is public, ArcGIS Hub compiles multiple metadata sources into a custom search index to support multiple content search and discovery services.

²⁰ <https://www.w3.org/TeamSubmission/turtle/>

²¹ <https://jena.apache.org/documentation/>



■ **Figure 4** Semantic annotations added to metadata, supporting search through query expansion.

The search index process, shown in Figure 4, includes three phases: harvesting, validation, and enrichment. During harvesting of a dataset, Hub collects metadata from the ArcGIS Online item information, associated formal metadata, the feature service and feature layer definition, and data attribute aggregate statistics. Validation includes heuristics to measure metadata completeness, support for secure connections with HTTPS, and query responsiveness, which determines if the data are actually accessible. During the enrichment phase, a dataset is decomposed into relevant keywords which are then sent to the semantic query service to retrieve new semantic tags that are then attached to the dataset metadata.

For example, Flood Zone data from Evansville, Indiana are tagged “Evansville, Vanderburgh County, Flood Zones, IN, environmental”. Using each of the terms from each of the tags results in a superset of synonyms, translations, broader terms, and narrower terms, shown in Figure 5. These terms are each added to the dataset record in the search index using an internal semantic annotation service. The semantic annotation service is an internal API that hosts the base vocabulary as a queryable API using the Apache Jena Fuseki server. This server supports defined requests to build a set of tags that expand the dataset metadata for broad, narrow, translated, and similar terms.

At query time, these additional terms can be used to match user queries such as “human health”, or “impact assessment” that may not have another similar word match in the dataset metadata collection but will now have results based on matching these new, additional semantic tags. The semantic tags also include translations such as “air pasang” (Indonesian) or “nousuvesi” (Finnish). Beyond similar terms, there are broader terms such as “Natural disasters”, and “Water” and narrower terms such as “Flood damage prevention” and “Forest influences” that can be used to recommend new search terms to the user for refining their search results.

4.4 Evaluating system performance for search

In practice, search for data is now semantically aided; related content, such as synonymous terms, can be retrieved when inferred as thematically related. For example, a search for traffic accidents can now return other content related to a broader concept of ‘transportation’ as pedestrian fatalities. While only a small fraction of data (about 8 percent) in ArcGIS Hub initially included formal metadata, semantic annotations added related metadata in the form of related terms, supporting data discoverability and integration.

In order to evaluate the contributions of our approach, we consider that semantically enabled search wasn’t previously possible: this informs our baseline criteria. Search efficacy is measured accordingly using several methods: conversion rates through usage analytics tracking, usability testing, and relevance judgment evaluation.

Usage analytics tracking measures all user interactions with the ArcGIS Hub web application. This includes search inputs, filter interactions and result selection. We define several conversion funnels corresponding to expected user outcomes, which include downloading the data, creating an information product such as a web map or a Story Map, or bookmarking a view of the data for later use. These conversions indicate that a good search result was returned. We can then compare conversion results with and without semantic annotation.

Usability testing includes defining a workflow that human test subjects perform while being monitored by researchers. Listening to stream of consciousness verbal evaluations and observing interface interactions denotes perceptions of different search modalities and outcomes. This testing may be performed in-house or in collaboration with stakeholders.

Lastly, relevance judgment evaluation asks a similar set of users to evaluate the quality of search results as: perfect, relevant, partially relevant, or irrelevant. The scores for each result are tallied and compared with the optimal result and rank ordering to define the quality of the search relevance, due to semantic annotations or without semantic annotations.

The results of these evaluation measures are forthcoming at this time of writing.

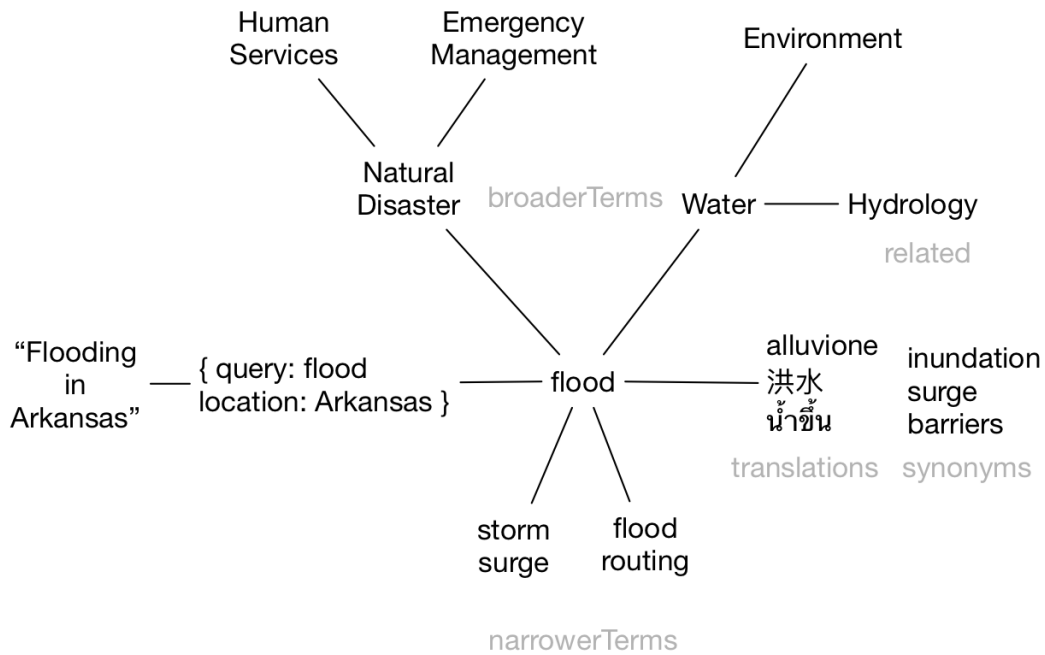
5 Results

Governments, academic institutions and other organizations publish open data to encourage the creative reuse of information for new purposes. ArcGIS Hub allows these organizations to create websites that enable search and discovery of their authoritative data, as well as recommend data shared through other groups. The Bureau of Transportation’s Geospatial Statistics site is shown as one such example in Figure 6. Visitors can perform simple searches through their web browser or mobile device, or request information through new digital media chatbots on Facebook and Amazon Alexa.

Extending dataset metadata with semantic annotations expands the discoverability of information through colloquial and multilingual search associations. Figure 4 illustrated how search queries use the semantic search index to parse and retrieve relevant datasets.

To use the semantic search API, Hub implements a REST HTTP API for structured queries from web browsers, mobile apps, and custom embeds; it uses a JSON-Schema self-documenting hypermedia API and includes search index attribute filters and facets. An API search query is first split into relevant parameters for keywords, time, location, and provider. The keywords are compared with the semantic annotation tags for similarity matches; the time, location and provider are used as filters. The result includes a relevance-ranked list of datasets as well as aggregate facets of topics, data types, and providers for further filtering.

The semantic annotations augment the search relevance matches by comparing search keywords with terms that may not have existed in the original metadata document, but describe the dataset with alternative labels that match these queries. Figure 5 shows an



■ **Figure 5** Search queries are parsed and compared with semantic annotations to expand matches and provide additional facets.

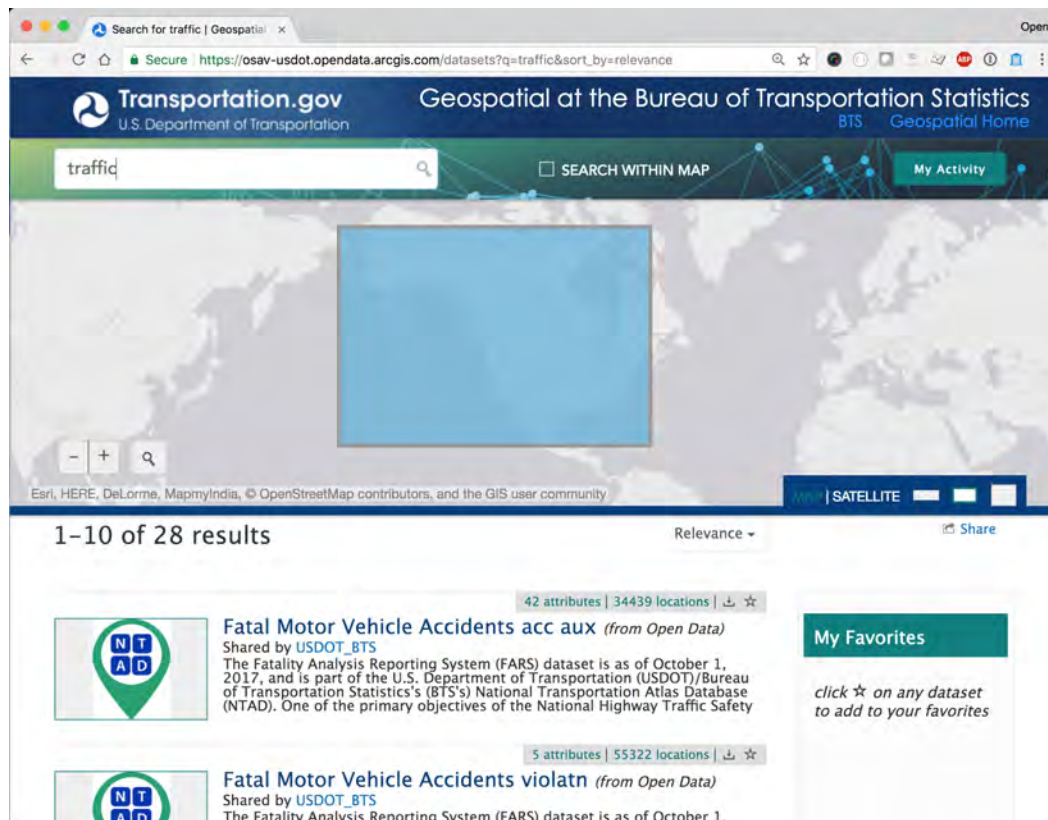
example of expanded semantic tags that are compared for relevance ranking, including multilingual terms, as well as the broader and narrower aggregate terms that can be used in search interface facets.

5.1 Building Data Networks

Semantic annotation supports additional use cases beyond metadata querying. ArcGIS Hub includes a global catalog of data from governments of various administrative levels: local council and departmental, metropolitan, provincial, regional, national, and multinational organizations. Each government follows a varying set of metadata and keyword standards that may not overlap with other governments, even if the organizations are geographically adjacent or coincident. This can make integration of data across municipal boundaries problematic, resulting in lost productivity or detriments to operations and safety.

Semantic annotations support data integration by organizing datasets into common thematic groupings, which increase the discovery and utilization of similar datasets across municipal data providers. By way of example, consider several civic datasets provided by neighboring municipalities such as road networks, public schools, moving violations (e.g. vehicle speeding citations), and reported crashes between vehicles, bicycles or people. Additionally, there are regional and national datasets provided by agencies that also include transit networks (bus stops and train stations): FARS (Fatality Analysis Reporting System).

In order to track progress toward thematic community initiatives, such as “Vision Zero”, discovery of relevant data must be possible across all levels of government. Vision Zero is a strategy to eliminate all traffic fatalities and severe injuries, while increasing safe, healthy, equitable mobility for all. Potential Federal data sources for tracking a “Vision Zero” Initiative are shown in Figure 6. However, without semantic annotation, there is uncertainty as to



■ **Figure 6** U.S. Department of Transportation traffic related datasets sorted by relevance.

whether a search for traffic data will return relevant results across other Hub sites at a state, county, or municipal level.

ArcGIS Hub builds the search index that includes each of the four example local municipal datasets from each municipality. This includes the original metadata and the additional semantic annotations on the datasets that associate them with related thematic groupings. Searching just the category term has mixed, or missing, results from some provider catalogs. Figure 7 compares search results across the GIS catalogs of the District of Columbia, State of Maryland, and County of Arlington, exposed through ArcGIS Hub.

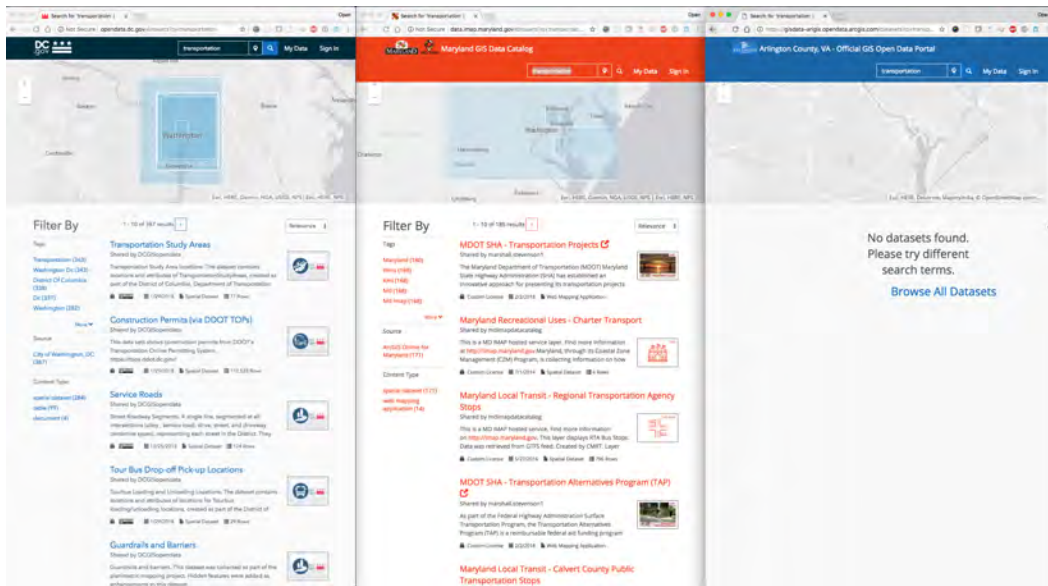
By comparison, when colloquial terms are used, there are similar results from all local providers. Figure 8 compares search results across the same GIS catalogs for related terms.

6 Discussion and Outlook

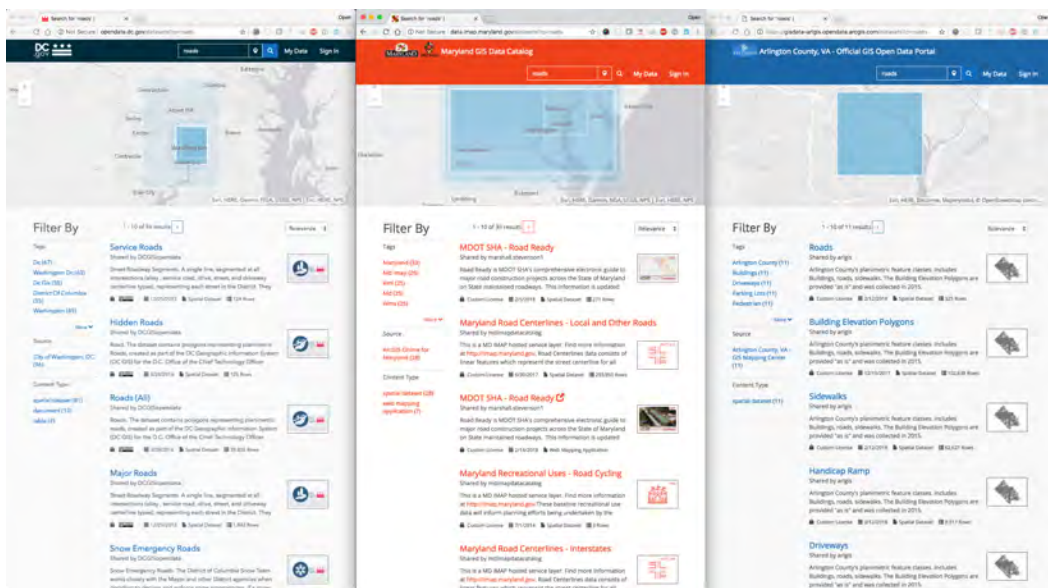
The work presented in this paper improves data discovery through the application of semantic annotations to civic data, which facilitate transparency and coordination of work; semantic search enables the exploration and discovery of relationships among organizations' data that were previously unknown.

Several areas of research are continuing from this work. We plan to expand and refine the base vocabulary to better support bi-directional term expansion. This will allow users to discover new datasets by improving traversal of the base vocabulary's relations, like broad and narrow terms, for both thematic and geographic concepts. We anticipate that alignment with new ontologies, such as the U.N.'s Sustainable Development Goals Ontology,

9:14 Improving Civic Data Discovery



■ Figure 7 Comparing searches for “transportation” before adding semantic annotations.



■ Figure 8 Comparing searches for related term “roads” after adding semantic annotations.

and application of our methods in related domains, such as academic libraries, will continue to improve data discovery across organizational repositories.

On a larger scale, the lessons learned from our research can be applied to new domains and extended along the following dimensions.

The Sustainable Development Goals (SDGs) are the results of an ambitious global initiative to improve the health and well-being of people and communities. They consist of 17 goals, 169 targets and 232 data indicators that will measure and monitor progress towards the SDG. These targets and indicators include a semantic graph that relate to socioeconomic terms, municipal planning, and other related governance sectors. Work is ongoing with

several national mapping agencies and the United Nations to integrate their semantic graphs with the base vocabulary presented in this paper.

We are also applying the methods developed in this paper to data discovery in the context of digital research libraries. While libraries have long been the traditional brokers of knowledge, today's queries are largely mediated by commercial digital search engines [12]. Yet, libraries are taking on new roles, facilitating discovery, and often co-production, of knowledge [8]. Semantically annotated data can be more easily discovered and retrieved via queries that traverse knowledge graphs, regardless of the endpoints where they are hosted. Academic libraries are poised to serve as a semantically-neutral meeting ground where domain data can be aggregated and made spatially and thematically discoverable, similar to ArcGIS Hub.

References

- 1 Sean Bechhofer, David De Roure, Matthew Gamble, Carole Goble, and Iain Buchan. Research objects: Towards exchange and reuse of digital knowledge. *Nature Precedings*, 2010. doi:10.1038/npre.2010.4626.1.
- 2 Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific american*, 284(5):34–43, 2001.
- 3 Stefan Bischof, Athanasios Karapantelakis, Cosmin-Septimiu Nechifor, Amit P Sheth, Alessandra Mileo, and Payam Barnaghi. Semantic modelling of smart city data. In *Report of the W3C Workshop on the Web of Things 2014*, 2014. URL: <https://www.w3.org/2014/02/wot/papers/karapantelakis.pdf>.
- 4 Wade Bishop and Tony H Grubestic. Geographic information, maps, and gis. In *Geographic Information*, pages 11–25. Springer, 2016.
- 5 Yaser Bishr. Overcoming the semantic and other barriers to gis interoperability. *International journal of geographical information science*, 12(4):299–314, 1998.
- 6 Christophe Debruyne, Éamonn Clinton, Lorraine McNerney, Atul Nautiyal, and Declan O'Sullivan. Serving ireland's geospatial information as linked data. In *International Semantic Web Conference (Posters & Demos)*, 2016.
- 7 Rob Kitchin. The real-time city? big data and smart urbanism. *GeoJournal*, 79(1):1–14, 2014.
- 8 Sara Lafia, Jon Jablonski, Werner Kuhn, Savannah Cooley, and F Antonio Medrano. Spatial discovery and the research library. *Transactions in GIS*, 20(3):399–412, 2016.
- 9 Matthew S Mayernik. Research data and metadata curation as institutional issues. *Journal of the Association for Information Science and Technology*, 67(4):973–993, 2016.
- 10 George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
- 11 Barry Smith and Mark Jensen. The unep ontologies and the obo foundry. In *ICBO/BioCreative*, 2016.
- 12 Elaine Svenonius. *The intellectual foundation of information organization*. MIT press, 2000.
- 13 Open Research Data Taskforce. Research data infrastructures in the uk : Landscape report. Technical report, Universities UK, 2017.
- 14 Andrew Turner. Desire paths to open data. <http://highearthorbit.com/articles/desire-paths-to-open-data>, 2014.
- 15 Anneke Zuiderwijk and Marijn Janssen. Open data policies, their implementation and impact: A framework for comparison. *Government Information Quarterly*, 31(1):17–29, 2014.

Local Co-location Pattern Detection: A Summary of Results

Yan Li

Department of Computer Science & Engineering, University of Minnesota, Minneapolis, MN, USA
lix4266@umn.edu

Shashi Shekhar

Department of Computer Science & Engineering, University of Minnesota, Minneapolis, MN, USA
shekhar@umn.edu

Abstract

Given a set of spatial objects of different features (e.g., mall, hospital) and a spatial relation (e.g., geographic proximity), the problem of local co-location pattern detection (LCPD) pairs co-location patterns and localities such that the co-location patterns tend to exist inside the paired localities. A co-location pattern is a set of spatial features, the objects of which are often related to each other. Local co-location patterns are common in many fields, such as public security, and public health. For example, assault crimes and drunk driving events co-locate near bars. The problem is computationally challenging because of the exponential number of potential co-location patterns and candidate localities. The related work applies data-unaware or clustering heuristics to partition the study area, which results in incomplete enumeration of possible localities. In this study, we formally defined the LCPD problem where the candidate locality was defined using minimum orthogonal bounding rectangles (MOBRs). Then, we proposed a Quadruplet & Grid Filter-Refine (QGFR) algorithm that leveraged an MOBR enumeration lemma, and a novel upper bound on the participation index to efficiently prune the search space. The experimental evaluation showed that the QGFR algorithm reduced the computation cost substantially. One case study using the North American Atlas-Hydrography and U.S. Major City Datasets was conducted to discover local co-location patterns which would be missed if the entire dataset was analyzed or methods proposed by the related work were applied.

2012 ACM Subject Classification Information systems → Geographic information systems, Information systems → Data mining

Keywords and phrases Co-location pattern, Participation index, Spatial heterogeneity

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.10

1 Introduction

Given instances of different spatial features (e.g., mall, hospital) and a spatial relation (e.g., geographic proximity), the problem of local co-location pattern detection (LCPD) pairs co-location patterns and localities such that the co-location patterns tend to exist inside the paired localities. A co-location pattern is a set of spatial features, the instances of which are often related to each other. The LCPD problem is one of the variants of co-location pattern detection problem, which focuses on detecting co-location patterns globally in the entire dataset [9]. Intuitively, if a co-location pattern is infrequent relative to all input instances, it may be neglected in the entire dataset, but more easily found in a subset of the dataset around its spatial footprint. The uneven distribution of spatial features in the space, i.e., spatial heterogeneity, is common, so the local existence of co-location patterns in an area is



© Yan Li and Shashi Shekhar;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 10; pp. 10:1–10:15

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

not unusual. For example, high NOx emissions from buses may occur with certain engine events only around the bus depot where the route starts, since the engines have not warmed enough to perform efficiently. Other examples include high NOx emission and elevation change in rural areas as illustrated in the Volkswagen emissions scandal [8], and assault crimes and drunk driving events near bars [10]. Because of its societal importance, LCPD has attracted growing attention recently.

In this paper, we will focus on detecting local co-location patterns with the locality defined using minimum orthogonal bounding rectangles (MOBRs). An MOBR is a rectangle with sides parallel to the coordinate system. It is widely used as an approximation of complex shapes by minimally enclosing them [13]. However, the enumeration of MOBRs is computationally challenging. Given a set of spatial objects in a 2-dimensional space, the number of the set’s subsets is exponentially related to its cardinality. Each of the subsets has an MOBR, so the number of MOBRs is also exponentially related to the number of the input objects. Moreover, the relationship between the participation index, a widely adopted metric for co-location patterns [9], in any pair of localities cannot be determined without considering the distribution of spatial objects within them.

The related work on the LCPD problem falls into two categories. The first line of research applies data-unaware space-partitioning heuristics (e.g. Quadtree, grid), which ignores the spatial distribution of data and may break up potential localities. The second class defines localities using clusters of spatial objects or co-location instances, but neglects other localities without a cluster.

Contributions. To detect local co-location patterns in all rectangular localities with sides parallel to the coordinate system, we first formally define the LCPD problem. Then, we present a Quadruplet & Grid Filter-Refine algorithm that leverages an MOBR enumeration lemma, and a novel upper bound on the participation index. The experimental evaluation shows that the proposed algorithm reduces the computation cost substantially. One case studies on North American Atlas-Hydrography and U.S. Major City Datasets was conducted to discover local co-location patterns which would be missed if the entire dataset was analyzed or methods proposed by the related work were applied.

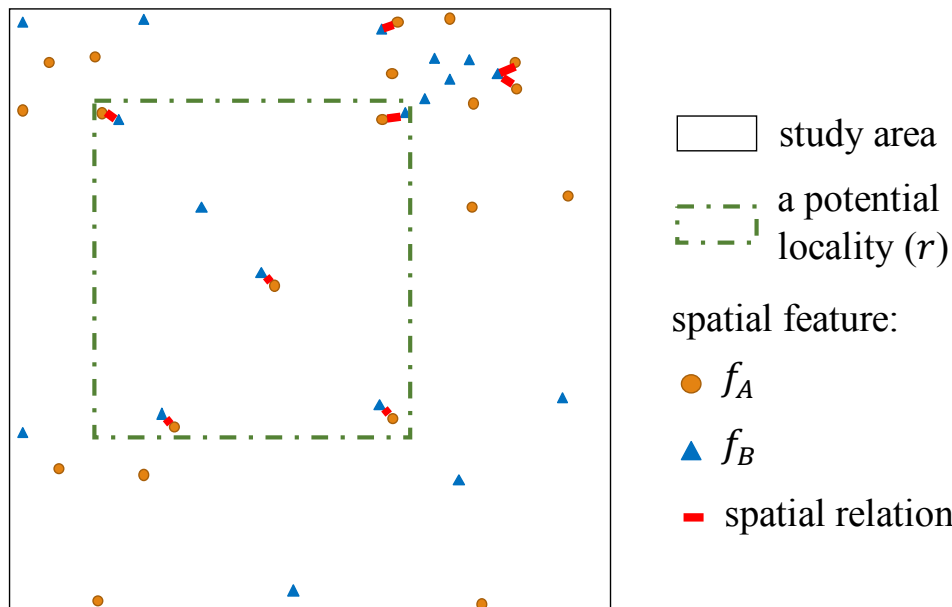
This paper is organized as follows: In §2, we explain the basic concepts and formally define our local co-location pattern detection problem. §3 reviews the related literature. §4 presents our algorithms for solving the problem, whose evaluation is given in §5. §6 concludes the paper and presents our future work.

2 Basic Concepts and Problem Statement

2.1 Basic Concepts

Huang et al. define the input, output and the interest measures for detecting co-location patterns globally through data in [9].

Each spatial **object**, composed of a boolean **feature** (e.g., mall, hospital) and a spatial location, can be related to others through a spatial **relation** (e.g., neighborhood). A **co-location pattern** is a set of features. An instance of a co-location pattern is a set of objects of every distinct feature in the pattern which can form a clique given the input relation. In the dataset shown in Figure 1, there are 20 objects of feature f_A (circle) and 18 objects of feature f_B (triangle), and the related objects are linked. Only one co-location pattern, $\{f_A, f_B\}$, exists, and it has 8 instances.



■ **Figure 1** A local co-location pattern $\langle \{f_A, f_B\}, r \rangle$.

The **participation ratio** of a feature f_i in a co-location pattern C , $pr(C, f_i)$, is the fraction of objects of the feature participating in instances of the pattern. The **participation index** of the pattern, $pi(C)$, is the minimal participation ratio of the features in the pattern. In Figure 1, for the co-location pattern $C = \{f_A, f_B\}$, $pr(C, f_A) = \frac{8}{20}$ and $pr(C, f_B) = \frac{7}{18}$, so $pi(C) = \frac{7}{18}$.

By extending these concepts, we introduce the following ones for the LCPD problem.

The **study area** is defined as the minimum orthogonal bounding rectangle (MOBR) of all input objects, whose subsets are **localities**. A **local co-location pattern** is a pair of a co-location pattern (C) and a locality (r), in the form of $\langle C, r \rangle$. Its instances and interest measure are the corresponding values of its co-location pattern in its locality. A locality where objects of features in a co-location pattern tend to be related to each other (determined by a participation index threshold) is called the pattern's prevalence locality.

In Figure 1, for a local co-location pattern $C_r = \langle \{f_A, f_B\}, r \rangle$, there are 5 instances, while $pr(C_r, f_A) = \frac{5}{5}$, $pr(C_r, f_B) = \frac{5}{6}$, and $pi(C_r) = \frac{5}{6}$. If the participation index threshold is 0.5, r is a prevalence locality of the pattern $\{f_A, f_B\}$.

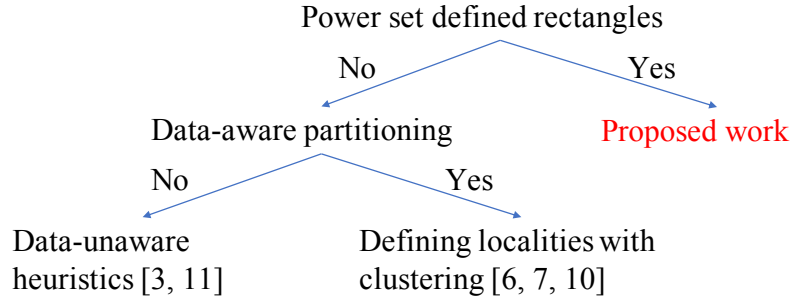
2.2 Problem Statement

Based on the above concepts, we can formally define the LCPD problem as follows:

Input:

- A set of spatial objects.
- A spatial relation on the objects.
- A participation index threshold θ .
- A co-location instance number threshold γ .

Output: Local co-location patterns with participation index $\geq \theta$ and the number of instances $\geq \gamma$.



■ **Figure 2** The related work.

Objective: Computational efficiency.

Constraints:

- Correctness and completeness of the result set.
- The co-location instance number threshold $\gamma \geq 2$.
- The locality of a local co-location pattern is the MOBR of its co-location instances.

If given the objects and relation in Figure 1, as well as thresholds $\theta = 0.5$ and $\gamma = 3$, $\langle \{f_A, f_B\}, r \rangle$ is one of the eligible results with a participation index of $\frac{5}{6}$ and 5 instances. The co-location instance number threshold is set to prevent the problem from degradation. A locality containing only one co-location instance may be a prevalence locality, but it is meaningless.

The MOBRs of a set of co-location instances, which are the localities detected by the algorithms, can be regarded as the representatives of the infinite number of arbitrarily rectangles with sides parallel to the coordinate system according to the following lemma.

► **Lemma 1.** *Given any arbitrarily rectangular prevalence locality of a co-location pattern with sides parallel to the coordinate system, the MOBR of the pattern's instances within it is also a prevalence locality of the pattern.*

Proof. For any feature f in a co-location pattern C , let n_r and n_{MOBR} denote the number of objects of f in an arbitrary rectangular prevalence locality r of C and the MOBR of C 's instances in r , while m_r and m_{MOBR} denote the number of those participating in C 's instances. Thus, $pr(\langle C, r \rangle, f) = \frac{m_r}{n_r}$, while $pr(\langle C, MOBR \rangle, f) = \frac{m_{MOBR}}{n_{MOBR}}$. According to the definition of MOBR, and that $MOBR \in r$, we have $m_r = m_{MOBR}, n_r \geq n_{MOBR}$, so $\frac{m_r}{n_r} \leq \frac{m_{MOBR}}{n_{MOBR}}$. Now that $\frac{m_{MOBR}}{n_{MOBR}} \geq \frac{m_r}{n_r} \geq pi(\langle C, r \rangle) \geq \theta$, the MOBR is a prevalence locality as well. ◀

3 Related Work and Limitations

In order to solve the LCPD problem, many methods have been proposed, which can be generalized into two steps. The first step is partitioning the study area into potential localities based on certain heuristics, which is followed by checking the eligibility of the localities. Based on whether the heuristics are data-aware, these methods belong to two classes (the right branch in Figure 2).

A good example using data-unaware heuristics is [3] in which Celik et al. use a QuadTree structure to divide the study area into localities, but it requires sophisticated domain knowledge to predefine localities. In another example, a grid is used to divide the study area

into cells, and arbitrary subgraphs of the cells' neighbor graph are regarded as localities [12]. Both approaches share the same limitation with others using data-unaware heuristics, that is, the partitioning scheme employed is independent of the spatial distribution of the data, which may break up potential localities [10].

The other class of methods using data-aware heuristics defines localities with clusters of spatial objects or co-location instances. In [7], localities grow from initial localities with high objects concentration. Mohan et al. define localities as areas delineated by neighbor graphs of spatial objects [10]. Deng et al. explore footprints of co-location instance clusters with an adaptive density threshold as localities [6]. These methods are not complete because localities without object or co-location instance concentrations may be eligible as well.

Our proposed work, on the other hand, detects local co-location patterns in all rectangular localities with sides parallel to the coordinate system, so the method will enumerate the MOBRs determined by all subsets of co-location instances (the elements in co-location instances' power set). Consider the dataset shown in Figure 1 as an example. If the participation index threshold is set as 0.6, the co-location pattern $\{f_A, f_B\}$ is not a eligible pattern globally through the data, because its participation index is $\frac{7}{18}$. However, our proposed work will find a prevalence locality for the pattern (green dash rectangles in Figure 3a), where the participation index is $\frac{5}{6}$. Contrarily, The participation index in the locality determined by the cluster of co-location instances shown in Figure 3b is $\frac{3}{7}$, while Figure 3c and 3d present the localities with the highest possible participation index if the study area is partitioned using the Quadtree and grid in them, where the participation index is $\frac{3}{7}$ in both cases. None of the currently available results in eligible patterns, so it is obvious that the proposed work will detect more complete results than the relate work.

4 Approach

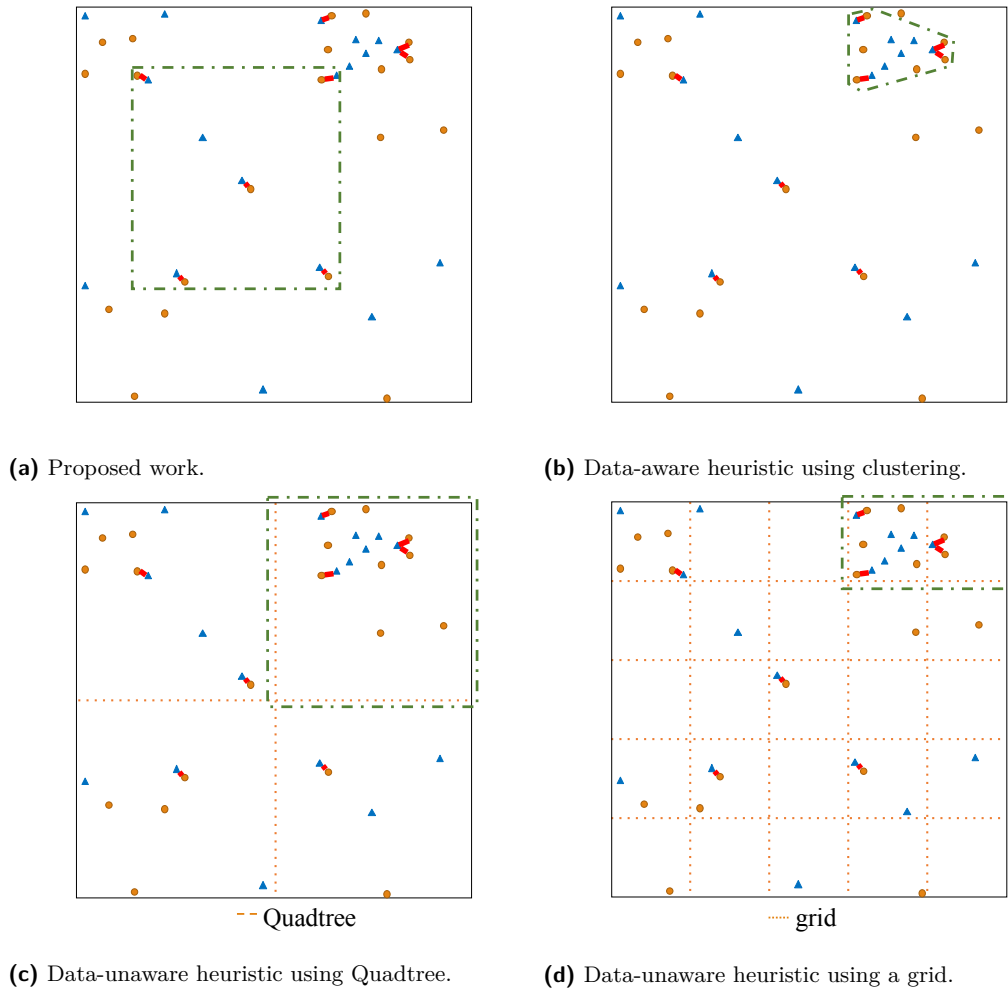
We begin this section by introducing a baseline algorithm for the LCPD problem. Then, we present two refinements: a Quadruplet (Quad) algorithm as well as a Quadruplet & Grid Filter-Refine (QGFR) algorithm, to reduce the computational cost without impairing correctness and completeness.

The pseudo-code of the general algorithm framework is shown in Algorithm 1. In this framework, all possible co-location patterns of the features associated with the input objects are enumerated in line 2-11. The instances of each co-location pattern are generated as the input of an MOBR-generating function `MOBRGenerator` (line 4), and the MOBRs obtained from this function are enumerated to detect the prevalence ones (line 4-10). Consider the dataset in Figure 1 as an example. In this case, F has two elements: f_A and f_B , so there is only one possible co-location pattern, $\{f_A, f_B\}$, whose 7 instances are saved in CI (line 3). The locality r is one of the MOBRs to be enumerated. There are 5 instances within it, and the participation index is $\frac{5}{6}$. Both metrics will be compared with the thresholds to determine whether $\langle \{f_A, f_B\}, r \rangle$ is an eligible result.

In this study, we focus on reducing the number of MOBRs enumerated for each co-location pattern (i.e., improving function `MOBRGENERATOR(·)`), but adopt Apriori-like algorithms to reduce the number of possible co-location patterns [9, 6], and the state-of-the-art algorithms to generate co-location instances [9, 14].

4.1 Baseline Algorithm

As already mentioned, we focus on localities defined as the MOBRs of subsets of co-location instances. In the function `MOBRGENERATOR(·)` of the baseline algorithm, we will enumerate all arbitrary subsets of the input co-location instances, and generate an MOBR for each of



■ **Figure 3** Comparison between related work. (Better in color.)

them. If each co-location pattern has n_{ci} instances on average, there will be $2^{n_{ci}}$ subsets, resulting in $2^{n_{ci}}$ MOBRs. Thus, the computational complexity of this baseline algorithm is $O(k2^{n_{ci}})$, where k is the number of possible co-location patterns.

4.2 Quad-Element Algorithm

Our first improvement is based on an MOBR enumeration lemma:

► **Lemma 2.** *Given a set s of n points in a two-dimensional plane, the set of MOBRs for arbitrary subsets of s is the same as the set of MOBRs for arbitrary subsets with cardinality ≤ 4 of s .*

Proof. Assume that there exists an MOBR for a subset (sub) with cardinality > 4 that is not an MOBR for a subset with cardinality ≤ 4 .

Let $x_{min}, x_{max}, y_{min}, y_{max}$ denote the minimum and maximum of the x, y coordinates of the points in sub . There must exist points $a, b, c,$ and d (which may be the same) in sub such that $x_a = x_{min}, x_b = x_{max}, y_c = y_{min}, y_d = y_{max}$. Thus, the MOBR for sub is the same as that for $\{a, b, c, d\}$, which is a subset of s with cardinality ≤ 4 , resulting in a contradiction with the assumption. ◀

Algorithm 1 General algorithm framework.

Require:

- Obj : A set of objects;
- R : A spatial relation over objects in Obj ;
- θ : Participation index threshold;
- γ : Co-location instance number threshold.

Ensure: Local co-location patterns with participation index $\geq \theta$ and the number of instances $\geq \gamma$.

```

1:  $F \leftarrow$  all spatial features in  $Obj$ ;
2: for all possible patterns  $C$  of  $F$  do
3:    $CI \leftarrow$  co-location instances of  $C$ ;
4:   for all  $mobr \in$  MOBRGENERATOR( $CI$ ) do
5:      $p \leftarrow$  the participation index of  $C$  in  $mobr$ ;
6:      $n \leftarrow$  the number of  $C$ 's instances in  $mobr$ ;
7:     if  $p \geq \theta$  and  $n \geq \gamma$  then
8:       Add  $\langle cp, mobr \rangle$  to the result.
9:     end if
10:  end for
11: end for

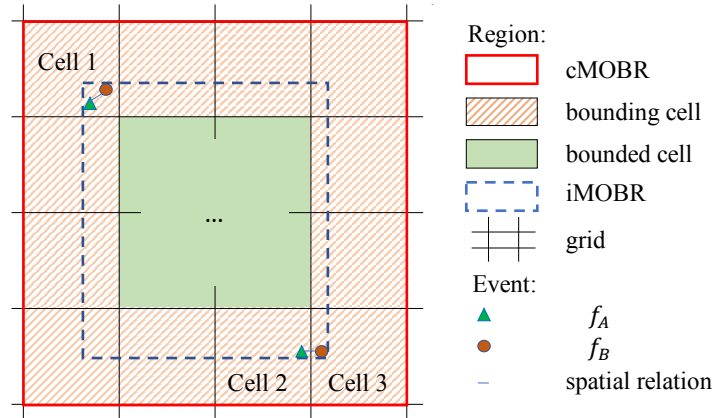
```

Lemma 2 indicates that the enumeration cost of a co-location pattern's MOBRs can be reduced from 2^n to n^4 without affecting completeness. By changing the function MOBRGENERATOR(\cdot) to generate the MOBRs of subsets with cardinality ≤ 4 of CI we can get the Quadruplet (Quad) algorithm with computational complexity of $O(kn_{ci}^4)$.

4.3 Quadruplet & Grid Filter-Refine Algorithm

Our definition of localities determines that a small displacement of any co-location instance that defines a locality's boundary will create a new locality, so there are lots of localities overlapping each other. If we can classify them into groups according to the areas they share, and apply a filter on each group instead of on individuals, the number of localities to be enumerated can be reduced further. Based on this idea, we proposed the second improvement: the Quadruplet & Grid Filter-Refine (QGFR) Algorithm.

The pseudo-code of the function MOBRGENERATOR(\cdot) in the QGFR algorithm is shown in Algorithm 2. Because a grid-based filter is applied, three new parameters are added, namely, a threshold of the participation index, a threshold of the number of co-location instances, and the cell size of the grid covering the entire study area. The first step of the function is saving the active cells of the input co-location pattern C (i.e., the cells overlapping C 's instances) in AC (line 2). A cell overlapping a co-location instance means that the intersection of the cell and the MOBR of this instance is nonempty. For example, Cells 1, 2, and 3 in Figure 4 are active cells of the pattern $\{f_A, f_B\}$. After getting the active cells, we will use their MOBRs (cMOBR) as an approximation of the MOBRs of C 's instances (iMOBR). The cells in a cMOBR are classified into two parts. The cells adjacent to the cMOBR's boundary are named as *bounding* cells, while the others are the *bounded* cells. In Figure 4, a cMOBR is delineated by a red solid rectangle, while its bounding and bounded cells are filled with a hash pattern and a solid color respectively. The boundary of each iMOBR has the following property:



■ **Figure 4** Grid cells and MOBRs (better in color).

► **Lemma 3.** *The boundary of any iMOBR must be within the bounding cells of one and only one cMOBR.*

The proof of this lemma is straightforward. If the boundary of an iMOBR is not within the bounding cells of a cMOBR, at least one of its four edges does not pass active cells, which is impossible. If two cMOBRs share the same bounding cells containing an iMOBR’s boundary, they must be the same. Therefore, we define that an iMOBR is in a cMOBR if its boundary is within the bounding cells of the cMOBR. For example, an iMOBR delineated by a dash rectangle in Figure 4 is in the plotted cMOBR. Because each iMOBR is in a unique cMOBR, by enumerating the iMOBRs in each cMOBR, we can enumerate all iMOBRs just once. In the pseudo-code, we enumerate all cMOBRs using Lemma 2 (line 3-10).

To eliminate the cMOBRs in which no iMOBR is eligible, we introduce an upper bound (MaxPI bound), $\eta(< C, \text{cMOBR} >)$, for the participation index of a local co-location pattern composed of a co-location pattern C and any iMOBR in a cMOBR of C . The MaxPI bound is based on an upper bound for the participation ratio, which can be stated as:

► **Lemma 4.** *The upper bound, $\zeta(< C, \text{cMOBR} >, f)$, for the participation ratio of a feature f in a local co-location pattern composed of a pattern C and any iMOBR in a cMOBR of C is*

$$\zeta(< C, \text{cMOBR} >, f) = \frac{po(C, f, \text{cMOBR})}{o(f, \text{bounded}) + po(C, f, \text{bounding})}$$

\forall iMOBR in cMOBR.

Table 1 describes the notation used in the above formula.

■ **Table 1** Symbols used in Lemma 4.

Number of objects of f in a locality r		
Participating in C	Not participating in C	All
$po(C, f, r)$	$npo(C, f, r)$	$o(f, r)$

where r can take values of “all cells” (cMOBR), “bounding cells” (bounding), or “bounded cells” (bounded) of the cMOBR, or the “actual iMOBR” (iMOBR), or the “intersection of iMOBR and bounding cells” (extra). The proof is as follows:

Proof.

$$\begin{aligned} pr(\langle C, iMOBR \rangle, f) &= \frac{po(f, C, iMOBR)}{o(f, iMOBR)} = \frac{po(f, C, bounded) + po(f, C, extra)}{o(f, bounded) + o(f, extra)} \\ &= \frac{po(f, C, bounded) + po(f, C, extra)}{o(f, bounded) + po(f, C, extra) + npo(f, C, extra)}. \end{aligned}$$

Because $npo(f, C, extra) \geq 0$,

$$pr(\langle C, iMOBR \rangle, f) \leq \frac{po(f, C, bounded) + po(f, C, extra)}{o(f, bounded) + po(f, C, extra)}.$$

Because $extra \in bounding$, $0 \leq po(f, C, extra) \leq po(f, C, bounding)$. Meanwhile, $\frac{po(f, C, bounded)}{o(f, bounded)} \leq 1$. Thus,

$$\begin{aligned} pr(\langle C, iMOBR \rangle, f) &\leq \frac{po(f, C, bounded) + po(f, C, bounding)}{o(f, bounded) + po(f, C, bounding)} \\ &= \frac{po(f, C, cMOBR)}{o(f, bounded) + po(f, C, bounding)}. \end{aligned} \quad \blacktriangleleft$$

Based on the definition of the participation index, we can define the MaxPI bound as the smallest upper bound of the participation ratio of any feature in the local co-location pattern, i.e.,

$$\eta(\langle C, cMOBR \rangle) = \min_{f_i \in C} (\zeta(\langle C, cMOBR \rangle, f_i)).$$

Given a participation index threshold θ , if $\eta(\langle C, cMOBR \rangle) < \theta$, there will not be any eligible iMOBR in this cMOBR. In the pseudo-code, the MaxPI bound of C in every one of its cMOBRs, together with the number of instances, is compared with the thresholds to determine whether enumerating the iMOBRs in the current cMOBR is necessary.

Algorithm 2 Function MOBRGenerator in QGFR algorithm.

Require:

- CI : A set of instances of a co-location pattern C ;
- θ : Participation index threshold;
- γ : Co-location instance number threshold;
- l : The size of each grid cell.

Ensure: MOBRs of CI 's subsets.

```

1: function MOBRGENERATOR( $CI, \theta, \gamma, l$ )
2:    $AC \leftarrow$  active cells of  $C$ ;
3:   for all  $subAC$  (with cardinality  $\leq 4$ )  $\subseteq AC$  do
4:      $cmobr \leftarrow$  the MOBR of  $subAC$ ;
5:      $\eta \leftarrow$  MAXPI( $C, cmobr$ );
6:      $n \leftarrow$  the number of  $C$ 's instances in  $cmobr$ ;
7:     if  $\eta \geq \theta$  and  $n \geq \gamma$  then
8:       Add iMOBRs in  $cmobr$  to the result.
9:     end if
10:  end for
11: end function

```

Assuming that each co-location pattern has n_{ac} active cells on average, and the number of iMOBRs in each cMOBR is q , the computational complexity is $O(kn_{ac}^4 q)$. If q can be

■ **Table 2** Parameters for the experiments.

Symbol	Meaning
n_{cp}	Number of core co-location patterns
n_{cc}	Core co-location patterns' cardinality
n_{ci}	Number of instances of each pattern
n_i	Number of input objects
n_f	Number of input features
Grid size	Cell's edge length of the grid used in the QGRF algorithm

treated as a constant, because n_{ac} is much less than n_{ci} in most cases, the computational cost of the QGRF algorithm is much lower than that of the Quad. Because we have proved that in this algorithm all MOBRs of co-location instances are evaluated once and only eligible results are returned, we maintain the correctness and completeness of the algorithm through the performance improvement.

5 Experimental Evaluation and Case Studies

In this section, we evaluate the baseline, Quad, and QGRF algorithm using synthetic data and a Chicago crime dataset [4], followed by one case study using the North American Atlas - Hydrography dataset from the U.S. Geological Survey [11] and the dataset of the U.S. major cities from Esri.

5.1 Experiments

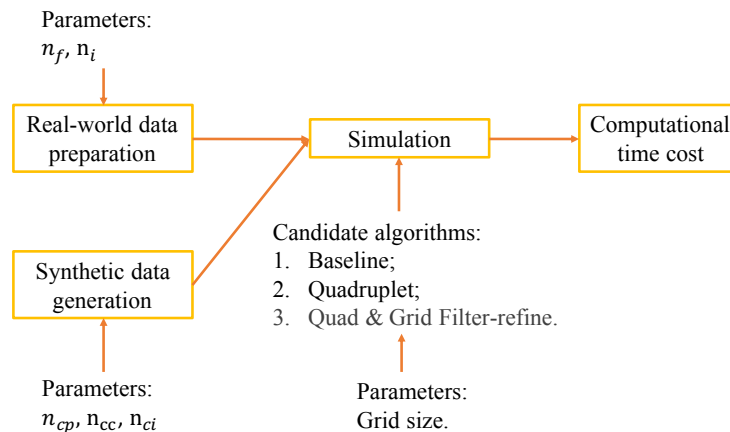
The goal of the experiments was twofold: (a) evaluate the effect of the performance refinements of the proposed Quad algorithm and QGRF algorithm compared with the baseline algorithm. (b) determine the robustness of the QGRF algorithm given different inputs.

According to our analysis in §4, the computational complexity of the three algorithms are $O(k2^{n_{ci}})$, $O(kn_{ci}^4)$, and $O(kn_{ac}^4q)$ respectively, where n_{ci} is the number of co-location instances per pattern, n_{ac} is the number of active cells per pattern, k is the number of co-location patterns, and q is the average number of iMOBR in each cMOBR. To evaluate the performance refinements, we studied the following two questions: (1) What is the effect of the number of co-location instances? (2) What is the effect of the number of co-location patterns? To determine the robustness, we asked how well the QGRF algorithm performed under different size of grid cells.

To answer these questions, we designed experiments as shown in Figure 5. The synthetic and the real-world data (a Chicago crime dataset) were generated with controlled parameters. In the simulation, three algorithms were executed with the grid cell size as a parameter. The performance was evaluated and compared using the run time of each algorithm. The platform for the simulation was Microsoft .NET Framework 4.5 on a computer with Intel(R) Core(TM) i7-4770 3.40 GHz CPU and 32 GB RAM. The parameters in the experiments are shown in Table 2.

5.1.1 Synthetic data generation

A point distribution with co-location patterns is often modeled as an aggregated point process [9, 2, 6]. Commonly used point processes include the Poisson cluster process [1] and Matérn's cluster process [5]. In order to ensure the existence of local co-location patterns, we made two changes on the steps used in [2], including:



■ **Figure 5** Experiment design.

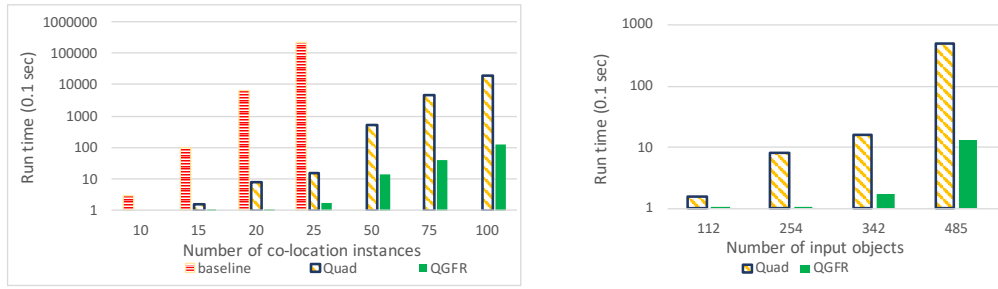
- Randomly select a rectangular region in the study area as a prevalence locality for each co-location pattern.
- In each co-location pattern's prevalence locality, ensure that at least 4 instances of the pattern are generated, and that no noise object of the features in the pattern is generated.

Because the subsets of a co-location pattern are also co-location patterns, when generating the synthetic data, we named the patterns which were not subsets of other patterns core patterns. The study area size was set to 10000×10000 . The spatial relation was a neighborhood with a radius of 10. The number of noise objects of each feature was set to $4 \times n_{ci}$.

5.1.2 Experimental results

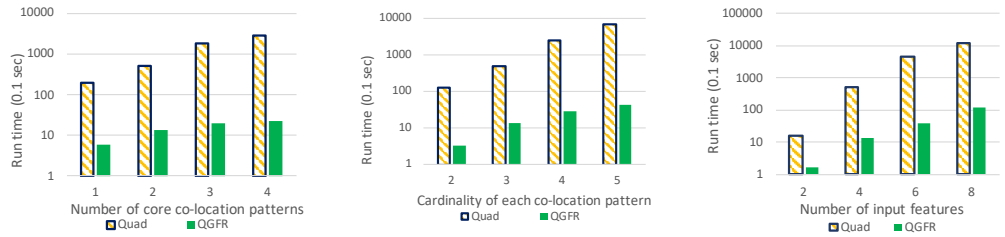
Effect of the number of co-location instances. The experiments were conducted with both synthetic and real-world data. The synthetic data was generated by fixing $n_{cp} = 2$ and $n_{cc} = 3$, but changing n_{ci} , whose results were shown in Figure 6a. The computational cost of the baseline algorithm, as expected, increased exponentially with n_{ci} , and was much larger than that of the two proposed algorithms, so its run time was not included when $n_{ci} = 50, 75$, or 100 . The run time of the Quad algorithm was much longer than that of the QGFR algorithm, and it also increased faster than the latter with increasing n_{ci} . The experiment with the Chicago crime dataset was conducted by fixing $n_f = 3$ but varying n_i . By increasing the number of input objects in a fixed study area, we increased the number co-location instances indirectly. The results (Figure 6b) also shown that the advantage of the QGFR algorithm increased as the number of input objects grew.

Effect of the number of co-location patterns. Since the number of co-location patterns is determined by both the number of core co-location patterns and their cardinalities, we conducted two controlled experiments with synthetic data and one with the Chicago crime dataset on them. Figure 7a and Figure 7b presented the results of experiments with the synthetic data. In Figure 7a $n_{cc} = 3$ and $n_{ci} = 50$ but n_{cp} changed, while in Figure 7b $n_{cp} = 2$ and $n_{ci} = 50$ but n_{cc} changed. Figure 7c shown the results using the real-world data, where the number co-location pattern was increased by increasing the number of input features. In all the cases, the growing number of co-location patterns increased the advantage of the QGFR algorithm over the Quad algorithm.



(a) Results with synthetic data. (b) Results with the Chicago crime dataset.

Figure 6 Effect of the number of co-location instances.



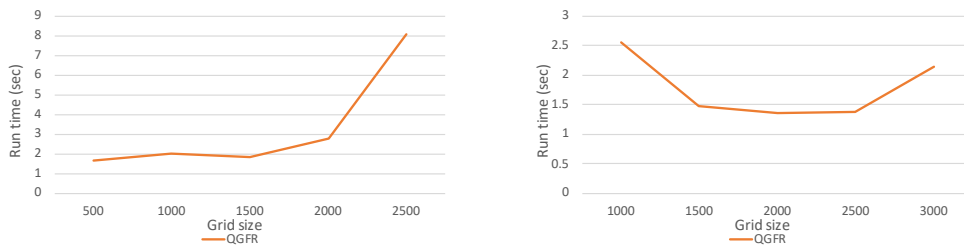
(a) Effect of the number of core co-location patterns. (b) Effect of each co-location pattern's cardinality. (c) Effect of the number of input features.

Figure 7 Effect of the number of co-location patterns

Effect of the size of grid cells. The sensitivity analysis was done through two controlled experiments where the same synthetic and real-world data but different grid cell size were used. The parameters for the synthetic data were $n_{cp} = 2, n_{cc} = 3, n_{ci} = 50$ and those for real-world data were $n_i = 485, n_f = 4$. According to the results shown in Figure 8, the QGFR algorithm was robust with changes in the grid cell size, since the fluctuation of its run time was small when the grid cell size changed. When the grid cell size was small, the number of active cells was not much smaller than the number of co-location instances, so the performance would be improved if a larger cell size was used. As the grid cell size increased, more iMOBRs resided in a single grid cell, so the performance improvement brought about by the MaxPI bound was weakened.

5.2 Case Study using North American Atlas-Hydrography and U.S. Major City Datasets

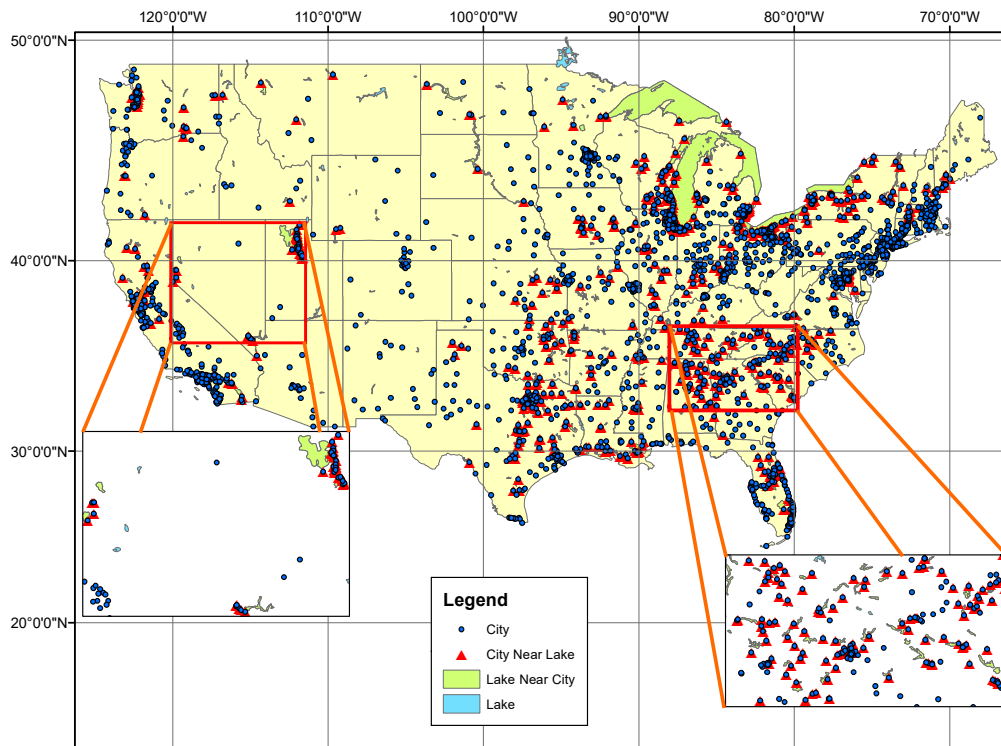
We conducted a case study using the North American Atlas - Hydrography dataset from the U.S. Geological Survey and the data of the U.S. major cities from Esri. Other inputs included a spatial relation specified by a neighborhood radius of 50 kilometers, a participation index threshold $\theta = 0.6$, and an instance number threshold $\gamma = 20$. There were 2610 cities which represent cities in the U.S. with population of more than 10 thousand in the dataset. The number of lakes was 394. The participation index of the co-location pattern $\{city, lake\}$ was 0.33, which meant major cities were not globally co-located with lakes in the U.S. However, our proposed QGFR algorithm detected some prevalence localities, two of which were shown in Figure 9 with the zoom-in maps. In the east locality, there were 163 cities, 109 of which were co-located with lakes, while 39 out of 41 lakes were near cities, so the participation index was about 0.67. This locality could be detected by the related work as well, because if



(a) Results with synthetic data.

(b) Results with the Chicago crime dataset.

■ **Figure 8** Effect of the size of grid cells.



■ **Figure 9** Case study with the hydrography and city data. Two prevalence localities of co-location pattern $\{city, lake\}$ are delineated by rectangles and shown in the zoom-in maps. (Better with color.)

we defined the density as the number of instances of a feature in a unit area, the density of both input objects and co-location instances was high (the ratio between the density of the co-location instances in the locality and that in the whole country was about 4.22). Contrarily, in the west locality, there were 35 out of 50 cities co-located with 7 out of 11 lakes, resulting in the participation index as about 0.63. In this locality, the density of the input objects and co-location instances was almost the same as that in the whole country (the ratio between the density of the co-location instances in the locality and that in the whole country was about 1.03), which meant that the locality could not be identified by the related work using clustering to define localities. The findings indicated that the co-location

pattern of major cities and lakes existed not only in the southeast of the U.S where lakes concentrated but also in the west where it was drier and lakes were more valuable of the cities.

6 Conclusion and Future Work

In this paper, we formally defined the local co-location pattern detection problem, and proposed two algorithms that can efficiently solve it. The effectiveness and efficiency of the algorithms were proved theoretically and validated experimentally on synthetic and real datasets. In addition, we presented the results of one case study using the North American Atlas-Hydrography and U.S. Major City Datasets.

During the study, we noticed that the distribution of spatial events (e.g., the auto-correlation between events of the same feature) may affect the results. Our future research will take this into consideration. In addition, the distribution of events related to humans may be strongly affected by road networks especially in urban areas. Defining regions as subsets of road networks may result in richer and more meaningful results. We plan to explore this idea in our future work.

References

- 1 Adrian Baddeley. Spatial Point Processes and their Applications. In *Stochastic Geometry*, volume 1892 of *Lecture Notes in Mathematics*, pages 1–75. Springer, Berlin, Heidelberg, 2007. doi:10.1007/978-3-540-38175-4_1.
- 2 S. Barua and J. Sander. Mining Statistically Significant Co-location and Segregation Patterns. *IEEE Transactions on Knowledge and Data Engineering*, 26(5):1185–1199, 2014. doi:10.1109/TKDE.2013.88.
- 3 Mete Celik, James M. Kang, and Shashi Shekhar. Zonal co-location pattern discovery with dynamic parameters. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 433–438. IEEE, 2007. URL: <http://ieeexplore.ieee.org/abstract/document/4470269/>.
- 4 Chicago Police Department. Crimes - 2001 to present, 2017. [Online; accessed 30-September-2017]. URL: <https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2>.
- 5 Sung Nok Chiu, Dietrich Stoyan, Wilfrid S. Kendall, and Joseph Mecke. *Stochastic Geometry and Its Applications*. John Wiley & Sons, 2013.
- 6 Min Deng, Jiannan Cai, Qiliang Liu, Zhanjun He, and Jianbo Tang. Multi-level method for discovery of regional co-location patterns. *International Journal of Geographical Information Science*, 31(9):1846–1870, 2017. doi:10.1080/13658816.2017.1334890.
- 7 Christoph F. Eick, Rachana Parmar, Wei Ding, Tomasz F. Stepinski, and Jean-Philippe Nicot. Finding Regional Co-location Patterns for Sets of Continuous Variables in Spatial Datasets. In *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '08, pages 30:1–30:10, New York, NY, USA, 2008. ACM. doi:10.1145/1463434.1463472.
- 8 Guilbert Gates, Jack Ewing, Karl Russell, and Derek Watkins. How Volkswagen's 'Defeat Devices' Worked. *The New York Times*, 2015. URL: <https://www.nytimes.com/interactive/2015/business/international/vw-diesel-emissions-scandal-explained.html>.
- 9 Yan Huang, Shashi Shekhar, and Hui Xiong. Discovering colocation patterns from spatial data sets: a general approach. *IEEE Transactions on Knowledge and Data Engineer-*

- ing*, 16(12):1472–1485, 2004. URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1350759.
- 10 Pradeep Mohan, Shashi Shekhar, James A. Shine, James P. Rogers, Zhe Jiang, and Nicole Wayant. A Neighborhood Graph Based Approach to Regional Co-location Pattern Discovery: A Summary of Results. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, pages 122–132, New York, NY, USA, 2011. ACM. doi:10.1145/2093973.2093991.
 - 11 USGS. North america rivers and lakes, 2018. [Online; accessed 13-February-2018]. URL: <https://www.sciencebase.gov/catalog/item/4fb55df0e4b04cb937751e02>.
 - 12 Song Wang, Yan Huang, and Xiaoyang Sean Wang. Regional Co-locations of Arbitrary Shapes. In *Advances in Spatial and Temporal Databases*, pages 19–37. Springer Berlin Heidelberg, 2013. doi:10.1007/978-3-642-40235-7_2.
 - 13 Jordan Wood. Minimum Bounding Rectangle. In Shashi Shekhar, Hui Xiong, and Xun Zhou, editors, *Encyclopedia of GIS*, pages 1232–1233. Springer International Publishing, 2 edition, 2017. doi:10.1007/978-3-319-17885-1_783.
 - 14 Jin Soung Yoo and S. Shekhar. A Joinless Approach for Mining Spatial Colocation Patterns. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1323–1337, oct 2006. doi:10.1109/TKDE.2006.150.

Detection and Localization of Traffic Signals with GPS Floating Car Data and Random Forest

Yann Méneroux

Univ. Paris-Est, LASTIG COGIT, IGN, ENSG, Saint-Mandé, France
yann.meneroux@ign.fr

Hiroshi Kanasugi

CSIS, Institute of Industrial Sciences, The University of Tokyo, Japan

Guillaume Saint Pierre

Centre for Studies and Expertise on Risks, Mobility, Land Planning and the Environment (Cerema), Toulouse, France

Arnaud Le Guilcher

Univ. Paris-Est, LASTIG COGIT, IGN, ENSG, Saint-Mandé, France

Sébastien Mustière

Univ. Paris-Est, LASTIG COGIT, IGN, ENSG, Saint-Mandé, France

Ryosuke Shibasaki

CSIS, Institute of Industrial Sciences, The University of Tokyo, Japan

Yugo Kato

Transport Consulting Division, NAVITIME JAPAN Co., Ltd

Abstract

As Floating Car Data are becoming increasingly available, in recent years many research works focused on leveraging them to infer road map geometry, topology and attributes. In this paper, we present an algorithm, relying on supervised learning to detect and localize traffic signals based on the spatial distribution of vehicle stop points. Our main contribution is to provide a single framework to address both problems. The proposed method has been experimented with a one-month dataset of real-world GPS traces, collected on the road network of Mitaka (Japan). The results show that this method provides accurate results in terms of localization and performs advantageously compared to the *OpenStreetMap* database in exhaustivity. Among many potential applications, the output predictions may be used as a prior map and/or combined with other sources of data to guide autonomous vehicles.

2012 ACM Subject Classification Computing methodologies → Machine learning, Information systems → Global positioning systems, Information systems → Data mining

Keywords and phrases Map Inference, Machine Learning, GPS Traces, Traffic Signal

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.11

Acknowledgements We would like to express our gratitude to Paul Chapron, Prof. Harutoshi Yamada and Prof. Yukio Sadahiro for their precious advice and constructive criticism.

1 Introduction

As one of the main supports for citizen mobility, roads are deservedly considered as a major cartographic theme in maps. Therefore, it is not surprising that most national mapping agencies allocate considerable amount of resources to keep road network databases as detailed,



© Yann Méneroux, Hiroshi Kanasugi, Guillaume Saint Pierre, Arnaud Le Guilcher, Sébastien Mustière, Ryosuke Shibasaki, and Yugo Kato;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 11; pp. 11:1–11:15



Leibniz International Proceedings in Informatics

LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

accurate and up-to-date as possible [14, 4]. This is generally done by stereorestitution on aerial orthoimages [17], completed with field surveys to get details that cannot be captured in the images. Recently, automatic detection of roads has dramatically improved, especially when combined with machine learning algorithms [28], and now achieves very good performance even on satellite images. However, if the whole process tends to get less expensive and less time-consuming, it still suffers from a major drawback: road map timeliness is inevitably limited by the frequency of aerial image release [7].

Nowadays, with the spread of connected terminal devices equipped with a Global Positioning System (GPS) receiver, an increasing number of vehicle trajectories are becoming available. *Map inference*, which aims at leveraging this new source of data to extract geographic information [3], is becoming popular and tends to complement, if not completely replace, traditional survey techniques. Among their main assets, GPS traces are recorded on a daily basis, which allows for short-delay update capabilities. Indeed, aerial picture campaigns are typically conducted every several years, notwithstanding an additional delay for image preprocessing and orthorectification. This substantial delay might be critical in applications relying on highly up-to-date reference networks, such as emergency routing or disaster mitigation.

Contrarily, with GPS traces analysis, modifications are potentially detectable as soon as enough traces are recorded on a suspicious point to ensure the statistical robustness of the notification. Ultimately, with connected devices, it is foreseeable that data will be recorded and processed by online algorithms, resulting in a much more reactive system that is capable of detecting ephemeral events (e.g., road works, detour or accidents) in quasi real-time. Moreover, data can be continuously recorded while drivers are commuting for example, which makes this solution much less expensive than aerial campaigns and field surveys. More anecdotally, since we may assume that for any consistent algorithm, the estimation is getting closer to the reality as the number of traces increases, the dataset sampling itself serves a logic of public utility: the most important itineraries are the most traveled, therefore those where road map inference is the most reliable.

Initially restricted to the construction of road geometry and topology, map inference is now getting attention for enriching pre-existing networks with attributes (number of lanes, speed limitations...) or infrastructure (traffic signals, speed bumps, bus stops...) [24, 18]. Most of these features are not accessible through aerial images, and utilizing GPS traces seems unavoidable. Moreover, aerial images may not be accessible in developing countries, or available only at prohibitive cost. Instead, access to data stemming from local fleets or collaborative transport smartphone applications, are producing large sets of GPS traces. This surrogate source of data may be used with map inference techniques to provide a cheap alternative solution for map construction.

An exhaustive and detailed knowledge of road infrastructure is a prerequisite to many applications. For example, autonomous cars are expected to appear on the market in the near future. Reliability and robustness of the information used by such vehicles to make decisions is a big concern. It is usually more reliable to know in advance the location and the type of object that should be detected and confirm detection with embedded sensors. Additionally, driving-assistance devices conception, road safety, eco-driving, urban traffic flow simulation or even accurate routing time computation are as many other examples of fields or applications where the knowledge of a road network needs to be completed with attributes and infrastructure [4, 26, 1].

In parallel, machine learning techniques are becoming all-pervasive in fields requiring to process a large amount of data, or simply when theoretical background is insufficient to build reliable predictive models. With this kind of approach, expert knowledge is no longer

required, and algorithms are trained on labeled data. However, machine learning is a relevant solution only if the two following conditions are met: firstly we must have an extensive and representative training dataset, and secondly, we must have a natural definition of cost that quantifies how close the generated road map is compared to the training data ground truth. A few years ago, some authors such as Liu *et al.* [14], have introduced numerical measures to assess the quality of maps produced by GPS traces, hence opening the way for a full machine learning resolution of the problem [3]. In this vein, Zhang and Sester [27] combined fuzzy logic and k-means clustering for incrementally inferring maps, while Fathi and Krumm [10] proposed to train an Adaboost classifier to recognize road intersections, based on the probability density function of trace headings. Similarly, Van Winden *et al.* [25] found that Support Vector Machines (SVM) and regression trees are the most adequate algorithms for speed limit inference. In some more sophisticated algorithms, traces are combined with external sources of data to get better results, for example in [12] where Twitter data and SVM are used for an automatic mining of street names. We believe that statistical learning is especially adapted to this problem, and that it guarantees the portability of the approach to other cities, countries and environments.

Among traffic control devices, traffic signals are unarguably the most effective to regulate jammed intersections [23]. They have a crucial impact both on traffic flow at the city scale and on the perceptions of individual drivers. Surprisingly, very few research works address the problem of utilizing a collection of probe vehicle traces coupled with machine learning algorithms to detect traffic signals. The most related research work is certainly the one of M. Munoz-Organero *et al.* [19], who used machine learning algorithms to detect in real-time several kinds of road infrastructures, based on an analysis of speed and acceleration signals, estimated from GPS positions. Despite providing very good results, the performance scores clearly exhibits some limitation on traffic signal detection, compared to the cases of street crossings, and roundabouts. Besides when the only source of measurement is a GPS receiver, speed-based analysis is only possible provided that the GPS positioning is accurate enough (for example if equipped with a Doppler speed measurement, when used in differential mode, or in open areas) and sampling frequency is high (over 1 point per second or so). Furthermore, a natural extension of [19] would be to use all vehicles which traveled at a specific location to detect infrastructure. In this work, we propose a method to detect and then localize traffic signals through a random forest classification and regression using the spatial distribution of stop points along the road.

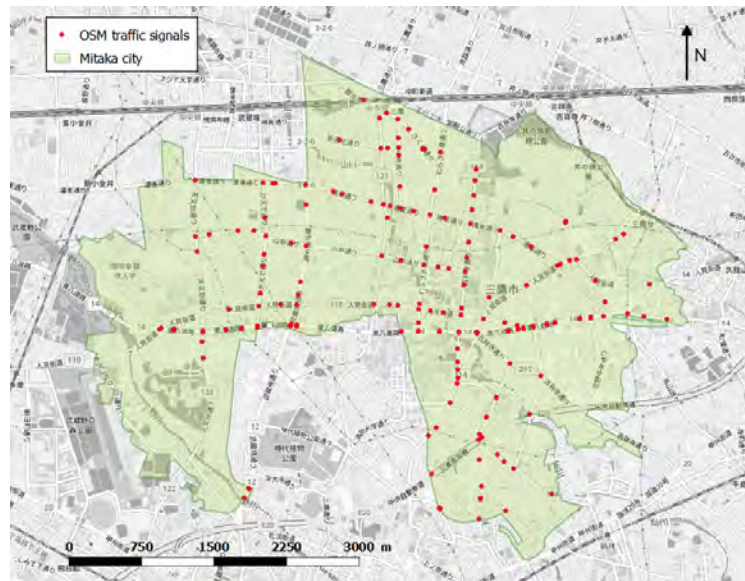
We must notice that localization is an important aspect of the problem. Even though we know that an intersection is controlled by a system of traffic lights, the positions of stop lines on each individual streets remain uncertain, and this is especially true since road network abstraction and generalization may introduce an additional component of uncertainty.

The remaining of the paper is structured as follows: the dataset and its preparation are briefly described in the next section, while our methodology to create instances, train and validate the model is detailed in section 3. Section 4 provides the results and discusses them. Eventually, section 5 concludes the paper.

2 Data and preparation

2.1 Study area

The experimentation was conducted in Mitaka (Japan), a commuter town located approximately 20 km west of central Tokyo, and covering an extent of 16 square kilometers. This choice was motivated by the fact that Mitaka contains a wide variety of urban aspects, ranging



■ **Figure 1** Mitaka city and OpenStreetMap traffic signal database.

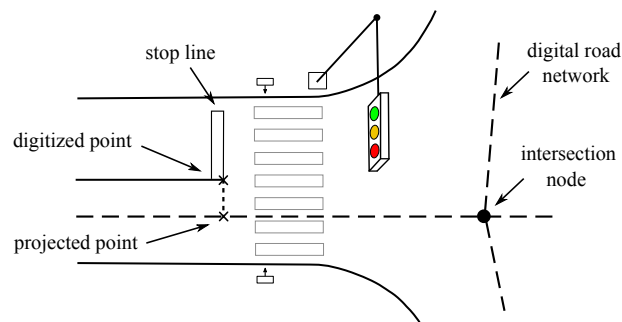
from dense downtown to inter-urban residential districts, including motorway environments and parks as well. Mitaka city is illustrated on figure 1, where traffic signal controlled intersections are depicted in red.

We extracted a routable road map from the national reference. The topological graph of a road map is often organized in such a way that a node is always located close to each traffic signal, even when no physical intersection is involved (e.g., traffic signal associated to pedestrian crossing in the middle of a road link). For this reason, we decided to remove degree-2 nodes, so that it may practically be assumed that digital road network has been created without any knowledge of traffic signal locations.

2.2 Ground truth data acquisition

As an application of machine learning, it is necessary to collect ground truth data, namely the positions of all traffic signals in Mitaka, to train and then validate the algorithm. Throughout this paper, a *stop line* is defined as the position along the road, where the front vehicle in queue is expected to stop while waiting for the signal to turn green.

We started from a base reference extracted from *OpenStreetMap* (figure 1). This source of data is not complete, and each point corresponds to an entire crossing controlled by a system of traffic lights, but no information is provided regarding the number of streets actually controlled by an individual signal, nor are the positions of stop lines on these streets. Using OSM basis and multiple sources of orthoimages (produced at different dates), positions of stop lines have been manually digitized, and then orthogonally projected onto the road network, as depicted here after on figure 2. At the end of this process, a total of 669 stop line positions have been digitized, which corresponds to 253 crossings controlled by traffic signals. Out of them, 177 (70%) are reported in OSM database. For each stop line, we also recorded a binary attribute to indicate which direction of flow is subject to stop at the traffic signal. It takes the value 0 if the stop line is directed to vehicles traveling from source node to target node, otherwise it is set equal to 1 (source and target node is arbitrarily defined by the road network database provider).



■ **Figure 2** Ground truth data acquisition on orthoimages and reference road network.

Eventually, since orthoimages might suffer from local distortions, we had to check that our ground truth dataset is accurate enough for our application. A positional accuracy control was carried out by uniformly sampling 30 stop lines at random and surveying them with a single phase low-cost GPS receiver [16]. This operation enabled to guarantee (with 95% confidence index) that stop line positions have been digitized with a sub-meter accuracy (root mean square error below 90 cm).

2.3 Floating Car Data

For this experimentation, we used GPS Floating Car Data (FCD) provided by NAVITIME JAPAN¹, a private company developing navigation technologies and providing various kinds of web application services such as route navigation, travel guidance, and other useful information services for moving people.

The sample dataset is covering the entire Japan and has been recorded over a one-month time span, in October 2015. Pedestrian trajectories have been priorly removed so that it contains only vehicle navigation data. We extracted all GPS records intersecting the Mitaka polygon shape. Each record (nominally one per second) contains the following entries: a user identification number, a route identification number, geographic coordinates (in decimal degrees) and a timestamp. A route is a set of records on an individual subtrip (i.e. during a GPS receiver session). Due to privacy issues, driver identification number is modified every day at midnight. Entries containing -1 in timestamp or coordinates (i.e. about 2% of records, corresponding to GPS signal lost or logging failure) have been removed. Coordinates (as well as network and traffic signal ground truth) have been converted into UTM 54N cartographic projection system. For convenience purposes, we also transformed timestamps into an integer number of epochs. This made computing traveled distances and elapsed time between records much easier.

Similarly to most studies related to GPS probe vehicles, map-matching, which consists in reconstructing the path traveled by a vehicle on a network, is an important pre-processing step and has two, possibly combined, main advantages: providing a mapping function between GPS positions and network links (which is necessary in our application case for updating the network) and enhancing positional accuracy. The latter is particularly important in urban environment, where GPS satellite signal is likely to be partially impeded by buildings. We used an algorithm based on Hidden Markov Models, developed by Newson and Krumm [20].

¹ <http://corporate.navitime.co.jp/en>

Since all traces are located on the same area, it is worthwhile to compute shortest path distances between every couple of nodes just once, then storing results in a look-up table, before map-matching all trajectories in a batch. Following this approach enabled to speed-up the process, and reach a pace of 10 traces map-matched per second (approximately 1500 faster than the naive solution requiring to process shortest paths online). However, for a road network containing a number n of nodes, since the time and space complexities of the look-up table computation are growing like $\mathcal{O}(n^2)$, it inevitably becomes necessary to find alternative solutions when the area of interest is large. One of them might be to use sparse matrix notation with hashtable data structure, and save only distances which are shorter than a predefined threshold.

Root mean square error of displacements induced by map-matching is equal to 8.3 m, which gives some insight regarding the average quality of GPS receivers. Overall, 99% of records have been map-matched (excluding outlier points). Eventually, we removed all traces map-matched with Chūō expressway, which runs the south-eastern part of Mitaka and, needless to say, does not contain any traffic light.

At the end of the pre-processing phase, a total of 11870 traces are remaining in the dataset, which represents slightly above 7 million records, about 42000 km and 3122 hours of driving data. The median trip runs 3 km and lasts 10 minutes. 95% of the dataset is recorded at a frequency higher than 0.2 Hz.

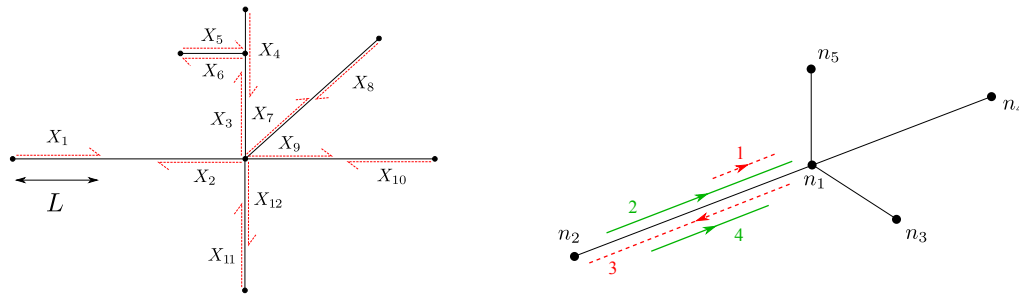
3 Methodology

In this section, we describe our methodology to build training and validation instances from GPS trajectories, then after a brief review of Random Forest algorithm, we present an extension to aggregate individual predictions, and infer the presence of traffic signals at the level of crossings.

3.1 Instance computation

In most machine learning problems, there is a natural definition of an instance. For example, in image recognition tasks, each individual image is an instance, and we may easily assume that they are independent to each other. In our application case, there is no such definition, since we are looking for objects located at unknown positions on a topological network. However, considering that most traffic signals are located near intersections, we decided to compute instances based on road segments starting from nodes. This choice was motivated by the fact that it results in mutually independent instances, hence facilitating split process into training and validation datasets. In turn, our algorithm will inevitably fail to detect traffic signals located far from road intersections. Since, it may be assumed that this represents a small proportion of all traffic signals, we believe that this choice would not have too much negative impact. Note that, as depicted on figure 3, each network edge is generating two instances (one starting from each node). Hence the total number of instances generated equals at most twice the number of edges in the road network (in fact, some of them might be empty of traces, consequently the actual number of instances is generally smaller). We will refer to this segment as a *frame* hereafter.

In order to get homogenous instances, frames have been set to a fixed length L . If an edge is longer than L , then only a portion of length L (starting from the node) is considered. On the opposite, if it is shorter than L , the frame is padded with zeroes (X_5 and X_6 on figure 3). The numerical value of L was set to 100 m, since there is no evidence to think that events located further than 100 m from a traffic signal, might be of any help for the detection.



■ **Figure 3** Left: frames generation (red dashed arrows) on the road network. Each frame is computed based on GPS traces moving towards the intersection node (i.e. in the opposite direction of the arrows). Right: selection of traces (see text for details).

We are interested in vehicles moving towards the intersection node, then only GPS traces *globally* moving towards the node are added up to the frame. More formally, the last record of the trace on the edge must be located closer from the intersection node than the first record (with respect to a distance metrics computed as a curvilinear abscissa along the edge geometry). Additionally, we required that the distance between both these extremal records is at least half of the edge length. For example, on the right part of figure 3, only traces 2 and 4 (solid lines) are taken into account in the frame generated from intersection n_1 (trace 1 is too short, while trace 3 is moving in the opposite direction). For the instance generated from node n_2 , traces 1, 2 and 4 are discarded. Once traces moving towards a given node have been identified, we can extract sequences of GPS records corresponding to vehicle stops.

► **Definition 1** (Stop sequence). Given a sequence of timestamped GPS points and two parameters: a maximal speed value $v_{max} \in \mathbb{R}^+$ and a minimal time duration $\tau_{min} \in \mathbb{R}^{+*}$, we define a *stop sequence* as a sub-sequence of consecutive records $S = \{(x_i, t_i) \mid p \leq i \leq q\}$ verifying the two following inequalities:

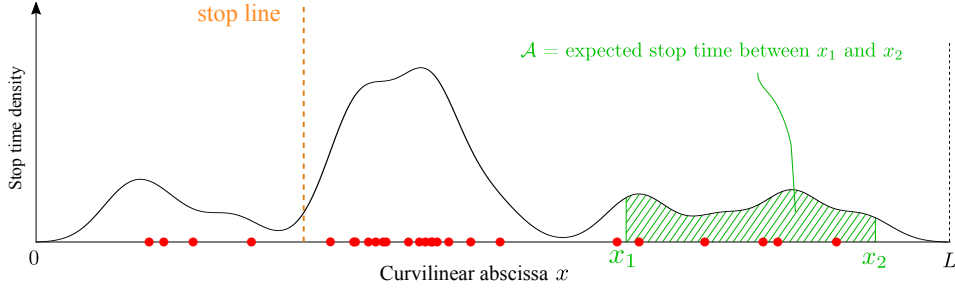
$$t_q - t_p \geq \tau_{min} \quad \text{and} \quad \forall i \in \llbracket p, q - 1 \rrbracket \quad \frac{|x_{i+1} - x_i|}{t_{i+1} - t_i} \leq v_{max}$$

where x is the curvilinear abscissa of GPS records along the edge. Simply put, for being qualified as a stop sequence, a portion of trajectory must be slow enough for a sufficiently long period of time. Also, note that p and q must be chosen in such a way that it is impossible to add new records to the sequence without breaking the inequalities stated above.

► **Definition 2** (Stop point). For a given stop sequence, a *stop point* is defined as the mean position of points in the sequence, associated with the total duration of stop.

For each instance, stop points have been extracted from the selected traces according to definitions 1 and 2 with the following parameters: $v_{max} = 0.5 \text{ m.s}^{-1}$ and $\tau_{min} = 5$ seconds.

Since the number of stop points is unpredictable, it is not a reasonable solution to train a classifier with a predefined number of stop points. Indeed, this solution would fatally imply that no prediction can be made on instances with too few stop points (for example in remote parts of the road map). Reversely, if too many stop points occurred on a given instance, there is no alternative but randomly selecting the appropriate number of stops to make it fit the model of classifier. A practical solution to this issue, is to estimate the distribution of stop durations along the road curvilinear abscissa with an adapted version of the kernel distribution estimation (KDE) method [22].



■ **Figure 4** Weighted kernel density estimation of stop points. Orange vertical dashed line stands for the position of the stop line associated to a traffic signal (controlling the entrance on an intersection located on the left of the graphics). Vehicles are moving from the right to the left.

Let K be a positive, real-valued symmetric function whose integral sums up to 1. Function K is called a kernel. Let $x_i \in [0, L]$ be a set of N stop point locations, associated to stop duration times $t_i \in \mathbb{N}$ (for reasons that will be detailed further, we assume that timestamps are precise up to the second, which means that stop durations may be considered as integers). We define the *weighted kernel density estimation* as :

$$\forall x \in [0, L] : \quad \hat{f}_h(x) = \frac{1}{Nh} \sum_{i=1}^N t_i K\left(\frac{x - x_i}{h}\right)$$

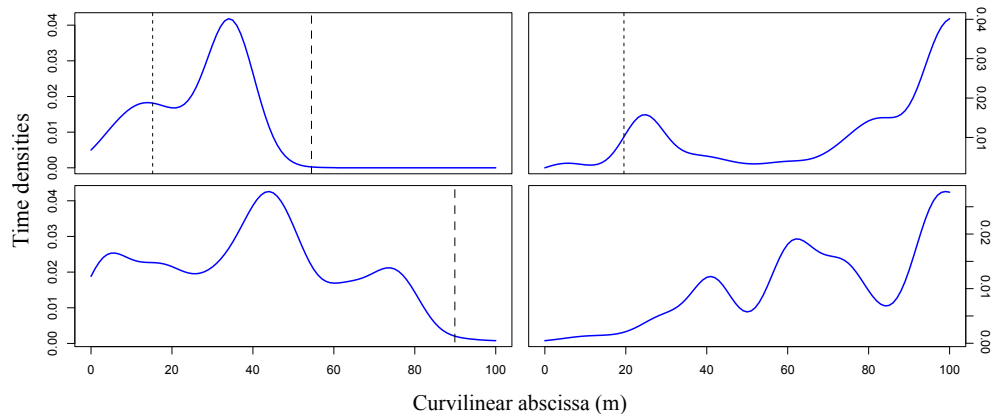
Note that this definition is slightly different from the standard KDE method, insofar as each kernel function centered in x_i is weighted by the corresponding stop duration t_i . As a consequence, \hat{f} is not normalized:

$$\int_0^L \hat{f}_h(x) dx \simeq \int_{-\infty}^{+\infty} \hat{f}_h(x) dx = \frac{1}{N} \sum_{i=1}^N t_i \int_{-\infty}^{+\infty} K(x - x_i) dx = \frac{1}{N} \sum_{i=1}^N t_i = \mathbb{E}[t]$$

where $\mathbb{E}[t]$ is the expected stop time of all vehicles in the frame (this holds provided that the bandwidth parameter h is small in front of the instance dimension L). Similarly, as illustrated on figure 4, the integral of \hat{f}_h over a given segment $[x_1, x_2]$ is equal to a theoretical amount of time vehicles are expected to stop between curvilinear abscissa x_1 and x_2 . Four examples of stop time distributions are depicted on figure 5 below.

Following a methodology inspired by [9], the resulting function is sampled at n evenly spaced locations to form the explanatory variable vector $X \in \mathbb{R}^n$. Eventually, target variables are computed. Binary classification variable $Y_1 \in \{0, 1\}$ denotes the presence of a traffic signal in the instance. If $Y_1 = 1$, regression variable $Y_2 \in [0, L]$ specifies the stop line location, measured as its distance to the intersection node (stop line abscissa on figure 4).

From a practical viewpoint, since we assumed stop durations are integer values, \hat{f}_h may be computed with any standard KDE library, simply by oversampling data in such a way that each point x_i is present a number t_i of times. Besides, given that in efficient implementations of KDE, computation is done with Fast Fourier Transform algorithm, it makes sense to set the numerical value of n as a power of 2. In our application case, we took $n = 64$. Though it may be demonstrated that the mean integrated squared error is minimal with Epanechnikov kernel, the choice of the kernel function is not critical. Therefore we used a gaussian kernel. The bandwidth parameter has been set independently for each instance, according to Silverman's rule [22], which is optimal for normally distributed observations.



■ **Figure 5** Examples of stop time distributions: the top two instances are positive (dashed line indicates traffic signal position), while the bottom two are negative. When the edge is shorter than 100 m, the thick dashed line denotes the end of the edge segment.

3.2 Training and validation

Given a set \mathcal{D} of training instances in $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^n$ and $\mathcal{Y} = \{0, 1\}$ denote input and output spaces, respectively, and a new feature vector $X_u \in \mathcal{X}$, whose label Y_u is unknown, the task of a classifier is to estimate the probability of a traffic signal presence $\mathbb{P}(Y_u = 1|X_u, \mathcal{D})$. X_u is classified as positive whenever the estimated value is greater than 0.5. For regression problems, $\mathcal{Y} = \mathbb{R}$, and the objective is to estimate the conditional expectation $\mathbb{E}[Y_u|X_u, \mathcal{D}]$.

Introduced by Breiman [6] two decades ago, Random Forests (RF) algorithm is a statistically robust version of decision trees, relying on ensemble method concept to reduce prediction variance of individual decision trees. Given a collection of T decision trees whose posterior probability estimate is P_t , the overall posterior estimation is calculated as an average of predictions made by each individual tree:

$$\mathbb{P}(Y|X) = \frac{1}{T} \sum_{t=1}^T P_t(Y|X)$$

This makes Random Forests a simple, fast and efficient classification and regression tool, often considered as robust to over-fitting and particularly useful in high-dimensional problems where one has no strong reason to believe that all features will be helpful for discriminating instances. Moreover, in his foundation paper, Breiman introduced as well parameters setting empirical rules, which makes the tuning process quite straight-forward. For more detailed information about RF, we recommend the complete and extensive works of Louppe [15] for the theoretical background or Criminisi *et al.* [8] for a presentation of some of its capabilities in a wide range of practical problems.

The final dataset contains 4611 instances, including 662 (14%) positive samples. While the entire dataset is not overwhelmingly labeled as negative, this significant imbalance in favor of negative instances may markedly penalize the training process [2]. To overcome this issue, we tried different strategies: down-sampling (randomly suppressing negative samples until dataset is balanced) and up-sampling (replicating positive samples: this second strategy has the advantage of keeping all the information available from the data, at the expense of increasing correlation between individual samples). We also tried SMOTE algorithm [5],

which is similar to up-sampling, but instead of replicating the minority class examples, new examples are generated by interpolation between randomly sampled neighbor instances of this class. We used $T = 500$ trees, and at each split $\sqrt{n} = 8$ features are taken into account. The model was validated by 10-fold cross validation, i.e. by training the algorithm on 90 % of the data, and validating it with the remaining 10 %, and repeating this process 10 times.

3.3 Inference on crossings

Given an intersection between a number n of incoming streets, each of them being classified by RF with a probability p_i of containing a traffic signal. We know that since the aggregated prediction relies on non-independent trees, and aggregation is calculated with a sum instead of a product, the values p_i are not strictly speaking probabilities. However, using the belief theory and Dempster-Shafer combination rule, it can be demonstrated through recurrence on n that the total belief towards the presence of a traffic signal on the intersection is:

$$\pi(p_1, p_2 \dots p_n) = \prod_{i=1}^n p_i \times \left(\prod_{i=1}^n p_i + \prod_{i=1}^n (1 - p_i) \right)^{-1}$$

The intersection is then classified as controlled by a traffic signal when $\pi \geq \frac{1}{2}$. Using this combination rule, we may aggregate predictions on individual streets into a unique probability on the entire crossing, trading granularity for precision.

4 Results and discussion

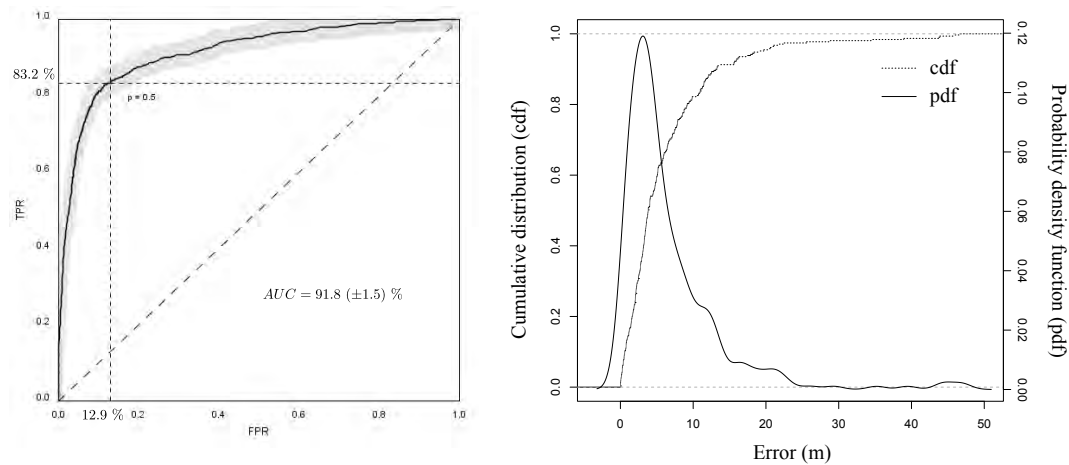
The whole experimental process has been implemented in R with *randomForest* package [13] and launched on an Intel Core(TM) i7-3770 processor (3.40 GHz RAM 8 Go). We computed the following performance scores: specificity (or 1 - false positive rate, which corresponds to the recall), sensitivity (or true positive rate), area under receiver operating curve (AUC), training time (for a single fold, i.e. on 90% of the data), and overall accuracy.

■ **Table 1** Detection performance scores for different way of balancing data.

Scores	Down-sampling	Up-sampling	Imbalanced	SMOTE
Specificity (%)	87.10	95.97	97.23	95.87
Sensitivity (%)	83.25	63.34	57.18	63.98
Accuracy (%)	86.57	91.49	91.73	91.49
AUC (%)	91.38	91.52	91.26	91.25
Training time (s)	1.35	6.98	3.83	7.18
Number of instances	1191	7108	4149	7108
OOB error rate (%)	14.46	2.00	8.23	2.36

Note that RF algorithm provides a practically unbiased error estimate during training phase (without validation dataset), called out-of-bag (OOB) estimate. Indeed, since training data are bootstrapped before used to grow decision trees, for a sufficiently large number of training data, it can be demonstrated that on average, each sample is not seen by a fraction $(1 - 1/n)^n \sim e^{-1}$ of trees. As a direct implication, each instance may be used as a training data for 63 % of trees, and passed through validation with the 37 % remaining trees.

From table 1, we observe that, as expected, the time complexity of the training process is roughly proportional to the number of training samples. Besides, area under curve (and then the overall performance) does not seem to depend upon the method selected for balancing the



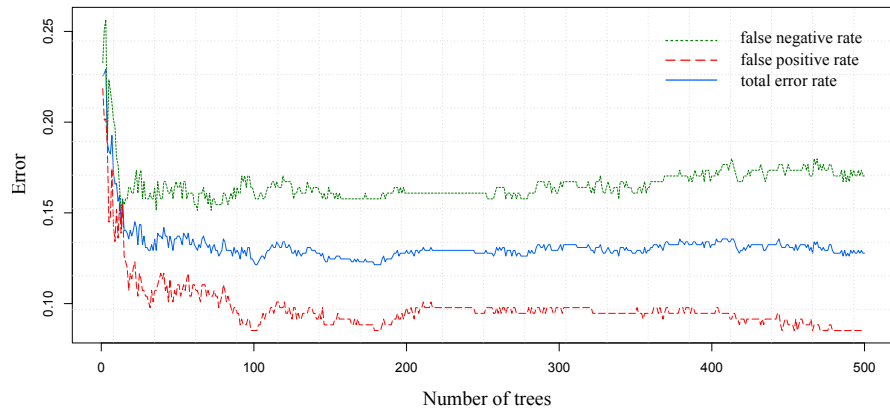
■ **Figure 6** Left: Receiver Operating Characteristics (ROC) curve of the classifier with 95% confidence bands (computed with bootstrap method). Right: probability density and cumulative distribution functions of regression errors.

data. Everything happens just as if the four classifiers above correspond to different selection threshold of the same classifier model. Therefore, in the remaining of this section, we will only use down-sampling since it decreases the number of instances to process, resulting in a minimal computation time. It is worth noticing, that while OOB estimate is often acknowledged as being quite reliable, it completely fails to provide realistic error estimate on up-sampling and SMOTE experiments. This may be explained by the fact that with these two balance procedures, two identical (or at least very similar) sample data may be in and out-of-bag, which amounts to validating a model with samples partially contained in training dataset.

Figure 6 depicts detection and localization performances for the down-sampling version of the algorithm. Area Under Curve index of the classifier is equal to 91.8 (±1.5) % which is considered as a fairly good result. Though specificity is not so high (compared to the number of potential false positive that might be detected on a typical road network), the ROC curve is remaining close to the *no false positive* vertical line even for decent value of true positive rate. This observation instills confidence in the possibility of building a semi-automatic process, achieving a satisfying recall, and entailing only few manual corrections. However, on the other side of the ROC curve, it seems difficult to get all traffic signals, without spending a lot of time separating true and false positive detections. From a more practical viewpoint, it is also worth noticing that our recall may be compared with OSM (with the substantial advantage that our algorithm performs detection on each individual traffic signal, not only on the entire crossing).

■ **Table 2** Localization performance scores. RMSE: root mean square error.

Scores	Mean error	Median error	Mode of errors	RMSE
Estimate (m)	6.22	3.82	2.65	9.51
Std. deviation (m)	±0.4	±0.3	±0.3	±0.8



■ **Figure 7** Out-Of-Bag (OOB) error estimate convergence versus number of trees T .

Besides, as depicted on the cumulative distribution function of regression errors, 82 % of errors are below 10 m, 60 % below 5 m, and 14 % as precise as 1 m. The root mean square error equals 9.51 (± 0.8) m, (which is to be put in perspective to the 20 m of the standard deviation of the explained variable before regression), while mean, median and mode values are much lower, indicating that the distribution is significantly right-skewed. This calls for a more general discussion over what *detection* means. It might be more reasonable to count outliers as undetected (a stop line detected with 50 m inaccuracy cannot be legitimately considered as detected), as a result, the recall would decrease slightly by 4 % and as a reward, the RMSE of localization drops to 6 m, and mean error to 4 m.

Similarly to many ensemble method algorithms, RF is robust to overfitting, and while there is no guidelines for selecting an adequate number of trees, it is admitted that an excessive number is not harmful to the prediction (at the expense of an additional burden in computation time at training and inference steps). Figure 7 depicts the evolution of the OOB error estimate as trees are grown in the model. It may be observed that the convergence of predictions has been reached with approximately 100 trees.

Detailed inspection of the results revealed that many false detections occurred on places where very few vehicles traveled, which implies that the algorithm has not reached convergence as far as the number of vehicles is concerned. With a more extensive dataset we could certainly get better results. It would be interesting, in future works, to study the impact of the number of traces on the prediction scores.

A limitation of our work is that, as stated in section 3.1, our choice of frame, located near the intersection node, makes it impossible to detect traffic signals located in the middle of edges. Indeed, a relatively important number of errors occurred on traffic signals activated by pedestrians push button. An interesting proposition to solve this issue would be to up-sample the network by creating artificially dummy nodes evenly spaced on long edges. This approach may be successful to capture the remaining traffic signals. Another strong limitation of this work is that only information extracted from GPS traces upstream of the intersection is used to create the features, although the behavior of drivers downstream of a traffic signal may exhibit some very specific pattern that could help discriminate from stop signs at jammed intersections.

Based on the posterior probability values estimated by the RF, and combining them with the method proposed in section 3.3, we classified crossings into two categories, depending on whether they are controlled by a system of traffic lights. This made sensitivity and

specificity increase to 87.9 % and 96.2 %, respectively, which is more than 8 % improvement in comparison to the per individual traffic light detection. This compares advantageously to OSM traffic signal database, particularly in terms of recall. Yet, specificity is not high enough to ensure fully automatic process without human supervision or post-processing corrections. Future research will try to leverage this correlation to improve results, even at the level of individual traffic signals. This can be done through relational learning techniques [21] and probabilistic graphical models [11], especially since we have a natural definition of network: the road map.

Apart from tuning more thoroughly the model parameters and the choice of features (additional data would preclude from over-fitting), among the main perspectives of improvement, we may attempt to use functional data analysis to decompose time distribution on an *ad hoc* basis of functions (e.g., wavelets, Karhunen-Loève transform...), in an attempt to minimize correlation between features. Extracting some other physical parameters such as speeds, accelerations, jerks... may also help discriminating traffic signals, as well as localizing it more precisely. This is possible, provided that GPS data speed profiles are smooth enough. Eventually, we may consider building *spatio-temporal* feature vectors, with a bi-dimensional kernel density estimation, where one dimension is the stop time and the second dimension is the stop position along the road axis.

5 Conclusion

Floating Car Data have been used so far in a wide variety of applications to infer the road network and its attributes. However, to the best of our knowledge, the method proposed in this paper is the first attempt to use multiple probe vehicle GPS traces along with statistical learning techniques to detect and localize traffic signals. Learning on a weighted-time distribution of stop points can reach up to 85 % detection scores, and approximately 5 m in positional accuracy. These results are promising for the future development but it is not yet sufficient at the moment to be used as a fully automatic detection system. Nonetheless, this algorithm might find some applications as it is, as a semi-automatic map inference algorithm with human post-process corrections, or when combined with other sources of data (e.g., sensors, embedded cameras, aerial images...) to provide a refined estimation with multi-source data fusion techniques. Future works will aim at improving detection scores by extracting more features from the data, and at extending this approach to other kinds of infrastructure elements. In the long run, one of the main prospects for this research is unquestionably autonomous cars, which, in addition to self-driving, would be self-mapping their environment and sharing information in a completely autonomous loop.

References

- 1 Cindie Andrieu, Guillaume Saint Pierre, and Xavier Bressaud. Estimation of space-speed profiles: A functional approach using smoothing splines. In *Intelligent Vehicles Symposium (IV), 2013 IEEE*, pages 982–987. IEEE, 2013.
- 2 Gustavo E. A. P. A. Batista, Ronaldo C. Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl.*, 6(1):20–29, 2004.
- 3 James Biagioni and Jakob Eriksson. Inferring road maps from global positioning system traces: Survey and comparative evaluation. *Transportation Research Record: Journal of the Transportation Research Board*, 2291:61–71, 2012. doi:10.3141/2291-08.

- 4 Olivier Bonin. *Modèle d'erreurs dans une base de données géographiques et grandes déviations pour des sommes pondérées ; application à l'estimation d'erreurs sur un temps de parcours*. Thèse de doctorat, spécialité mathématiques - statistique, Université Paris VI - Pierre et Marie Curie, mar 2002.
- 5 Kevin W. Bowyer, Nitesh V. Chawla, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *CoRR*, abs/1106.1813, 2011.
- 6 Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- 7 Yihua Chen and John Krumm. Probabilistic modeling of traffic lanes from gps traces. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 81–88, November 2010. doi:10.1145/1869790.1869805.
- 8 A Criminisi, J Shotton, and E Konukoglu. Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning. *Microsoft Research Cambridge, Tech. Rep. MSRTR-2011-114*, 5(6):12, 2011.
- 9 Mohamed el Habib Boukhobza and Malika Mimi. Classification automatique de la densité des tissus mammaires. *Traitement du Signal*, 33:441–460, 2016.
- 10 Alireza Fathi and John Krumm. Detecting road intersections from gps traces. In *International Conference on Geographic Information Science*, pages 56–69. Springer, 2010.
- 11 Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.
- 12 Jun Li, Qiming Qin, Jiawei Han, Lu-An Tang, and Kin Hou Lei. Mining trajectory data and geotagged data in social media for road map inference. *Transactions in GIS*, 19(1):1–18, 2015.
- 13 Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.
- 14 Xuemei Liu, James Biagioni, Jakob Eriksson, Yin Wang, George Forman, and Yanmin Zhu. Mining large-scale, sparse gps traces for map inference: Comparison of approaches. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 669–677, New York, NY, USA, 2012. ACM.
- 15 Gilles Louppe. Understanding random forests: From theory to practice. *arXiv*, 2014. arXiv:1407.7502.
- 16 Y Méneroux, D Manandhar, S Ranjit, G Saint Pierre, and R Shibasaki. Positional accuracy control in dense urban environment with low-cost receiver and multi-constellation gnss. In *Proc. 9th Multi-GNSS Asia – MGA Conference*, 2017.
- 17 Volodymyr Mnih and Geoffrey E. Hinton. Learning to detect roads in high-resolution aerial images. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision – ECCV 2010*, pages 210–223, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.
- 18 Ana Tsui Moreno and Alfredo García. Use of speed profile as surrogate measure: Effect of traffic calming devices on crosstown road safety performance. *Accident Analysis & Prevention*, 61:23–32, 2013.
- 19 Mario Munoz-Organero, Ramona Ruiz-Blaquez, and Luis Sánchez-Fernández. Automatic detection of traffic lights, street crossings and urban roundabouts combining outlier detection and deep learning classification techniques based on gps traces while driving. *Computers, Environment and Urban Systems*, 68:1–8, 2018.
- 20 Paul Newson and John Krumm. Hidden markov map matching through noise and sparseness. In *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems*, pages 336–343. ACM, 2009.
- 21 Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93, 2008.
- 22 Bernard W Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.


- 23 Mohit Dev Srivastava, Shubhendu Sachin Prerna, Sumedha Sharma, and Utkarsh Tyagi. Smart traffic control system using plc and scada. *International Journal of Innovative Research in Science, Engineering and Technology*, 1(2):169–172, 2012.
- 24 Leon Stenneth and Philip S. Yu. Monitoring and mining gps traces in transit space. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 359–368, 2013. doi:10.1137/1.9781611972832.40.
- 25 Karl Van Winden, Filip Biljecki, and Stefan Van der Spek. Automatic update of road attributes by mining gps tracks. *Transactions in GIS*, 2016.
- 26 Christopher K. H. Wilson, Seth Rogers, and Shawn Weisenburger. The potential of precision maps in intelligent vehicles. In *IEEE International Conference on Intelligent Vehicles*, pages 419–422. Citeseer, 1998.
- 27 Lijuan Zhang and Monika Sester. Incremental data acquisition from gps-traces. In *Geospatial Data and Geovisualization: Environment, Security, and Society; Special Joint Symposium of ISPRS Commission IV and AutoCarto*, 2010.
- 28 Qiaoping Zhang and Isabelle Couloigner. Automated road network extraction from high resolution multi-spectral imagery. In *Proceedings of ASPRS 2006 Annual Conference*, pages 01–05, 2006.

Heterogeneous Skeleton for Summarizing Continuously Distributed Demand in a Region

Alan T. Murray

Department of Geography, University of Santa Barbara, CA, USA


amurray@ucsb.edu

 <https://orcid.org/0000-0003-2674-6110>

Xin Feng

Department of Geography, University of Santa Barbara, CA, USA


xin.feng@geog.ucsb.edu

 <https://orcid.org/0000-0001-6434-3895>

Ali Shokoufandeh¹

Department of Computer Science, Drexel University, Philadelphia, PA, USA

ashokouf@cs.drexel.edu

 <https://orcid.org/0000-0002-3729-4490>

Abstract

There has long been interest in the skeleton of a spatial object in GIScience. The reasons for this are many, as it has proven to be an extremely useful summary and explanatory representation of complex objects. While much research has focused on issues of computational complexity and efficiency in extracting the skeletal and medial axis representations as well as interpreting the final product, little attention has been paid to fundamental assumptions about the underlying object. This paper discusses the implied assumption of homogeneity associated with methods for deriving a skeleton. Further, it is demonstrated that addressing heterogeneity complicates both the interpretation and identification of a meaningful skeleton. The heterogeneous skeleton is introduced and formalized, along with a method for its identification. Application results are presented to illustrate the heterogeneous skeleton and provides comparative contrast to homogeneity assumptions.

2012 ACM Subject Classification Applied computing → Operations research, Information systems → Geographic information systems, Theory of computation → Computational geometry

Keywords and phrases Medial axis, Object center, Geographical summary, Spatial analytics

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.12

1 Introduction

An area, polygon and/or region is often the byproduct of political, administrative or management delineation, but such an object can also be used to represent in situ phenomena and attributes. Irrespective of its origin, summary, explanation and characterization of the spatial extent of an area-based object can be very important. One approach for summary representation has been through the use of the skeleton, or medial axis among other names. Okabe et al. [16] note the ability of the skeleton to characterize the shape of a polygon. In cartography, the skeleton may be used for effective label placement, contributing to visual appeal and enhanced communication of a display and/or map. Bruck et al. [5] and Matisziw

¹ Ali SHokoufandeh's work supported in part by the National Science Foundation under Grant Number PFI:AIR-TT:1640366.



© Alan T. Murray, Xin Feng, and Ali Shokoufandeh;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

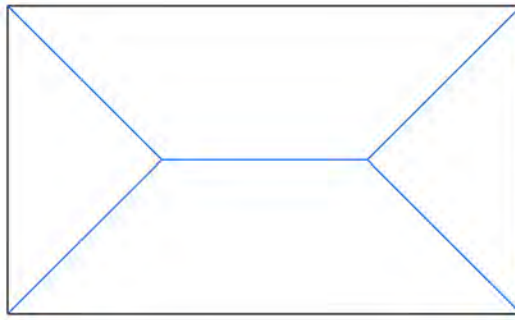
Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 12; pp. 12:1–12:11

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

12:2 Heterogeneous Skeleton



■ **Figure 1** Skeleton for a rectangle region.



■ **Figure 2** Heterogeneous attribute for the rectangle region.

and Murray [14] have demonstrated important spatial properties of the skeleton, as have others.

The skeleton is a line-based object that is represented by the locus of all points equidistant to at least two nearest locations on the polygon boundary it describes. Figure 1 depicts the skeleton (colored blue) for a rectangle region (colored black). Interestingly, definition of the skeleton has focused only on the polygon boundary, devoid of any other spatial attributes. In particular, one might consider an attribute distributed within a polygon as an important influencing factor, if such information is available. It may be that only the total value of an attribute for a polygon is known, and not its actual spatial distribution within the polygon. We know the total attribute value within the rectangle region in Figure 1 to be 58,217. Clearly in such a case, the standard definition based on polygon boundary makes sense. However, if the spatial attribute distribution within a polygon is indeed known, then this should influence the shape of the skeleton if it is to reflect both boundary and attribute information. Figure 2 depicts the spatial variability of the attribute in the rectangle region (Figure 1), where darker colors correspond to higher attribute values (greater population). The 58,217 people in this region are not uniformly distributed, but rather are non-uniform, with a high of 40 people in the left top corner cell and a low of one in right bottom corner cell. One can characterize the skeleton based only on polygon boundary as homogeneous, whereas a skeleton based on boundary and spatially varying attribute(s) within the polygon would be better described as heterogeneous. Figure 3 depicts the heterogeneous skeleton for the rectangle region, accounting simultaneously for both boundary and attribute variability.

In this paper we introduce the heterogeneous skeleton to simultaneously reflect boundary and attribute variability of a polygon. The idea is to provide enhanced summary and characterization, taking advantage of the greatest amount of information possible. The



■ **Figure 4** A polygon-based region, Φ .

a line-based object that is a summary of this polygon. The skeleton represents one such approach.

A set theoretic model for the skeleton can be structured. Assume we have a simple polygon object, Φ . Further, this polygon can be converted to its polyline representation, φ . Both objects are now used in the characterization of the skeleton:

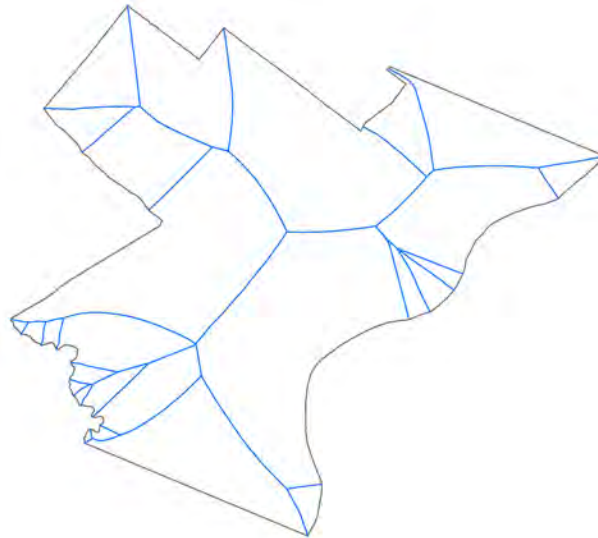
$$S = \left\{ p \in \mathbf{R}^2 \mid \forall r \in \mathbf{R}, \left(\delta(p, r) \subset \Phi \right) \wedge \left(|\delta(p, r) \cap \varphi| \geq 2 \right) \right\} \quad (1)$$

where p is a point in two-dimensional space, r is a distance (Euclidean), and $\delta(p, r)$ is a polyline object (circle) of distance r from point p . The skeleton results from an infinite collection of instances of p and r , where $\delta(p, r)$ is contained in Φ and $\delta(p, r)$ intersects φ in two or more tangent points. Further discussion of the skeleton can be found in Okabe et al. [16] and Matisziw and Murray [14].

Given this formal specification of the skeleton, it may be derived using a number of methods. There are different approaches that have been well documented for skeleton extraction of two- and three-dimensional objects. They can be grouped into three major categories based on their principles and object representation:

1. Voronoi - Algorithms based on the Voronoi diagram or continuous geometric approaches of point clouds, polygonal, or polyhedral representations of object boundaries. Based on properties of the Voronoi diagram, Voronoi edges or planes can be used to construct symmetry structures, or the skeleton.
2. Thinning - Algorithms that rely on the continuous evolution of object boundaries. For example, the object boundary is shrunk with the spread of fire starting at the boundary, the so called grassfire algorithm. The skeleton is formed at the location of singularities, referred to as the “quench points” where fires from different parts of the boundary meet.
3. Distance transformation - Algorithms using the principle of digital morphological erosion or location of singularities, e.g., local maxima, on a digital distance transform field.

Figure 5 illustrates the associated skeleton for polygon Φ shown in Figure 4. As noted previously, the skeleton is the byproduct of evaluation that considers only polygon boundary. As a result, there are no attribute oriented influences in the structure of the skeleton.



■ **Figure 5** Homogeneous skeleton of polygon Φ .

3 Heterogeneous Skeleton

A polygon region that has associated attribute detail about variability within it represents a rich source of information. While a standard assumption is to assume that a polygon attribute is uniformly distributed across the area it delineates, when ancillary information exists regarding the actual spatial distribution of an attribute, this is particularly valuable. The skeleton, S , defined using (1) assumes homogeneity and is derived solely on the basis of polygon boundary φ . Yet, more may be known about attribute variability, and this has the potential to provide greater spatial richness to a line-based summary. As an example, Figure 6 indicates population density for the study region. Darker shades indicate higher population density, and it is clearly not uniform across the polygon. Extending the skeleton/medial axis to account for both geographic boundary as well as heterogeneity in the distribution of attributes across Φ is important.

This means then that one must be able to explain and account for attribute variability. In continuous space the function $g()$ defines the attribute value for any point $q \in \mathbf{R}^2$.

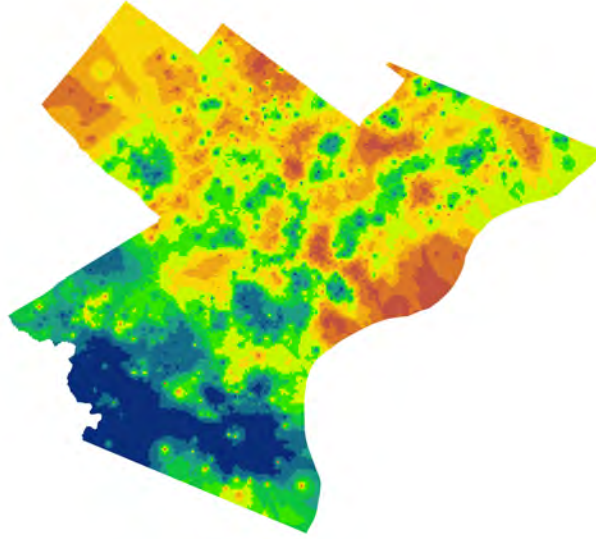
Using set theory notation, we introduce the heterogeneous skeleton as:

$$W = \left\{ \rho^* \in \mathbf{R}^2 \mid \forall p \in \mathbf{R}^2, r \in \mathbf{R}, \left(\delta(p, r) \subset \Phi \right) \wedge \left(|\delta(p, r)| \cap \varphi \geq 2 \right) \right. \\ \left. \wedge \min_{\rho^*} \iint_{q \in \delta(p, r)} g(q) \gamma(\rho^*, q) dq \right\} \quad (2)$$

where $\gamma(\rho^*, q)$ is the distance between ρ^* and q . Building on the homogeneous skeleton, S , definition in (1), the heterogeneous skeleton in (2) adds the additional condition that the inscribed circle, $\delta(p, r)$, serves as an object for which the best representative point is sought. This representative point then helps to define the proposed skeleton variant.

The subproblem communicated in (2) is:

$$\min_{\rho^*} \iint_{q \in \delta(p, r)} g(q) \gamma(\rho^*, q) dq \quad (3)$$



■ **Figure 6** The attribute variability within region Φ .

This is actually a continuous space optimization problem (see Church and Murray [6]). A discrete variant of (3) is what Rogerson [17] and others refer to as the weighted median center. With demand in $\delta(p, r)$ distributed according to the function $g(\cdot)$, the distance $\gamma(\rho^*, q)$ from point q to the optimal median center ρ^* reflects the weighted distance. That is, we seek the optimal ρ^* for each inscribed circle, $\delta(p, r)$, such that the total weighted (attribute) distance is minimized. It therefore is the most efficient or most representative center point for $\delta(p, r)$. The collection of optimal ρ^* for all points $p \in \mathbf{R}^2$ satisfying (2) results in the heterogeneous skeleton.

Often the attribute function $g(\cdot)$ is approximated in some way (Yao and Murray [22]), where $\delta(p, r)$ is delineated into smaller reporting units or cells. The index i , $i \in \{1, \dots, n\}$, is used to refer to discrete points/units in $\delta(p, r)$, where (x_i, y_i) are the coordinates of unit i . Naturally, g_i represents the observed attribute value for unit i . If the coordinates of ρ^* are (X, Y) , then these are the subproblem decision variables. The weighted median center is therefore the following problem:

$$\min_{(X, Y)} \sum_{i=1}^n g_i \sqrt{(x_i - X)^2 + (y_i - Y)^2} \quad (4)$$

The distance function, $\gamma(\cdot)$, in this case is the Euclidean metric. As noted in Wesolowsky [21] and Church and Murray [6], (4) is nothing other than the Weber problem and can be solved using the Weizfeld algorithm.

With the problem description and details, an approach to solve (2) is possible. Pseudo code for the solution process is as follows:

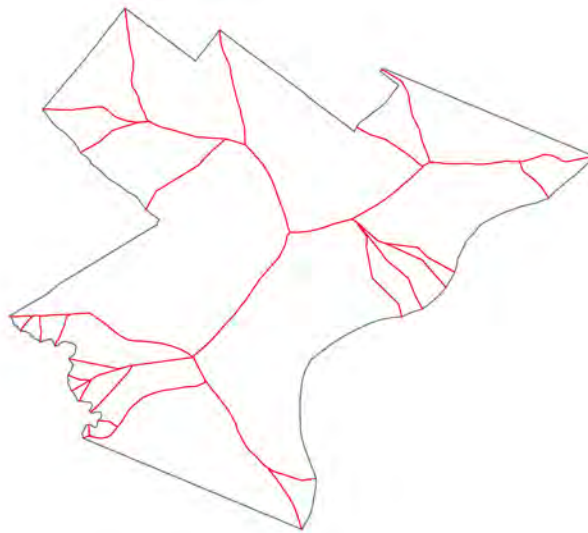
Effectively, the proposed approach must first identify each inscribed circle, as done for S in (1). However, the point to include on the heterogeneous skeleton, denoted as W in (2), is defined based upon the weighted median center criteria, (4). Depending on the structure of polygon Φ , as the number of ρ^* defining W increases, the associated heterogeneous skeleton results.

Algorithm 1 Overview of heterogeneous skeleton derivation.

```

for  $\delta(p, r)$  in  $\Phi$  do
   $\delta(p, r) \subset \Phi$ 
   $|\delta(p, r) \cap \phi| \geq 2$ 
  Find  $\rho^*$ , or rather  $(X, Y)$  using (4)
end for

```



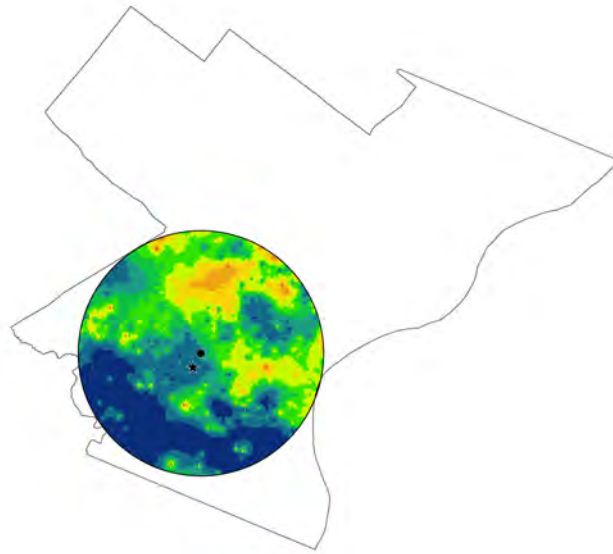
■ **Figure 7** Heterogeneous skeleton (boundary and attribute variability).

4 Results

The models were implemented in the Python platform using `arcpy`, `pysal` and `sympy` libraries, among others, on a Windows 10 Enterprise server with an Intel Xeon E5-2650 v3 (2.3GHz) 64 bit CPU and 64 GB of RAM. ArcGIS was utilized for data creation, management, manipulation, analysis, and visualization. Reported findings required only seconds or minutes to derive.

The heterogeneous skeleton is shown in Figure 7 for the study region (Figures 4 and 6). In comparison to the homogeneous skeleton (Figure 5), there is much variability in the line-based object in terms of precisely where the skeletal line segments are located. The reason for this is highlighted in Figure 8, where an inscribed circle is depicted, $\delta(p, r)$ and helps to form the derived skeleton. The unweighted median center is shown using the symbol \bullet . This is the feature which is used to define the homogeneous skeleton, S , in equation (1). In contrast, the weighted median center, (4), is shown using the symbol \star . That is what is being used to define the heterogeneous skeleton, W , in equation (2). Accordingly, the two skeletons are different based upon their resulting line segments. This happens because of the added influence of attribute and its spatial variability, i.e., the higher population density areas are effectively pulling the skeleton to create a shape and location that is more representative of the distribution of the underlying attribute.

One final question to consider is how distinct the heterogeneous and the classic homogeneous skeletons are. While visual inspection and comparison highlights significant differences, aspects of quantification are possible. One distinction can be made in terms of how far apart



■ **Figure 8** Inscribed circle with associated attribute variability.

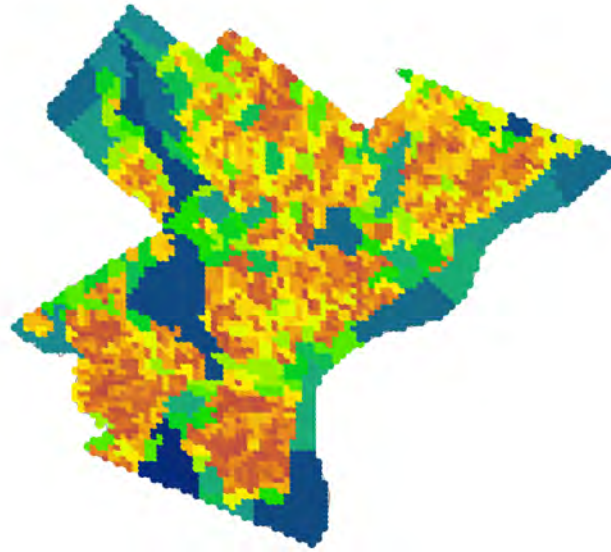
the unweighted and weighted median center are from each other for each inscribed circle used in defining the skeleton. In this case, the distance between the homogeneous center and the corresponding heterogeneous center ranges from 3.43 to 3,593.07 ft in this case. The mean distance is 1220.54 ft, with a standard deviation 939.55 ft. As the region is nearly 100 square miles, such differences are highly significant.

5 Discussion and Conclusions

There are a number of issues worth further investigation associated with the heterogeneous skeleton. First, a polygon may have many possible associated heterogeneous skeletons, one for each one of its attributes. For example, if there are m attributes referenced using $j \in \{1, \dots, m\}$, then any unit i would have m unique attribute values g_{ij} . As a result, depending on the spatial variability of the attribute, one could anticipate m unique heterogeneous skeletons that reflect attribute variation along with the influence of boundary footprint. Figure 9 illustrates a second attribute for polygon Φ , with a decidedly different pattern of spatial variability. Figure 10 indicates the associated skeleton in this case. As is evident through visual inspection, the skeleton in Figure 10 is much different from the case where population is considered (Figure 7). Thus, many different heterogeneous skeletons may be possible depending on associated spatial attributes.

The derivation of the heterogeneous skeleton detailed here was based on the notion of inscribed circles, $\delta(p, r)$. Here, the main motivation was to maintain a connection to the original construction of skeletal representation. This also ensures that one can account for boundary and attribute variability. While this is theoretically sound, other definitions of the heterogeneous skeleton too may be appropriate and meaningful. It is conceivable that approaches based on modified differential grassfire operators or distance transform may be mathematically intuitive or computationally more efficient [8, 4].

The paper introduced the heterogeneous skeleton to help simultaneously characterize boundary and attribute variability of a polygon-based region. The classic definition of a skeleton was reviewed, highlighting the focus on the defining boundary only. Taking into



■ **Figure 9** Second regional attribute.



■ **Figure 10** Second regional attribute.

account attribute information in the formalization of the skeleton has many potential benefits given the wide array of already established application areas. In particular, the heterogeneous skeleton represents an approach for summarizing multi-dimensional information that includes both spatial detail as well as locational attributes. The work here represents an initial attempt to define and derive the heterogeneous skeleton.

References

- 1 Sylvain Airault, Oliver Jamet, and Frederic Leymarie. From manual to automatic stereoplotting: evaluation of different road network capture processes. *International Archives of Photogrammetry and Remote Sensing*, 31:14–18, 1996.

- 2 Harry Blum. A transformation for extracting new descriptors of shape. *Models for Perception of Speech and Visual Forms, 1967*, pages 362–380, 1967.
- 3 Gunilla Borgefors. On digital distance transforms in three dimensions. *Computer vision and image understanding*, 64(3):368–376, 1996.
- 4 Heinz Breu, Joseph Gil, David Kirkpatrick, and Michael Werman. Linear time euclidean distance transform algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5):529–533, 1995.
- 5 Jehoshua Bruck, Jie Gao, and Anxiao Jiang. Map: Medial axis based geometric routing in sensor networks. *Wireless Networks*, 13(6):835–853, 2007.
- 6 Richard L Church and Alan T Murray. *Business Site Selection, Location Analysis, and GIS*. Wiley, 2009.
- 7 Christopher M Cyr and Benjamin B Kimia. 3d object recognition using shape similarity-based aspect graph. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 1, pages 254–261. IEEE, 2001.
- 8 James Damon. Smoothness and geometry of boundaries associated to skeletal structures i: sufficient conditions for smoothness (la lissité et géométrie des bords associées aux structures squelettes i: conditions suffisantes pour la lissité). In *Annales de l'institut Fourier*, volume 53, pages 1941–1985, 2003.
- 9 M Fatih Demirci, Ali Shokoufandeh, and Sven J Dickinson. Skeletal shape abstraction from examples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):944–952, 2009.
- 10 Eric Ferley, Marie-Paule Cani-Gascuel, and Dominique Attali. Skeletal reconstruction of branching shapes. In *Computer Graphics Forum*, volume 16, pages 283–293. Wiley Online Library, 1997.
- 11 Stefan Funke. Topological hole detection in wireless sensor networks and its applications. In *Proceedings of the 2005 joint workshop on Foundations of mobile computing*, pages 44–53. ACM, 2005.
- 12 Peter J Giblin and SA Brassett. Local symmetry of plane curves. *The American Mathematical Monthly*, 92(10):689–707, 1985.
- 13 Benjamin B Kimia, Allen R Tannenbaum, and Steven W Zucker. Shapes, shocks, and deformations i: the components of two-dimensional shape and the reaction-diffusion space. *International journal of computer vision*, 15(3):189–224, 1995.
- 14 Timothy C Matisziw and Alan T Murray. Area coverage maximization in service facility siting. *Journal of Geographical Systems*, 11(2):175–189, 2009.
- 15 David Milman. The central function of the boundary of a domain and its differentiable properties. *Journal of Geometry*, 14(2):182–202, 1980.
- 16 Atsuyuki Okabe, Barry Boots, Sugihara Sugihara, Kokichi, and Sung Nok Chiu. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams, second condition*. John Wiley & Sons, 2000.
- 17 Peter A Rogerson. *Statistical Methods for Geography: a student's guide, fourth edition*. Sage, 2015.
- 18 Ali Shokoufandeh, Diego Macrini, Sven Dickinson, Kaleem Siddiqi, and Steven W Zucker. Indexing hierarchical structures using graph spectra. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7):1125–1140, 2005.
- 19 Kaleem Siddiqi, Juan Zhang, Diego Macrini, Ali Shokoufandeh, Sylvain Bouix, and Sven Dickinson. Retrieving articulated 3-d models using medial surfaces. *Machine vision and applications*, 19(4):261–275, 2008.
- 20 Luc Vincent and Pierre Soille. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(6):583–598, 1991.


- 21 George O Wesolowsky. The weber problem: history and perspectives. *Location Science*, 1(1):5–23, 1993.
- 22 Jing Yao and Alan T Murray. Continuous surface representation and approximation: spatial analytical implications. *International Journal of Geographical Information Science*, 27(5):883–897, 2013.
- 23 Yosef Yomdin. On the local structure of a generic central set. *Compositio Math*, 43(2):225–238, 1981.

A Network Flow Model for the Analysis of Green Spaces in Urban Areas

Benjamin Niedermann

Institute of Geodesy and Geoinformation, University of Bonn, Germany


niedermann@igg.uni-bonn.de

 <https://orcid.org/0000-0001-6638-7250>

Johannes Oehrlein

Institute of Geodesy and Geoinformation, University of Bonn, Germany


oehrlein@igg.uni-bonn.de

 <https://orcid.org/0000-0003-0478-4298>

Sven Lautenbach

Institute of Geodesy and Geoinformation, University of Bonn, Germany


sven.lautenbach@igg.uni-bonn.de

 <https://orcid.org/0000-0003-1825-9996>

Jan-Henrik Haunert

Institute of Geodesy and Geoinformation, University of Bonn, Germany

haunert@igg.uni-bonn.de

 <https://orcid.org/0000-0001-8005-943X>

Abstract

Green spaces in urban areas offer great possibilities of recreation, provided that they are easily accessible. Therefore, an ideal city should offer large green spaces close to where its residents live. Although there are several measures for the assessment of urban green spaces, the existing measures usually focus either on the total size of green spaces or on their accessibility. Hence, in this paper, we present a new methodology for assessing green-space provision and accessibility in an integrated way. The core of our methodology is an algorithm based on linear programming that computes an optimal assignment between residential areas and green spaces. In a basic setting, it assigns a green space of a prescribed size exclusively to each resident such that the average distance between residents and assigned green spaces is minimized. We contribute a detailed presentation on how to engineer an assignment-based method such that it yields reasonable results (e.g., by considering distances in the road network) and becomes efficient enough for the analysis of large metropolitan areas (e.g., we were able to process an instance of Berlin with about 130 000 polygons representing green spaces, 18 000 polygons representing residential areas, and 6 million road segments). Furthermore, we show that the optimal assignments resulting from our method enable a subsequent analysis that reveals both interesting global properties of a city as well as spatial patterns. For example, our method allows us to identify neighborhoods with a shortage of green spaces, which will help spatial planners in their decision making.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases urban green, transportation problem, maximum flow, linear program

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.13



© Benjamin Niedermann, Johannes Oehrlein, Sven Lautenbach, and Jan-Henrik Haunert; licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 13; pp. 13:1–13:16

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

The existence as well as the spatial distribution of green spaces in a city have a large impact on the quality of life. Therefore, spatial planners are interested in quantitative measures for the assessment of cities with respect to their green spaces. Different indicators have been suggested for this purpose. In particular, indicators for *green-space accessibility* and *green-space provision* have been described [9]. We argue, however, that the one cannot reasonably be assessed without the other. If, for example, a small green-space exists in the center of a city, it may be accessible for many residents but not at all sufficient to satisfy their demand. Large green spaces at the boundary of a city that are difficult to access, on the other hand, may lead to a positive assessment with respect to green-space provision, although they are of limited use for the city's residents. Therefore, we introduce a new methodology to analyze green-space accessibility and green-space provision in an integrated way.

Our basic idea is to assign a certain amount of green space *exclusively* to each resident, meaning that each green space can supply only a limited number of residents and, thus, is assumed to have a certain *capacity*. We compute the assignments such that a prescribed per-capita demand is satisfied for each resident and the average distance in a road network between residents and assigned green spaces is minimized. We use this *average distance to assigned green spaces* (i.e., the objective value of the solution) as a global quality measure and approximation for the accessibility of the green spaces. For the sake of simplicity, we do not require each resident to be assigned to a single green space but consider the population of a residential area as a quantity that can be split into arbitrary fractions which can be assigned to different green spaces. Such assignments are modeled as a flow from the residential areas via the road network to the green spaces.

Although we consider the average distance to assigned green spaces particularly interesting, we will introduce a more general objective function that allows us to distinguish different types of green spaces and residential areas of different demands. Besides, we will show that the solutions that we obtain provide interesting information on spatial patterns within a city. In particular, since the result of our method depends on several parameters, such as the per-capita demand, we are interested in studying the influence of these parameters on an optimal assignment. A green space far away from any residential area, for example, will be assigned to no resident unless the per-capita demand is set to a very high value. Hence, we can measure the importance of a green space by identifying the smallest per-capita demand for which it is used in the assignment. By visualizing the green spaces with colors representing those values, we obtain a map that highlights important green spaces.

To put our general idea to practice, several design decisions have to be made and technical obstacles have to be overcome. For example, the data set has to be reasonably selected to include all relevant green spaces. Furthermore, green spaces and residential areas are usually given as sets of isolated polygons with no direct connection to the segments of a road data set and, thus, additional links have to be established. The number of residents a green space can satisfy does not only depend on the size of the green space but also on its type (e.g., parks have higher recreational values than forests) and, therefore, needs to be modeled adequately. Moreover, since the polygons representing residential areas and green spaces may be too large and complex to reasonably argue about the distances between them, it may be necessary to partition the polygons into smaller units. All of these aspects are considered in our method in the sense that it offers parameters that should be set by domain experts (e.g., spatial planners). We discuss in detail how these parameters are considered in our method. However, we use rather basic methods and parameter settings in our experiments.

In algorithmic terms, we adapt the *transportation problem* [13], which has been studied frequently to decide how to ship a commodity from a set of suppliers to a set of consumers [8]. For assessing green spaces, however, it has not been applied yet. The transportation problem can be solved with specialized algorithms [5] or via linear programming (LP) [6]. We choose the latter since it can be implemented easily with a mathematical solver and since an LP formulation can be extended easily, for example, to incorporate additional constraints.

The rest of the paper is structured as follows. After discussing related work (Section 2), we introduce a generic network flow model that constitutes the core of our methodology (Section 3). We further present how to deploy this model overcoming several technical obstacles (Section 4) and how to use it for the analysis of green spaces (Section 5). We finally conclude the paper with a short outlook on future work (Section 6).

2 Related Work

Urban green spaces affect the quality of life in a variety of manners. In different fields, researchers stressed the significance of green space to cities considering socio-cultural (e.g., [17, 18]), medical (e.g., [2, 3]), ecological (e.g., [14, 15]), or economic aspects (e.g., [12, 19]). Consequently, there is an increasing interest in measuring and assessing the green-space supply of an urban area (e.g., [7, 10, 19]).

Baycan-Levent et al. [1] make clear that assessing the green space of a city is a complex problem. They perform an analysis on several criteria considering various aspects mentioned above. With their approach, only the green spaces of an urban area themselves are assessed without taking the residential areas into account: The sheer existence of a high-quality green space improves the rating for a city regardless of whether its residents are able to access it. But, especially for benefits arising from visiting a green space its accessibility is crucial.

Comber et al. [4] examine the green-space supply of a city with respect to its residential areas. They perform a *road-network analysis* in order to determine the accessibility of urban green spaces. With respect to the road network, they consider the percentage of citizens living within a certain radius of green spaces exceeding a minimum size. Their approach lacks the complexity of the analysis of Baycan-Levent et al. and a more differentiated global view on the situation in the city. Comber et al. detect for residential areas whether a green space of adequate size is within a certain distance d or not. If not, no further differentiation takes place: For their assessment methodology, it does not matter whether residents have to walk slightly more than d to the next suitable green space or several times the distance. In order to handle this problem, Comber et al. repeat their analysis for various settings concerning the distance to and the minimum size of the considered green spaces.

Sister et al. [16] use a road-network analysis in order to examine *park pressure*, the ratio of the number of people assigned to a park to its area. They use mean park pressure in order to assess the green-space supply of a city. Their method uses Voronoi diagrams for assigning residents. Considering the average, a positive overall rating may hide a park with immense pressure as parks in this model have unlimited capacity. Furthermore, with Voronoi diagrams, each resident is assigned to the closest green space. Sister et al. are aware of this simplification but pursued their strategy since proximity plays an important role to residents for the selection of a park to visit. Nevertheless, this assumption leads to distorted assessments. With this measure, the assessment of the green-space supply of a city can be improved by abolishing small green spaces close to residential areas in order to assign the residents to a different (slightly more distant) and, above all, more capacious green space. Improving the green-space supply by abolishing existing green spaces without replacement is counter-intuitive and, thus, on the downside of this approach.

In a recent work, Grunewald et al. [9] suggested indicators considering both green-space accessibility and provision. For accessibility, they compute the share of inhabitants living within a certain distance from green space. Concerning provision, they examine the green-space area per capita both globally and in walking distance from residential areas. A city with green spaces accessible for many but of insufficient capacity, e.g. in high-density residential areas, earns a high rating with respect to accessibility; A city with large green spaces accessible only for few, e.g. on the outskirts, gets a high provision rating. A combination of both leads to a high overall rating although the city's inhabitants are not satisfied. The problem is that Grunewald et al. rather accumulate than combine accessibility and provision criteria. In this paper, we consider green-space provision and accessibility in an integrated manner.

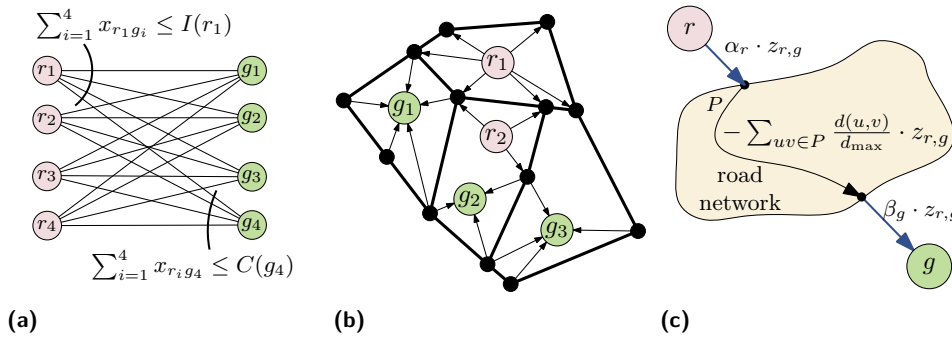
3 Methodology

In this section, we describe the core of our methodology. We first describe the underlying concepts and ideas informally (Section 3.1). Then, we present a formal model implementing these ideas (Section 3.2). This model is rather generic and allows different instantiations that can be adapted for versatile purposes. Finally, we describe a specialization of the model that assumes that residents prefer nearby green spaces (Section 3.3).

3.1 Basic Concepts and Ideas

As discussed in Section 2, several approaches have been suggested to measure and assess the supply of green space in urban areas. One of the simplest approaches is certainly computing the area of green space that is available per resident. However, this measure does not take any information about the structure of the urban area into account. Green spaces far away from residential areas contribute in the same way as green spaces easily accessible by the residents. Hence, as an alternative one may consider the average distance between residential areas and their nearest green space. This, however, ignores the restricted capacity of green spaces. For example, small parks in the city center may not serve all residents, but the typically larger green spaces outside the city boundaries may also be needed to satisfy the demand of the residents. Moreover, while both approaches break down the assessment of green space into an easily comparable number, both do not support a differentiated, spatial analysis on the distribution of green space. However, for urban planning this is precisely essential to answer questions about the importance and accessibility of particular green spaces as well as about the supply of green space to individual residential areas.

We introduce a methodology that interweaves both measures and overcomes their shortcomings. We assume that for each residential area we are given its number of residents and for each green space we are given its capacity, i.e., the maximum number of people that can be served by this area. Intuitively, larger spaces may serve more people than smaller spaces, but this number may also rely on other criteria such as the type of the green space (e.g., a park may serve more people than a forest of the same size). The overall idea of our methodology is to assign the residents of the residential areas to the green spaces such that the average *happiness* of the residents is maximized, while the capacities of the green spaces are respected. We model happiness by rating for each residential area and each green space how much the residents of the residential area prefer that particular green space. This rating typically relies on the distance between the residential area and the green space, but other factors such as the demography of the residential area and the type of the green space may be taken into account. We say a high rating causes high happiness and, altogether, aim for



■ **Figure 1** Assignment Model. Residential areas are represented by red vertices and green spaces by green vertices. (a) Illustration of a generic assignment model. (b) Service network $N = (V \cup R \cup G, E \cup F)$ based on the road network $H = (V, E)$ (black vertices and fat edges), the residential areas R and the green spaces G . (c) Flow $z_{r,g}$ is transmitted from the residential area r to the green space g through the road network on the shortest path P . The flow creates the value given in Equation (7).

an assignment that maximizes the average happiness of all residents. The strength of the model lies in the possibility of applying a detailed spatial analysis on the result; we perform such an analysis in Section 5.2.

3.2 Generic Assignment Model

We now describe how we model the problem formally. We assume that we are given an urban area that consists of a set R of residential areas and a set G of green spaces. Each residential area $r \in R$ has a number $I(r)$ of residents and each green space $g \in G$ has a number $C(g)$ of residents that can be served; we call $C(g)$ the *capacity* of g . We aim to find an assignment such that no green-space capacity is exceeded and the average happiness of the residents is maximized. We formalize this as follows. For a residential area r and a green space g we interpret the triple (r, g, i) such that i residents of r are assigned to g . We call $A \subseteq R \times G \times \mathbb{R}^+$ an *assignment* for (R, G) if it maintains the supply and capacities of the residential areas and green spaces, respectively. That is, we require $\sum_{(r,g,i) \in A} i \leq I(r)$ for all $r \in R$ and $\sum_{(r,g,i) \in A} i \leq C(g)$ for all $g \in G$. However, not every assignment is equally good, but its quality may be affected by multiple criteria such as distances, the type of the green spaces, the mobility of the residents of a residential area, etc. Therefore, we introduce the rating function $h: R \times G \rightarrow [0, 1]$ that describes the preferences of the residents. The higher the value of $h(r, g)$, the more the residents of r prefer the green space g . Altogether, we aim to find an assignment A such that the *total happiness* $\sum_{(r,g,i) \in A} h(r, g) \cdot i$ is maximized; we call that problem **GREENSPACEASSIGNMENT**. For any assignment A we assume that it only contains triples that contribute to the objective, i.e., there is no $(r, g, i) \in A$ such that $h(r, g) = 0$. We note that there might be residents that are not assigned to any green space; we say that these are *unsatisfied*, while all others are *satisfied*.

From a computational point of view, **GREENSPACEASSIGNMENT** can be easily reduced to finding a maximum flow in a complete bipartite graph formed by R and G ; see Figure 1(a). For the convenience of the reader we present the corresponding LP formulation at this point. For each pair $(r, g) \in R \times G$ we introduce a variable $x_{r,g}$. We interpret $x_{r,g}$ as the number of residents of r assigned to g . Subject to

$$\sum_{g \in G} x_{r,g} \leq I(r) \text{ for all } r \in R \quad (1) \quad \text{and} \quad \sum_{r \in R} x_{r,g} \leq C(g) \text{ for all } g \in G \quad (2)$$

we maximize $\sum_{r \in R} \sum_{g \in G} x_{r,g} \cdot h(r,g)$. The assignment is $A = \{(r,g,x_{r,g}) \mid r \in R \wedge g \in G\}$.

In Section 3.3 we describe one possible variant of this highly general model in more detail in order to demonstrate its application. In Section 6 we sketch further variants.

3.3 Network-Based Assignment Model

We now introduce a specialization of our model in which green spaces are assessed by their attractiveness and their accessibility. We assume that residents prefer nearby and attractive green spaces and are not willing to use green spaces that are further away than a certain distance d_{\max} ; we call this distance the *scope* of the residents. Further, we assume that the mobility of the residents may vary from residential area to residential area. To model the mobility of residents and the attractiveness of green spaces, we introduce for each residential area $r \in R$ and each green space $g \in G$ the weights α_r and β_g , respectively. A higher value corresponds with a higher mobility of the residents in r and a higher attractiveness of g , respectively. To assess the accessibility of a green space g from a residential area r , we take the distance $d(r,g)$ between r and g into account. We obtain this distance from the road network of the considered urban area. For a residential area r we then rate the green space g by $h(r,g) = \alpha_r + \beta_g - \frac{d(r,g)}{d_{\max}}$. We note that $h(r,g)$ may become negative. However, in this case no resident of r is assigned to g because we consider a maximization problem. Consequently, a negative value corresponds with setting $h(r,g) = 0$.

GREENSPACEASSIGNMENT can be solved using the LP formulation above. While this works out for small and medium sized cities, it easily exceeds the storage of a modern server system for large cities because it uses a quadratic number of variables. Instead, we introduce a specialized formulation based on the given road network. This formulation uses a number of variables that is linear in the number of green spaces, residential areas and the size of the road network. This allows us to consider metropolitan cities.

We assume that we are given the road network as a directed geometric graph $H = (V, E)$. From H we derive the *service network* $N = (V \cup R \cup G, E \cup F)$ by adding a vertex for each residential area and each green space; see Fig. 1(b). These vertices are connected to the remaining graph by means of the additional edges in F . More precisely, there is an edge $rv \in F$ with $r \in R$ and $v \in V$ if and only if v is an *access point* of the residential area r . Similarly, there is an edge $ug \in F$ with $g \in G$ and $u \in V$ if and only if u is an access point of the green space g . A vertex of the road network is an access point of a region if a resident may access the region via this point; in Section 4 we describe a simple tool to compute access points of residential areas and green spaces.

We set the length d of the edges in N as follows. For an edge $e \in E$ we define its length $d(e)$ as its geodesic length in the road network. For edges $rv \in F$ incident to a residential area r we define $d(rv) = \alpha_r$. Finally, for edges $ug \in F$ incident to a green space g we define $d(ug) = \beta_g$. Depending on the application we may define d differently, e.g., as travel time.

We are now ready to introduce our LP formulation for this specialized model. For each edge $e \in E \cup F$ we model a flow on e with a variable x_e . This represents the number of residents using edge e . We introduce the following linear constraints.

$$\sum_{rv \in F} x_{rv} \leq I(r) \quad \text{for all residential areas } r \in R \quad (3)$$

$$\sum_{uv \in E \cup F} x_{uv} = \sum_{vw \in E \cup F} x_{vw} \quad \text{for all road network vertices } v \in V \quad (4)$$

$$\sum_{ug \in F} x_{ug} \leq C(g) \quad \text{for all green spaces } g \in G \quad (5)$$

Subject to these constraints we maximize the following objective

$$\sum_{r \in R} \sum_{rv \in F} \alpha_r \cdot x_{rv} + \sum_{g \in G} \sum_{ug \in F} \beta_g \cdot x_{ug} - \sum_{uv \in E} \frac{d(u,v)}{d_{\max}} \cdot x_{uv} \quad (6)$$

The first constraint states that for each residential area r the flow on the outgoing edges does not exceed the number of residents of r . The second constraint preserves the flow within the road network, i.e., flow entering a road network vertex $v \in V$ also needs to leave v on its outgoing edges. Finally, the last constraint ensures that the flow on the incoming edges of a green space does not exceed the capacity of the green space. Put differently, the number of residents that are assigned to a green space does not exceed the capacity of the green space.

The intuition behind the objective can be explained as follows. Consider the flow $z_{r,g}$ of a residential area r that is absorbed by a green space g . As the number of residents using the same edge in N is not limited, we can assume without loss of generality that the flow $z_{r,g}$ is not split anywhere in the flow network. Since each edge $uv \in E$ has cost $-\frac{d(u,v)}{d_{\max}}$ and since we consider a maximization problem, the flow from r uses a shortest path P in the road network to reach g ; see Figure 1(c). Hence, the flow has value

$$\alpha_r \cdot z_{r,g} + \beta_g \cdot z_{r,g} - \sum_{uv \in P} \frac{d(u,v)}{d_{\max}} \cdot z_{r,g} = h(r,g) \cdot z_{r,g}. \quad (7)$$

Consequently, the value of the flow in total is $\sum_{r \in R} \sum_{g \in G} h(r,g) \cdot z_{r,g}$, which corresponds with the objective of GREENSPACEASSIGNMENT.

4 Deployment

We now describe the deployment of the network-based model (Section 3.3) in experiments and practical applications. This is just one way to apply our methodology, but it easily can be adapted to other scenarios. We assume that we are given the residential areas R and the green spaces G of an urban area as simple polygons. Each residential area has a number of residents. The road network is given as a graph $H = (V, E)$ with geometric embedding. We apply two phases. In the first phase, we preprocess the data in 5 steps obtaining an instance of GREENSPACEASSIGNMENT. In the second phase, we solve that instance.

First Phase – Preprocessing

Step 1. Since the polygons representing green spaces may be too large and complex to reasonably argue about the distances between them and polygons representing residential areas, it may be necessary to partition these polygons into smaller units. We use an approach by Haurert and Meulemans [11]. They decompose a simple polygon into a minimum number of simple polygons such that each of the resulting polygons is sufficiently compact, with respect to a measure of dilation from graph theory. We obtain a new set of green spaces formed by these compact polygons that replaces the green spaces in G .

Step 2. We determine the access points of the green spaces and the residential areas. To that end, we buffer each polygon; in our experiments we use an offset of 100 m. Hence, roads closely passing by the original polygon intersect the buffered polygon. Each vertex of the road network in the buffered polygon then is an *access point* of the original polygon.

Step 3. We construct the service network based on the road network H . We add the residential areas in R and green spaces in G as vertices to the road network. For each residential area $r \in R$ and each access point v of r , we introduce the edge rv . Similarly, we introduce for each green space $g \in G$ and each access point u of g the edge ug . We denote the set of edges incident to vertices representing residential areas and green spaces by F . Altogether, we obtain the service network $N = (V \cup R \cup G, E \cup F)$.

Step 4. To reduce the graph's complexity, we iteratively remove any degree-2 vertex by replacing its two edges with a single edge connecting its neighbors; the length of the new edge is derived from the two incident edges. Since we do not use the geometric embedding of H in the subsequent steps, this is a valid operation to speed up shortest path queries.

Step 5. In our model, we assume that residents only use shortest paths. Hence, for each vertex of the service network we compute whether it lies on a shortest path between a residential area and a green space. If this is not the case, we remove the vertex from the road network. Otherwise, we annotate the vertex with the smallest distance between it and any residential area; we call this distance the *accessibility* of the vertex. We use this distance in the second phase to prune the network.

Second Phase – Linear Programming

In this phase, we process the instance of GREENSPACEASSIGNMENT that we have created in the previous phase. To that end, we systematically explore different choices of capacities of green spaces as well as different scopes. More precisely, we assume that there is a demand γ of green space made by each resident; we call γ the *per-capita demand*. The capacity of a green space is then $\frac{\text{area of green space}}{\text{per-capita demand}}$. In our experiments, we not only consider one choice of γ but a set Γ of per-capita demands. Similarly, for the scope we consider a set D of distances. For each pair $(\gamma, d) \in \Gamma \times D$ we solve GREENSPACEASSIGNMENT on the respective instance. That is, we set the capacities of each green space g to $\frac{\text{area of } g}{\text{per-capita demand}}$. Applying $d_{\max} := d$, we then use the LP formulation to solve GREENSPACEASSIGNMENT on the corresponding instance. In the LP formulation, we only consider vertices whose accessibility does not exceed d_{\max} . As result we obtain for each pair (γ, d) the average distance between a resident's residential area and the assigned green space. Besides, for each residential area, we obtain the number of residents that were assigned to a green space. Analogously, for each green space, we obtain the number of residents assigned to this area.

Further, for each per-capita demand $\gamma \in \Gamma$ we compute the smallest scope $D_\gamma \in \mathbb{R}$ for which all residents are satisfied. We compute this distance using a simple parametric search.

5 Experiments

In this section, we describe our experimental evaluation that we use to assess our methodology. We emphasize that the aim of this evaluation is not primarily to find new insights into the structure of specific cities but to demonstrate that the methodology works in general and yields a manifold tool set to analyze the supply of green spaces.

5.1 Data and Experimental Setup

In our evaluation, we have considered 53 urban areas in Germany. As data basis, we use the Urban Atlas 2012¹. For a selection of cities, this atlas provides detailed information about land use in the urban area. It particularly distinguishes between the city and its surroundings. For each city, we extract its residential areas as simple polygons excluding its surroundings. In this atlas, a residential area typically represents one housing block separated from others by roads. The data basis further provides for each residential area an estimated number of residents resulting from downscaling census data. Similarly, we extract green spaces as simple polygons for each city including its surroundings. In contrast to residential areas, green spaces may describe vast regions constituting large parts of the urban area. For our experiments, we only take green spaces tagged with *forest*, *green urban area*, or *sports and leisure facility*. Columns 1–3 of Figure 2 give an overview of the analyzed urban areas. The number of residents ranges from 33 thousand to 2.4 million; the cities have 285 thousand residents on average. The area of considered green spaces ranges from 8.6 km² to 6560 km²; on average there are 659 km² of green space in the urban area. In addition, Column 3 yields information about the area of green space that is available per resident.

The road network is taken from OpenStreetMap². We have chosen the extent of the road network such that any shortest path between residential areas and green spaces is included.

We configure the second phase of our approach as follows. To keep the evaluation simple, we choose $\alpha_r = 1$ for any residential area $r \in R$ and $\beta_g = 0$ for any green space $g \in G$. Hence, for any resident it yields the same gain to leave the according residential area, but there is no reward for entering specific green spaces. This particularly implies that any resident reaches any green space within the globally defined scope, but no resident may exceed that distance. In order to define the capacities and scopes as described in Section 4, we define the per-capita demands as $\Gamma = \{50 \cdot i \mid 1 \leq i \leq 20\} \cup \{1, 10\}$ in m² and the scopes as $D = \{0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 40, 50, 60, \infty\}$ in km. As described above, we solve GREENSPACEASSIGNMENT for all pairs $(\gamma, d) \in \Gamma \times D$. Further, for each $\gamma \in \Gamma$, we compute the smallest scope D_γ such that all residents are satisfied.

We solve the LP formulations using Gurobi 7.0.2³. For the LP formulations, we use continuous variables instead of integer variables. Hence, residents may be distributed on multiple areas. Since we are not interested in the specific assignment of a resident to a green space but aim to maximize the average happiness of the residents, this is a reasonable assumption improving the running time of the applied solver.

The experiments were performed on an Intel[®] Xeon[®] CPU E5-1620 processor. The machine is clocked at 3.6 GHz and has 32 GB RAM. The first phase of our approach is implemented in Python utilizing QGIS 2.18.14⁴. The second phase is written in Java.

5.2 Evaluation

In this section, we sketch different analysis techniques that can be used to assess the green-space supply of urban areas. To that end, we use the following measures.

- For each residential area its *largest satisfiable per-capita demand*: the largest per-capita demand $\gamma \in \Gamma$ such that every resident of that residential area is satisfied.

¹ ©European Union, Copernicus Land Monitoring Service 2018, European Environment Agency (EEA). <http://www.land.copernicus.eu>

² <http://www.openstreetmap.org>

³ <http://www.gurobi.com>

⁴ <http://www.qgis.org>

13:10 A Network Flow Model for the Analysis of Green Spaces in Urban Areas

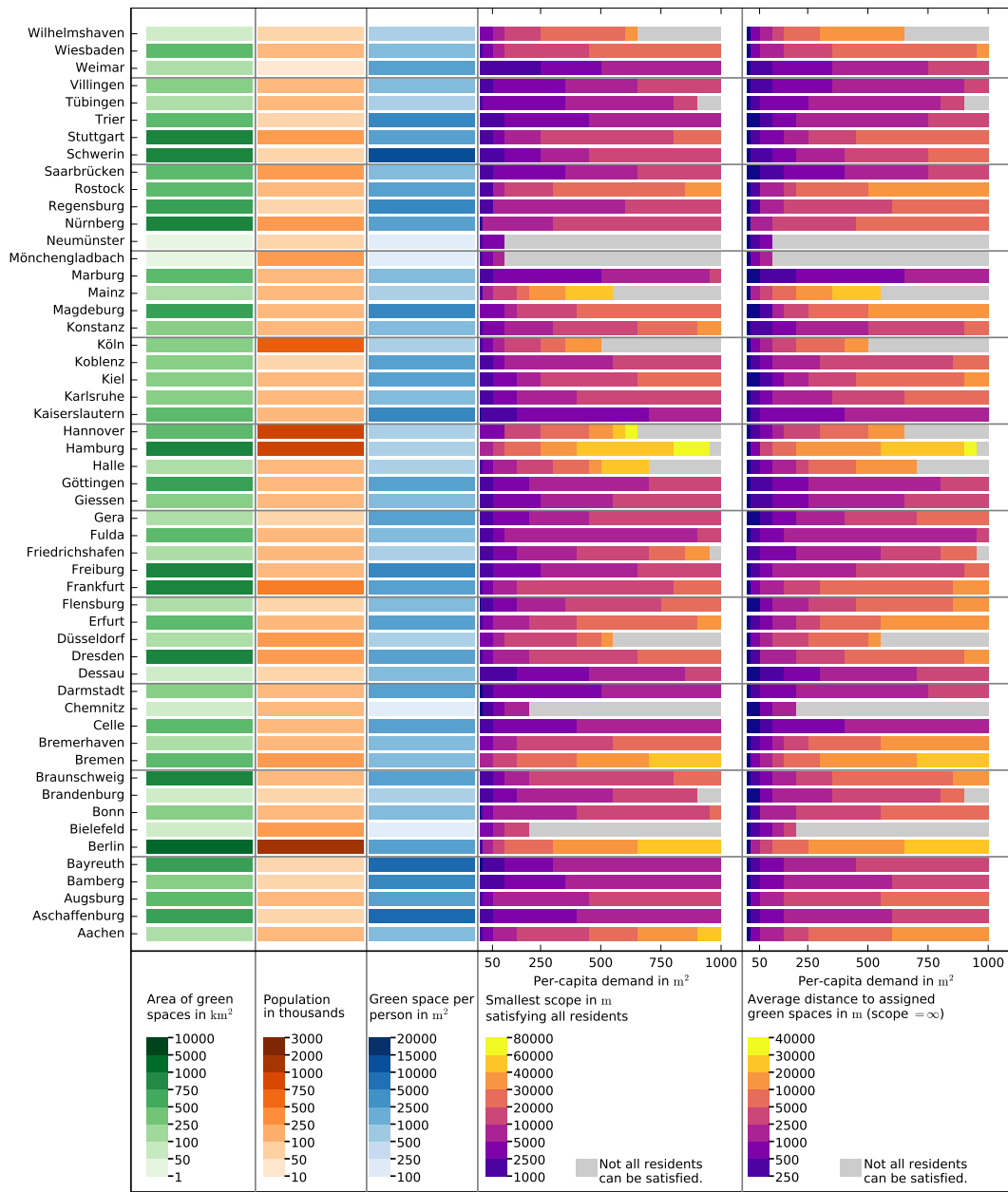
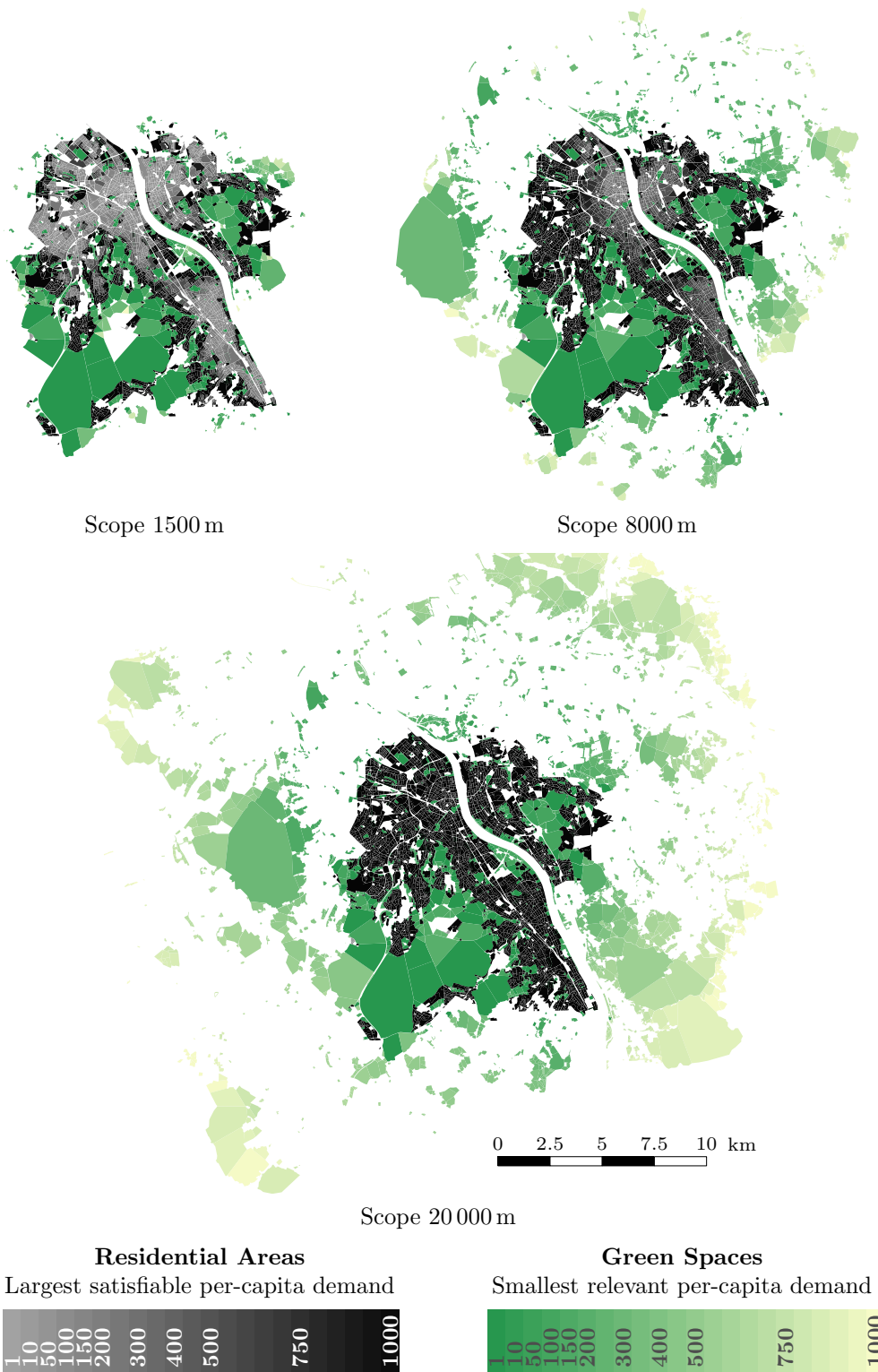


Figure 2 Results for 53 urban areas in Germany. The first three columns give some basic information about the urban areas while the two last columns summarize our results.

- For each green space its *smallest relevant per-capita demand*: the smallest per-capita demand $\gamma \in \Gamma$ such that the green space is used in the assignment.
- For each $\gamma \in \Gamma$ the *smallest scope satisfying all residents*: smallest scope such that all residents of all residential areas are satisfied.
- For each $\gamma \in \Gamma$ the *average distance to assigned green spaces*: the average distance to assign all residents to green spaces considering an infinitely large scope.



■ **Figure 3** Green space supply of Bonn, Germany. An interactive illustration for every scope and every considered city is found on <http://www.geoinfo.uni-bonn.de/urbanarea>.

Green Space Supply of a Single Urban Area

In this section, we discuss the analysis of a single urban area. To that end, we exemplarily consider the urban area of the city of *Bonn*; see Figure 3. As a medium-sized city in Germany its extent can be printed using a reasonable resolution. Using a tool with the possibility of zooming into the map the analysis may also be done on larger cities.⁵

Figure 3 shows the urban area of Bonn with respect to the scopes 1500, 8000 and 20 000 in meters. For each scope we have drawn all residential areas as well as all the green spaces to which residents are assigned; all other green spaces are omitted. Consequently, with increasing scope, more green spaces are shown.

Furthermore, we color each green space with respect to its smallest relevant per-capita demand; see Figure 3. The higher the saturation of the color of a green space, the lower is the smallest relevant per-capita demand. Hence, the saturation of the color shows the *importance* of a specific green space. Similarly, we paint each residential area with respect to its largest satisfiable per-capita demand. The lighter the gray of the residential area is, the lower is the highest per-capita demand for which all residents can be satisfied. Hence, light grays indicate residential areas with poor access to green spaces while dark grays indicate residential areas with easy access to green spaces.

We observe that for the scope of 1500 m there are two regions in Bonn that have full access to green spaces only for small per-capita demands; see light gray regions in Figure 3. With increasing scope the green space supply is apparently improved because the residents begin to reach green spaces further away from the city. However, for the comparatively large scope of 8000 m, there are still residential areas that are only completely satisfied for small per-capita demands. We particularly note that our methodology is robust against small green spaces in the city center. They only impact some nearby residential areas, but do not influence the overall impression that the city center lacks green space supply. Further, the maps indicate that the green spaces on the south side of the city play a particularly important role as local recreation areas.

Comparing the Green Space Supply of Multiple Urban Areas

In our evaluation, we consider 53 cities of different size. Column 4 of Figure 2 shows the smallest scope that is sufficient to satisfy all residents of the considered urban area. The result of a specific urban area can be interpreted as the *robustness* of its green-space supply, which we motivate as follows. For 39 urban areas even a per-capita demand of 1000 m² can be realized without leaving a resident unsatisfied. Hence, their green-space supply is hardly affected even for high per-capita demands. In contrast, there are 14 urban areas whose green-space supply collapses for smaller per-capita demands.

Considering the 39 urban areas in more detail, further differences of large extent are observable. There are 8 urban areas (e.g., *Aschaffenburg*, *Bamberg* and *Bayreuth*) whose scope does not exceed 10 km even if each resident requires 1000 m². In contrast, for 16 of the 39 urban areas a scope of at least 20 km is necessary to satisfy all residents with per-capita demand of 1000 m²; with 48 km Berlin requires the largest scope among those cities.

Considering the 14 urban areas whose green-space supply collapses for per-capita demands smaller than 1000 m², we observe that there are urban areas whose green-space supply already collapses for rather small per-capita demands up to 250 m². For example, for *Neumünster* and

⁵ Illustrations for all considered scopes and cities are found at <http://www.geoinfo.uni-bonn.de/urbanarea>.

Mönchengladbach a per-capita demand of 150 m^2 is not realizable without leaving residents unsatisfied. In these cases, the small scopes indicate that the diameter of the considered surrounding area is not sufficient. In contrast, there are urban areas whose green-space supply collapses only for higher values. For *Hamburg*, for example, all residents can be satisfied up to a per-capita demand of 950 m^2 . However, this requires a scope of 74 km. Hence, the robustness of its green-space supply is dearly bought by a large scope.

Column 5 of Figure 2 shows the average distance to assigned green spaces with respect to the per-capita demands; in case that not all residents can be satisfied the average distance is not presented. The result of a specific urban area can be interpreted as the *accessibility* of its green-space supply, which we motivate as follows. With increasing per-capita demand, the average distance increases depending on the green-space supply of the urban area. For cities with large nearby green spaces, the average distance increases more slowly than the average distance for cities with small nearby green spaces. Hence, for the latter, the local green-space supply becomes easily insufficient for satisfying all residents. For the urban area of *Marburg*, for example, the average distance to assigned green spaces increases slower than the average distance for the urban area of *Wiesbaden*. We emphasize that both regions have a similar population size and a similar total area of green space. Still, on average, the residents of *Marburg* need to cover smaller distances than the residents of *Wiesbaden*, which implies that the green spaces of *Marburg* are more easily accessible than the green spaces of *Wiesbaden*.

Running Time

A typical interactive scenario using our methodology could be as follows. The first phase is applied only once in order to create the service network at the very beginning of the scenario. Once the service network is created, its structure is not changed anymore, but the user gains the possibility of assigning to each residential area and green space attributes (e.g., number of residents, preferences, mobility, etc.). Instead of doing this only once, the user may repeatedly change the attributes to interactively explore the influence of single residential areas and green spaces. Each time, the second phase is executed. Hence, the performance of the repetitively executed second phase is clearly more crucial than the performance of the first phase. With this in mind, we have therefore focused on the second phase.

For the first phase, we put together standard algorithms without engineering their performance. For the urban area of *Berlin* (with 130 000 polygons representing green spaces, 18 000 polygons representing residential areas, and 6 million road segments our largest instance) the first phase takes about 3 minutes.

Solving the LP formulations used by far the greatest portion of the running time of the second phase. In our experiments, we measured the running time for solving $|C| \cdot |D| = 484$ LP formulations per region. Solving a single LP formulation, which we call a *run*, takes 46 seconds in maximum and 5 seconds on average. Over 95 % of all runs took at most 14 seconds. About 89 % of the runs took at most 10 seconds. These running times indicate that our approach does not allow real-time animations, but is usable in interactive systems where the user can update the assignment on demand. Apart from interactive systems, our approach can also be used for the systematic and automatic evaluation of green spaces. Accumulating the running times of all runs of a single urban area yields 3.3 hours in maximum and 40 seconds on average. In total, 35 hours were necessary to process for all 53 cities.

Summary

The presented evaluation demonstrates the strength of our methodology, which stands out by the following features.

- Detailed spatial analysis of single urban areas.
- Simultaneous evaluation of single residential areas and large regions with intuitive maps.
- Easy identification of local recreation areas.
- Robustness against small residential areas and green spaces.
- Sophisticated analysis of multiple urban areas with respect to different measures.
- Practical running times for interactive scenarios and the analysis of multiple urban areas.

We emphasize that domain experts from urban development confirmed the great use of this tool. They particularly highlighted the possibility of spatially analyzing single urban areas.

6 Conclusion & Outlook

We have presented a highly general model for the evaluation of green spaces of urban areas. It is based on the idea of assigning residents to green spaces maximizing the overall happiness of the residents while capacity constraints for green spaces are respected. We have described a specialization of the model and its deployment in detail. It utilizes the underlying road network for computing the assignment. The advantage of this specialization is the better performance obtained by the linear number of variables. This provides the possibility of considering metropolitan cities such as Berlin. In an exemplary evaluation, we demonstrated that the presented methodology can be used for analyzing a single urban area specifically as well as large sets of urban areas in general. Our approach not only yields abstract parameters describing the green-space supply, but it supports a spatial analysis based on the level of single residential areas and green spaces. A discussion panel with domain experts from urban development yielded that our approach will be of great use for urban planning to easily assess existing green-space supply as well as to plan future land usage. Especially, the methodology is of great use in interactive scenarios for urban planning. By means of our approach, an urban planner may interactively explore the influence of potential residential areas and green spaces using maps such as in Figure 3. They may change the importance of green spaces, the preference of residential areas, or even introduce new regions. Each time, our model is updated and the result is visualized. Thus, the user can easily assess the impact of the changes made.

In Section 3.3, we have described one specialization of the generic assignment model. However, the generality of our model provides many different variants. Among others, the following specializations and research questions arise.

- We kept our experiments simple to evaluate the core of our methodology. In practice, it lends itself to use a more complex parameterization reflecting reality more accurately like using travel times instead of geodesic distances in the road network. Further, one may differentiate the mobility of residents and the attractiveness of green spaces by adapting the weights α_r and β_g , respectively. Additionally, introducing further types of recreational areas such as lakes, rivers and open spaces promises a detailed evaluation.
- An interesting followup question is to analyze the utilization of the road network in detail. Which roads are used more than others? May these insights help in traffic planning, especially for weekends? A closer look at the computed flow may give insights.
- The network-based model anonymizes the assignment in the sense that we can not keep track of single residents, but we only obtain how many residents per residential area are assigned to specific green spaces. In some cases, however, it may be useful to analyze the exact assignment. In that case, one may use the generic model of Section 3.2.

- Our approach may also be used to evaluate the accessibility of public services. For example, the coverage of hospitals, medical practices, schools, playgrounds, etc., can be analyzed with our approach as well. In particular, depending on the accuracy of the given data, residential areas may be differentiated by their type of demands.

Altogether, we have presented a generic tool for the assessment of green spaces in urban areas. It can be easily adapted for different applications. For future work, we are planning to apply our methodology on concrete use cases in urban planning.

References

- 1 T. Baycan-Levent, R. Vreeker, and P. Nijkamp. A multi-criteria evaluation of green spaces in European cities. *European Urban and Regional Studies*, 16(2):193–213, 2009.
- 2 A. L. Bedimo-Rung, A. J. Mowen, and D. A. Cohen. The significance of parks to physical activity and public health: A conceptual model. *American Journal of Preventive Medicine*, 28(2):159–168, 2005.
- 3 D. A. Cohen, T. L. McKenzie, A. Sehgal, S. Williamson, D. Golinelli, and N. Lurie. Contribution of public parks to physical activity. *American Journal of Public Health*, 97(3):509–514, 2007.
- 4 A. Comber, C. Brunsdon, and E. Green. Using a GIS-based network analysis to determine urban greenspace accessibility for different ethnic and religious groups. *Landscape and Urban Planning*, 86(1):103–114, 2008.
- 5 L. R. Ford Jr. and D. R. Fulkerson. A simple algorithm for finding maximal network flows and an application to the hitchcock problem. Technical report, RAND Corp., 1955.
- 6 L. R. Ford Jr. and D. R. Fulkerson. Solving the transportation problem. *Management Science*, 3(1):24–32, 1956.
- 7 R. A. Fuller and K. J. Gaston. The scaling of green space coverage in European cities. *Biology Letters*, 5(3):352–355, 2009.
- 8 S. I. Gass. On solving the transportation problem. *Journal of the Operational Research Society*, 41(4):291–297, 1990.
- 9 K. Grunewald, B. Richter, G. Meinel, H. Herold, and R.-U. Syrbe. Proposal of indicators regarding the provision and accessibility of green spaces for assessing the ecosystem service “recreation in the city” in Germany. *International Journal of Biodiversity Science, Ecosystem Services & Management*, 13(2):26–39, 2017.
- 10 K. Gupta, A. Roy, K. Luthra, S. Maithani, and Mahavir. GIS based analysis for assessing the accessibility at hierarchical levels of urban green spaces. *Urban Forestry & Urban Greening*, 18(Supplement C):198–211, 2016.
- 11 J.-H. Haunert and W. Meulemans. Partitioning polygons via graph augmentation. In *Proc. Int. Conf. Geographic Information Science (GIScience 2016)*, pages 18–33. Springer, 2016.
- 12 F. Kong, H. Yin, and N. Nakagoshi. Using GIS and landscape metrics in the hedonic price modeling of the amenity value of urban green space: A case study in Jinan City, China. *Landscape and Urban Planning*, 79(3-4):240–252, 2007.
- 13 J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, 1957.
- 14 D. E. Pataki, M. M. Carreiro, J. Cherrier, N. E. Grulke, V. Jennings, S. Pincetl, R. V. Pouyat, T. H. Whitlow, and W. C. Zipperer. Coupling biogeochemical cycles in urban environments: Ecosystem services, green solutions, and misconceptions. *Frontiers in Ecology and the Environment*, 9(1):27–36, 2011.
- 15 U. Sandström, P. Angelstam, and G. Mikusiński. Ecological diversity of birds in relation to the structure of urban green space. *Landscape and Urban Planning*, 77(1-2):39–53, 2006.

13:16 A Network Flow Model for the Analysis of Green Spaces in Urban Areas

- 16 C. Sister, J. Wolch, and J. Wilson. Got green? Addressing environmental justice in park provision. *GeoJournal*, 75(3):229–248, 2010.
- 17 A. F. Taylor, A. Wiley, F. E. Kuo, and W. C. Sullivan. Growing up in the inner city: Green spaces as places to grow. *Environment and Behavior*, 30(1):3–27, 1998.
- 18 L. Tyrväinen, K. Mäkinen, and J. Schipperijn. Tools for mapping social values of urban woodlands and other green areas. *Landscape and Urban Planning*, 79(1):5–19, 2007.
- 19 J. R. Wolch, J. Byrne, and J. P. Newell. Urban green space, public health, and environmental justice: The challenge of making cities ‘just green enough’. *Landscape and Urban Planning*, 125(Supplement C):234–244, 2014.

Continuous Obstructed Detour Queries

Rudra Ranajee Saha

Department of CSE, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
darklord.saha@gmail.com

Tanzima Hashem

Department of CSE, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
tanzimahashem@cse.buet.ac.bd

Tasmia Shahriar

Department of CSE, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
shahriartasmia@gmail.com

Lars Kulik

Dept of CIS, University of Melbourne, Melbourne, Australia
lkulik@unimelb.edu.au

Abstract

In this paper, we introduce Continuous Obstructed Detour (COD) Queries, a novel query type in spatial databases. COD queries continuously return the nearest point of interests (POIs) such as a restaurant, an ATM machine and a pharmacy with respect to the current location and the fixed destination of a moving pedestrian in presence of obstacles like a fence, a lake or a private building. The path towards a destination is typically not predetermined and the nearest POIs can change over time with the change of a pedestrian's current location towards a fixed destination. The distance to a POI is measured as the summation of the obstructed distance from the pedestrian's current location to the POI and the obstructed distance from the POI to the pedestrian's destination. Evaluating the query for every change of a pedestrian's location would incur extremely high processing overhead. We develop an efficient solution for COD queries and verify the effectiveness and efficiency of our solution in experiments.

2012 ACM Subject Classification Information systems → Location based services

Keywords and phrases Obstacles Continuous Detour Queries Spatial Databases

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.14

Acknowledgements I This research was partially supported under the Australian Research Council's Discovery Projects funding scheme (project number DP180103332).

1 Introduction

Efficient processing of location-based queries in the presence obstacles like a river, a fence or a private property has become an important research area in recent years. Obstructed space is different from road networks and the Euclidean space, which ignore the obstacles in the space. It is not possible to adapt the query processing algorithms for the Euclidean space or road network settings to the obstructed space as the presence of obstacles brings new challenges for processing location-based queries in real time. Considering the importance of the applications of obstructed location-based queries for pedestrians, in the last few years, researchers have developed solutions [1, 5, 16, 20] for variant location-based queries in the obstructed space that were previously addressed in the Euclidean space or road networks.



© Rudra Ranajee Saha, Tanzima Hashem, Tasmia Shahriar, and Lars Kulik;
licensed under Creative Commons License CC-BY

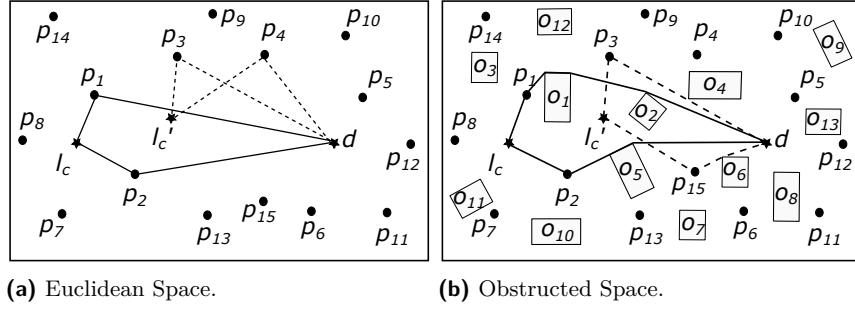
10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 14; pp. 14:1–14:16

Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



■ **Figure 1** An Example of a Continuous Detour Query for $k = 2$.

We introduce a Continuous Obstructed Detour (COD) query that allows a moving pedestrian continuously monitor the POI with the smallest obstructed detour distance, which is measured as the summation of distances from the user’s current location to the POI, and from the POI to the user’s destination by avoiding the obstacles. For example, a tourist enjoying a scenic view may not follow a predetermined walking path and instead want to visit a restaurant or a souvenir shop before arriving at the hotel. A pedestrian roaming around the city may want to buy a medicine from a pharmacy before she goes to her usual bus stop to home. In both scenarios, users have fixed destinations but do not have a predetermined path to reach the destination, and need to visit a POI before reaching the destinations.

A COD query can be extended to a $COkD$ query that continuously returns k POIs with the k smallest detour distances for a moving user heading towards a fixed destination. Figure 1 shows an example of a continuous detour query for $k = 2$ in both Euclidean and obstructed space. The Euclidean distance is measured as the length of the direct line connecting two locations. In Figure 1(a), when a user is at l_c , POIs p_1 and p_2 are the 1st and 2nd nearest detour POIs based on the Euclidean distances. When the user moves to l'_c , the answer changes, and p_3 and p_4 become the 1st and 2nd nearest detour POIs based on the Euclidean distances. In Figure 1(b), the obstacles are shown using rectangles. The obstructed distance is the length of the shortest path between two locations without crossing any obstacle. Figure 1(b) shows that p_2 and p_1 are the 1st and 2nd obstructed nearest detour POIs when the user is at l_c . When the user moves to l'_c , the answer changes, and p_{15} and p_3 become the 1st and 2nd obstructed nearest detour POIs.

The $COkD$ query cannot be modeled and processed as a continuous obstructed nearest neighbor (POI) query because of the presence of a destination. Although the obstructed distance of a POI to the destination is constant, it differs for multiple POIs, and the obstructed nearest detour POI is determined with respect to both current location and destination of the moving pedestrian. Hence, the solution [10] for moving k nearest neighbor (kNN) queries in the obstructed space is not applicable for $COkD$ queries.

Since the path to reach the destination is not predefined, for a $COkD$ query, the obstructed nearest detour POIs need to be re-evaluated in real time with respect to every changed location and the destination location of the moving user. Thus, a $COkD$ query can be processed with the repeated evaluation of obstructed k detour (OkD) queries, where an OkD query returns k obstructed nearest detour POIs with respect to a user’s current location and destination. Researchers have proposed obstructed k group nearest neighbor ($OkGNN$) algorithms [15, 16] that return k POIs having k smallest obstructed aggregate distances with respect to multiple query locations. An $OkGNN$ query is same as an obstructed k detour (OkD) query when the number of query location is two. However, the straightforward

application of the OkD algorithm for processing a $COkD$ query is not feasible as it would incur extremely high processing overhead, specially in the obstructed space as the computation of the number of obstructed distance increases with the increase of the number of the query re-evaluation. The search for the obstructed nearest detour POIs independently using OkD queries accesses the same POIs and obstacles multiple times. Thus, the major challenges for processing a $COkD$ query efficiently is to reduce the frequency of the query re-evaluation and the retrieval of the same POIs and obstacles from the database.

To address the challenges for a $COkD$ query, we develop a safe region [10, 13] based solution that avoids the re-evaluation of the query as much as possible. The key idea of our safe region based approach is to retrieve the obstructed nearest detour POIs from a database with respect to a user's current location and destination, and then identify the regions, obstructed integrated safe region (OISR) and obstructed safe regions (OSRs) with respect to the retrieved POIs. We exploit geometric properties to compute such regions. If a user resides in the OISR, the user's movement does not change the order of already retrieved k obstructed nearest detour POIs. Thus, the computation of an OISR avoids the re-computation of the query answer. If a user leaves an OISR, we compute obstructed safe regions (OSRs) with respect to the retrieved POIs to check whether new POIs are required to be retrieved from the database. Computation of OSRs allows us to avoid the retrieval of the same POIs multiple times, which in turn decreases the number of same obstacles retrieved for computing the obstructed distances of the POIs.

To further improve the efficiency of our approach, we propose two algorithms: a single point retrieval method (SPRM) and a multiple point retrieval method (MPRM), to retrieve new POIs from the database when a moving user leaves the current safe region. The aim of SPRM and MPRM is to refine the POI search space, i.e., reduce the number of the retrieval of new POIs, for identifying k obstructed nearest detour POIs with respect to the current location l_c and destination d of a moving user. A smaller number of retrieved POIs reduces the computational overhead and I/O cost for retrieving obstacles from the database.

The key difference between SPRM and MPRM is that SPRM incrementally retrieves obstructed nearest detour POIs with respect to the first location and the destination of the moving user (e.g., l_c and d in Figure 1), whereas for MPRM the obstructed nearest detour POIs are retrieved with respect to few of the current locations and destination of the moving user (e.g., l_c and d , and l_c' and d in Figure 1). SPRM does not retrieve the same POI multiple times but may retrieve additional POIs, whereas MPRM reduces the retrieval of additional POIs in return of increasing the number of obstructed distance computations with respect to multiple locations (e.g., l_c and l_c' in Figure 1).

We summarize our key contributions below:

- We introduce and formulate $COkD$ queries in spatial databases that allow pedestrians to monitor the nearest detour POIs in the presence of obstacles.
- We develop an efficient safe-region based solution for processing $COkD$ queries. To the best of our knowledge, we are the first to address the problem of $COkD$ queries.
- We develop two algorithms, SPRM and MPRM, to refine the POI search space and retrieve new POIs in the refined search space.
- We perform extensive experiments using a real data set to show the efficiency and effectiveness of our proposed solution.

■ **Table 1** A List of Symbols.

Notation	Description	Notation	Description
k	The number of required nearest detour POIs	x	The number of auxiliary POIs
l_c	The current location	d	The destination
l_f	The location from where a user starts to move	o_i	An obstacle
l_s	The location used to compute safe regions	O	The set of obstacles
z	The $(k+x)^{th}$ nearest POI of l_s	P	The set of all POIs
p_i	A POI	L	The list (set) of $(k+x)$ obstructed nearest detour POIs
A	The set of k obstructed nearest detour POIs for l_c and d	T_p	POI R-tree
$d_e(p, q)$	Euclidean distance between p and q	$d_{\Delta}(p, q)$	Obstructed distance between p and q
$s_e(a, b, c)$	Summation of $d_e(a, b)$ and $d_e(b, c)$	$s_{e\Delta}(a, b, c)$	Summation of $d_e(a, b)$ and $d_{\Delta}(b, c)$
$s_{\Delta}(a, b, c)$	Summation of $d_{\Delta}(a, b)$ and $d_{\Delta}(b, c)$	T_o	Obstacle R-tree

2 Problem Formulation

In a $COkD$ query, initially, a moving user provides her current location l_c , a destination d and the number k of desired nearest (detour) POIs. Later the moving user periodically updates her current location l_c . The obstructed space may include obstacles like buildings, parks, lakes, etc. An obstructed path is calculated as the shortest path between two points in the obstructed space, where a path does not intersect the interior of an obstacle. The obstructed distance between two points is the length of the obstructed path between those points. The obstructed detour distance $s_{\Delta}(l_c, p_i, d)$ of a POI p_i is measured as the summation of the obstructed distances from p_i to l_c and p_i to d . Similar to existing work in the obstructed space [16, 15], the POIs and obstacles are indexed using two separate R -trees [7], POI R -tree and obstacle R -tree in the database. Table 1 summarizes the symbols used in the paper.

A $COkD$ query is formally defined as follows.

► **Definition 1. A Continuous Obstructed k Detour Query:** Given a set of POIs P and a set of obstacles O , the current location l_c of a moving user, a destination d , and the required number of the obstructed nearest detour POIs k , a $COkD$ query returns A , a set of k obstructed nearest detour POIs that have k smallest obstructed detour distances with respect to every instance of l_c and d , i.e., $s_{\Delta}(l_c, p_i, d) \leq s_{\Delta}(l_c, p_j, d)$ for $p_i \in A$ and $p_j \in P - A$.

3 Related Work

Efficient approaches have been proposed in the literature for variants of spatial queries in the obstructed space. Processing spatial queries in the presence of obstacles has been first addressed in [19]. In [6, 17, 19], the authors developed algorithms to find the nearest POIs with respect to a static location in the obstructed space. In [5], the authors developed an algorithm to process continuous obstructed nearest neighbor queries. In [4] and [1], the authors developed solutions for efficient processing of obstructed reverse nearest neighbor queries and obstructed optimal sequenced route queries, respectively. In [20], the authors addressed obstructed range nearest neighbor queries. Obstructed group nearest neighbor (OGNN) queries that return a POI with the minimum obstructed aggregate distance have been addressed in [15, 16]. An OGNN query transforms to an obstructed detour query if the number of query location is two (i.e., a user's current location and destination). This paper focuses on the $COkD$ query, which is different from all of the above mentioned queries.

In [19], the authors proposed the first algorithm to compute the obstructed distance between two locations. Instead of directly applying the obstructed distance computation algorithm between two locations, in [16], the authors developed an algorithm to efficiently compute multiple obstructed distances with respect to a single point without retrieving same obstacles multiple times. To compute the obstructed detour distance, we need to compute two obstructed distances from a common POI, and thus, we use the algorithm in [16].

Continuous nearest neighbor queries [3, 11] and continuous detour queries [12, 14, 18] have been addressed in road networks that ignore the presence of the obstacles. In [12], the authors proposed an incremental approach using a shortest path tree to process continuous detour queries in the road network. In [14, 18], the authors developed a solution for detour queries with an assumption that a user travels in a predetermined path towards a destination. In CO k D queries, a pedestrian's path towards a destination is not known before and can be obstructed by the obstacles.

Researchers have already shown that computing safe regions can significantly reduce the query processing overhead for processing moving nearest neighbor queries [8, 10, 13]. However, none of these approaches take the destination into account, and thus, the computed safe regions are not applicable for a CO k D query, where a pedestrian moves towards a fixed destination.

4 Safe Regions

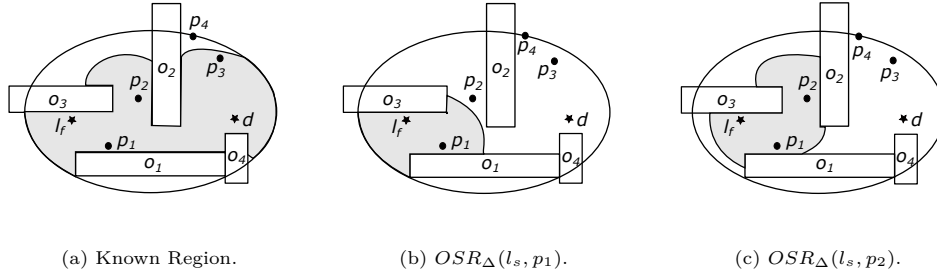
We develop a safe region based approach for processing CO k D queries. The underlying idea is to identify the safe regions based on already retrieved POIs, *obstructed integrated safe region (OISR)* and the intersection of *obstructed safe regions (OSRs)*, where the query answer does not change and any new POI does not need to be retrieved from the database for a moving user, respectively. These regions can help us to reduce the computational overhead and the retrieval of same POIs multiple times from the database. The larger the safe regions, the smaller is the number of times POIs need to be retrieved from the database. Considering this issue, we retrieve auxiliary POIs in addition to the required number (k) of POIs with an intuition that additional POIs can reduce the processing overhead. The number of auxiliary POIs x is decided in experiments.

Suppose that L is a set of ordered $k + x$ obstructed nearest detour POIs that have been retrieved from the database with respect to a moving user's location l_s and a fixed destination d for $x \geq 0$. An OSR of a retrieved POI represents the area, where a user's movement cannot incur another POI that has not yet been retrieved from the database to have a smaller obstructed detour distance than the retrieved POI. Thus, additional POIs are not retrieved from the database if a user moves inside the intersection of the OSRs of the retrieved POIs. An OISR represents an area where the current CO k D answer for a moving user does not change. To compute the OISR, in addition to OSRs we need to know the obstructed fixed rank region (OFRR) that represents the area where a user's movement does not change the relative ranking (based on the obstructed detour distance) of the retrieved POIs in L .

In Sections 4.1 and 4.2, we show how the presence of a fixed destination d makes the computation of OSRs and OFRR different from the existing OSR and OFRR computation techniques for obstructed nearest neighbor queries [10]. In Section 4.3, we combine OSRs and OFRR to compute an OISR.

4.1 Obstructed Safe Region (OSR)

Let z represent the POI that has the $(k + x)^{th}$ smallest obstructed detour distance with respect to l_s and d . Based on the retrieved $k + x$ obstructed nearest POIs with respect to l_s and d , we first define the obstructed known region: a set of points that have equal or smaller obstructed detour distances than that of z with respect to l_s and d . Figure 2(a) shows an obstructed known region for $k = 2$ and $x = 1$, where p_1 , p_2 , and p_3 have been retrieved as $k + x$ obstructed nearest detour POIs with respect to l_f and d . Note that l_f is the location from where a user starts to move, and l_s is the location used to compute safe regions. Thus,



■ **Figure 2** (a) Known Region, and (b)-(c) OSRs.

the first time when a safe region is computed, both l_f and l_s point to the same location. In all figures, we only show l_f and we assume that the safe regions are computed for the first time and l_s points to l_f .

Let p_o be a POI located outside the obstructed known region that has not yet been retrieved from the database. For a POI p_i in the obstructed known region, the obstructed safe region with respect to p_i , denoted by $OSR_{\Delta}(l_s, p_i)$, is defined as follows:

$$\begin{aligned} OSR_{\Delta}(l_s, p_i) &= \{l | s_{\Delta}(l, p_i, d) \leq s_{\Delta}(l, p_o, d)\} \\ &= \{l | d_{\Delta}(l, p_i) + d_{\Delta}(p_i, d) \leq d_{\Delta}(l, p_o) + d_{\Delta}(p_o, d)\} \end{aligned} \quad (1)$$

Here l refers to a point location. Thus $OSR_{\Delta}(l_s, p_i)$ is a set of points, where each point l satisfies $s_{\Delta}(l, p_i, d) \leq s_{\Delta}(l, p_o, d)$.

From the definition of the known region, $d_{\Delta}(l_s, p_o) + d_{\Delta}(p_o, d) \geq d_{\Delta}(l_s, z) + d_{\Delta}(z, d)$. Rearranging we have, $d_{\Delta}(l_s, p_o) \geq d_{\Delta}(l_s, z) + d_{\Delta}(z, d) - d_{\Delta}(p_o, d)$. On the other hand, according to the triangular inequality, $d_{\Delta}(l_s, l) + d_{\Delta}(l, p_o) \geq d_{\Delta}(l_s, p_o)$. By rearranging and replacing $d_{\Delta}(l_s, p_o)$ with its tighter bound, we have the tighter bound of $d_{\Delta}(l, p_o)$ as $d_{\Delta}(l_s, z) + d_{\Delta}(z, d) - d_{\Delta}(p_o, d) - d_{\Delta}(l_s, l)$.

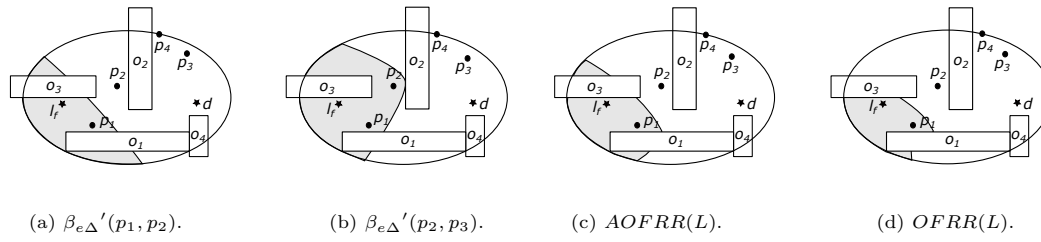
In Equation 1, if we can guarantee that $(d_{\Delta}(l, p_i) + d_{\Delta}(p_i, d))$ is less than or equal to a tighter bound of $(d_{\Delta}(l, p_o) + d_{\Delta}(p_o, d))$, i.e., $d_{\Delta}(l_s, z) + d_{\Delta}(z, d) - d_{\Delta}(l_s, l)$, then $d_{\Delta}(l, p_i) + d_{\Delta}(p_i, d) \leq d_{\Delta}(l, p_o) + d_{\Delta}(p_o, d)$ is satisfied. Thus, we can redefine $OSR_{\Delta}(l_s, p_i)$ as follows:

$$\begin{aligned} OSR_{\Delta}(l_s, p_i) &= \{l | d_{\Delta}(l, p_i) + d_{\Delta}(p_i, d) \leq d_{\Delta}(l_s, z) + d_{\Delta}(z, d) - d_{\Delta}(l_s, l)\} \\ &= \{l | d_{\Delta}(l, p_i) + d_{\Delta}(l_s, l) \leq d_{\Delta}(l_s, z) + d_{\Delta}(z, d) - d_{\Delta}(p_i, d)\} \end{aligned} \quad (2)$$

Figures 2(b) and 2(c) show OSRs for p_1 and p_2 , respectively for the same example shown in Figure 2(a). According to Equation 2, if a moving user's current location l_c satisfies $d_{\Delta}(l_s, l_c) + d_{\Delta}(l_c, p_i) \leq d_{\Delta}(l_s, z) + d_{\Delta}(z, d) - d_{\Delta}(p_i, d)$, then the user is inside the OSR of p_i , $OSR_{\Delta}(l_s, p_i)$, and any POI p_o outside the obstructed known region cannot have a detour distance smaller than that of p_i with respect to l_c and d . If the user's current location l_c is inside the intersection of the OSRs for all $(k+x)$ POIs in the obstructed known region, i.e., $\bigcap_{i=1}^{k+x} OSR_{\Delta}(l_s, p_i)$, then it is guaranteed that any POI p_o outside the obstructed known region cannot have a detour distance smaller than those for $(k+x)$ POIs in the obstructed known region with respect to l_c and d .

4.2 Obstructed Fixed Rank Region (OFRR)

The OFRR represents an area where the ranking of k obstructed nearest detour POIs in L does not change. We compute an OFRR using the concept of a dominant region. In [10],



■ **Figure 3** (a)-(b) Dominant Regions, (b) Approximate OFRR (c), and (d) OFRR (Shaded Areas).

for a moving obstructed nearest POI query, an obstructed dominant region of POI p_i over POI p_j is defined as $\beta_{\Delta}(p_i, p_j) = \{l | d_{\Delta}(l, p_i) \leq d_{\Delta}(l, p_j)\}$. We modify the definition of a dominant region for a COkD query as follows:

$$\beta_{\Delta}(p_i, p_j) = \{l | s_{\Delta}(l, p_i, d) \leq s_{\Delta}(l, p_j, d)\} \quad (3)$$

For a COkD query, an OFRR for an ordered POI set L can be computed as follows:

$$OFRR(L) = \bigcap_{i=1}^{|L|-1} \beta_{\Delta}(p_i, p_{i+1}) \quad (4)$$

To reduce the complexity of the computation of OFRRs, we first approximate a dominant region of POI p_i over POI p_j as $\beta_{e\Delta}'(p_i, p_j) = \{l | s_{e\Delta}(l, p_i, d) \leq s_{e\Delta}(l, p_j, d)\}$.

Using the approximate dominant regions, we compute the approximate OFRR (AOFRR) for L as follows:

$$AOFRR(L) = \bigcap_{i=1}^{|L|-1} \beta_{e\Delta}'(p_i, p_{i+1}) \quad (5)$$

We continue with the same example shown in Figure 2(a). Figures 3(a) and 3(b) show the approximate dominant region of p_1 over p_2 , $\beta_{e\Delta}'(p_1, p_2)$ and the approximate dominant region of p_2 over p_3 , $\beta_{e\Delta}'(p_2, p_3)$, respectively.

After computing the $AOFRR(L)$ using Equation 5, we identify the non visible region inside $AOFRR(L)$ for every POI in L . Let NVR_i be a non visible region for a POI p_i and NVR be the union of non visible regions with respect to all POIs in L . Thus, an OFRR for an ordered POI set L can be computed from the approximated OFRR as follows:

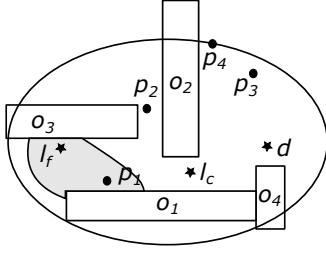
$$OFRR(L) = AOFRR(L) - NVR \quad (6)$$

Figures 3(c) and 3(d) show $AOFRR(L)$ and $OFRR(L)$, respectively, where $L = \{p_1, p_2, p_3\}$. $AOFRR(L)$ (shaded area) in Figure 3(c) is computed as the intersection areas between the shaded areas, $\beta_{e\Delta}'(p_1, p_2)$ and $\beta_{e\Delta}'(p_2, p_3)$, in Figures 3(a) and 3(b), respectively. $OFRR(L)$ in Figure 3(d) is computed by removing the nonvisible regions of p_1 , p_2 , and p_3 from $AOFRR(L)$, i.e., $OFRR(L) \subseteq AOFRR(L)$.

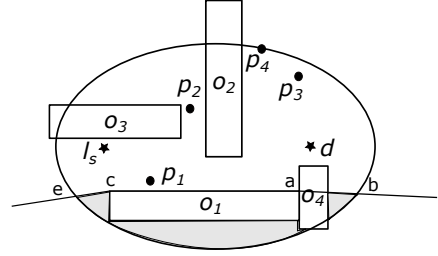
4.3 Obstructed Integrated Safe Region (OISR)

The obstructed integrated safe region, denoted by OISR, is the area, where a user's movement does not change the COkD query answer. It is the intersection of OSR and OFRR. Formally the OISR can be defined as follows:

$$OISR(l_s, L) = OFRR(L) \cap \bigcap_{i=1}^k OSR_{\Delta}(l_s, p_i) \quad (7)$$



■ **Figure 4** OISR.



■ **Figure 5** Non Visible Region for p_1 .

The shaded area in Figure 4 shows the OISR for the same example shown in Figure 2(a). However, computing intersections of safe regions for every POI is expensive. The following theorem shows that the intersection of $OFRR(L)$ and $OSR_{\Delta}(l_s, p_k)$ is enough to generate $OISR(l_s, L)$.

► **Theorem 2.** *Given a set of retrieved ordered POIs L with respect to a moving user's locations l_s and d , the obstructed safe region $OSR_{\Delta}(l_s, p_i)$ for every i^{th} nearest POI p_i of l_s in L , the obstructed fixed rank region $OFRR(L)$, then $OFRR(L) \cap \bigcap_{i=1}^k OSR_{\Delta}(l_s, p_i) = OFRR(L) \cap OSR_{\Delta}(l_s, p_k)$.*

Proof. Suppose l_c is a location in $OFRR(L) \cap \bigcap_{i=1}^k OSR_{\Delta}(l_s, p_i)$. Since l_c is a location inside $OFRR(L)$, for $i \in [1..k-1]$, the following equation also holds:

$$d_{\Delta}(l_c, p_i) + d_{\Delta}(p_i, d) \leq d_{\Delta}(l_c, p_k) + d_{\Delta}(p_k, d) \quad (8)$$

Since $l_c \in OSR_{\Delta}(l_s, p_i)$, from Equation 2, we have

$$d_{\Delta}(l_c, p_i) + d_{\Delta}(p_i, d) \leq d_{\Delta}(l_s, z) + d_{\Delta}(z, d) - d_{\Delta}(l_s, l_c) \quad (9)$$

Now if Equation 9 holds for location l_c and $i = k$, then according to Equation 8, Equation 9 also holds for l_c and $i \in [1..k-1]$. Thus, $OFRR(L) \cap \bigcap_{i=1}^k OSR_{\Delta}(l_s, p_i) = OFRR(L) \cap OSR_{\Delta}(l_s, p_k)$. ◀

5 Algorithms

In this section, we present our $COkD$ query processing algorithm (Algorithm 1) using safe regions computed in Section 4. The input to the algorithms are a current location l_c , a destination d , the number of required nearest detour POIs k , and the number of auxiliary POIs x . The output of the algorithm is the set of k obstructed nearest detour POIs A . Both l_c of a moving user and A are updated periodically. The algorithm uses a priority queue Q_p and lists L and L' to process a $COkD$ query. Q_p is used to store already accessed R -tree nodes and POIs. L is a set of ordered $k+x$ obstructed detour POIs with respect to l_c and d . The list L' includes POIs that are not in L but have been retrieved from the database for finding $k+x$ obstructed nearest detour POIs.

Algorithm 1 starts with initializing l_f and l_s as l_c , where l_f is a moving user's start location and l_s is a location used to compute the last safe regions. Then the algorithm retrieves $k+x$ obstructed nearest detour POIs with respect to l_f and d using the function *RetrievePOIs* in L (Line 2), adds first k obstructed nearest detour POIs in L to A (Line 3), and sends A to the

Algorithm 1 COkD_Process.

Input: l_c, d, k, x **Output:** A

```

1:  $l_f, l_s \leftarrow l_c$ 
2:  $L \leftarrow \text{RetrievePOIs}(l_f, d, k, x)$ 
3:  $A \leftarrow \text{FindAnswer}(l_c, d, k)$ 
4:  $\text{Send}(A)$ 
5: for every update of  $l_c$  do
6:    $\text{flagOISR}, L \leftarrow \text{CheckOISR}(l_s, l_c, d, k, x, L)$ 
7:   if  $\text{flagOISR} = 1$  then
8:      $\text{Send}(A)$ 
9:   else
10:     $\text{flagOSR} \leftarrow \text{CheckOSRs}(l_s, l_c, d, k, x, L)$ 
11:    if  $\text{flagOSR} = 0$  then
12:       $L \leftarrow \text{RetrieveNextPOIs}(l_f, l_c, d, k, x, L)$ 
13:       $l_s \leftarrow l_c$ 
14:    end if
15:     $A \leftarrow \text{FindAnswer}(l_c, d, k)$ 
16:     $\text{Send}(A)$ 
17:  end if
18: end for

```

user (Line 4). The function *RetrievePOIs* incrementally retrieves Euclidean nearest detour POIs with respect to l_c and d from the database until $k + x$ obstructed nearest detour POIs for l_c and d have been identified. After every update of the current location l_c , the algorithm checks whether the current location l_c is in $OISR(l_s, L)$ using the function *CheckOISR*. The function returns 1 if $l_c \in OISR$, 0 otherwise. The steps of the function *CheckOISR* are discussed in Section 5.1.

If the function *CheckOISR* returns 1 (i.e., $l_c \in OISR$), then Algorithm 1 sends A to the user without any further computation (Lines 7-8). On the other hand, if the function *CheckOISR* returns 0 (i.e., $l_c \notin OISR$), then Algorithm 1 checks whether l_c is in the intersection of OSRs of POIs $\{p_1, p_2, \dots, p_k\}$ in L using the function *CheckOSRs* (Line 10). The function checks the condition stated in the last line of Equation in 2 to determine whether $l_c \in OSR(p_i)$ of a POI p_i . If the condition is false for the OSR of any POI in $\{p_1, p_2, \dots, p_k\}$, the function returns 0. If the condition is true for all POIs, then the function returns 1, i.e., l is in the intersection of OSRs of POIs $\{p_1, p_2, \dots, p_k\}$ in L .

If $\text{flagOSR} = 1$, then the algorithm does not need to retrieve any new POI. If $\text{flagOSR} = 0$, then the algorithm retrieves $k + x$ nearest detour POIs in L with respect to l_c and d using the function *RetrieveNextPOIs* (Line 12). For the function *RetrieveNextPOIs*, we develop two efficient methods: SPRM (Section 5.2) and MPRM (Section 5.3) with the aim to minimize the number of the retrieval of POIs for finding $k + x$ nearest detour POIs with respect to l_c and d . After retrieving new POIs using the function *RetrieveNextPOIs*, l_s is updated as l_c (Line 13). Finally, Algorithm 1 adds first k obstructed nearest detour POIs in L to A from L and sends A to the user (Lines 15-16).

Algorithm 2 CheckOISR.

Input: l_s, l_c, d, k, x, L **Output:** $flagOISR$ and L

```

1:  $NVR \leftarrow ComputeNVR(L, l_s, d)$ 
2: if  $l_c \in NVR$  then
3:   return  $0, L$ 
4: end if
5:  $flag, L \leftarrow CheckPOIOrder(L, l_c, d)$ 
6: if  $flag = 1$  then
7:   return  $0, L$ 
8: else
9:   return  $l_c \in OSR(p_k), L$ 
10: end if

```

5.1 Function CheckOISR

The steps of this function is shown in Algorithm 2. The inputs to the algorithm are l_s, l_c, d, k, x , and L . The outputs are a flag $flagOISR$ and the L . From Equation 7, we know that $OISR(l_s, L)$ is the intersection of $OFRR(L)$ and $OSR(p_k)$. The function first computes non visible region NVR as the union of non visible regions with respect to all POIs in L (Line 1). if the direct path between a location and a POI is obstructed then the location is non-visible from the POI. Thus any location of a non visible region for a POI does not have a direct path to that POI. Figure 5 shows an example of non visible region (represented with two lines ab and ce) for POI p_1 with respect to obstacle O_1 by ignoring the presence of other obstacles. Non visible regions can be computed using a visibility graph [2, 9]. The vertices of a visibility graph represent POIs and corner points of the obstacles, and there is an edge between two vertices if the direct path between those vertices is not obstructed. To reduce the computational overhead, after computing a non visible region NVR_i for a POI p_i , it is stored and reused in the query evaluation process unless any new obstacle is retrieved.

Since $OFRR(L)$ can be computed as $AOFRRL(L) - NVR$ (Equation 6), if l_c in NVR then l_c is not in $OFRR(L)$. Again from Equation 7, $OISR(l_s, L) = OFRR(L) \cap OSR_\Delta(l_s, p_k)$. Thus, if l_c in NVR then l_c is also not in $OISR(l_s, L)$. In such a scenario, Algorithm 2 returns $flagOISR$ as 0 and L without any modification (Lines 2-4).

Otherwise, using the function $CheckPOIOrder$, Algorithm 2 computes obstructed detour distances of POIs in L with respect to l_c and d , and sorts the POIs in L , if the order of POIs based on computed obstructed detour distances changes (Line 5). If the order is changed, $flag$ is set to 1 and Algorithm 2 returns $flagOISR$ as 0 and updated L (Lines 6-7). Otherwise, Algorithm 2 checks whether $l_c \in OSR(p_k)$ using the condition stated in the last line of Equation in 2 and returns $flagOISR$ as 1 or 0 and L without any modification, if the condition stated in the last line of Equation in 2 is satisfied or not, respectively.

5.2 SPRM

POIs are indexed using an R -tree in the database. To identify $(k + x)$ obstructed nearest detour POIs for l_c and d , SPRM incrementally retrieves Euclidean detour POIs with respect to l_f and d , where from l_f the user starts to move towards a destination d . A priority queue Q_p stores already accessed R -tree nodes and POIs in order of the minimum Euclidean detour

■ **Table 2** Experiment Settings.

Parameter	Range	Default value	Parameter	Range	Default value
k	1-20	10	x	1-20	12
Query Range R	500-3000 units	1500 units	$ P / O $ Ratio	50-350	200

distances with respect to l_f and d . To avoid the retrieval of the same POIs multiple times and reduce I/O access, SPRM does not start the search for the POIs from the root node of the R -tree while incrementally retrieving the POIs with respect to l_f and d .

The POI search space that has been already traversed is an ellipse with foci at l_f and d and the major axis having the length equal to the Euclidean detour distance of the last retrieved POI with respect to l_f and d from Q_p . SPRM determines the current $(k+x)^{th}$ smallest obstructed detour distance of l_c and d based on the already retrieved POIs. The ellipse expands with the retrieval of new POIs from Q_p . With the retrieval of a new POI, SPRM updates the current $(k+x)^{th}$ smallest obstructed detour distance of l_c and d if it becomes smaller. The search ends when the minimum Euclidean detour distance of l_c and d from the boundary of the ellipse becomes greater than or equal to the current $(k+x)^{th}$ smallest obstructed detour distance of l_c and d .

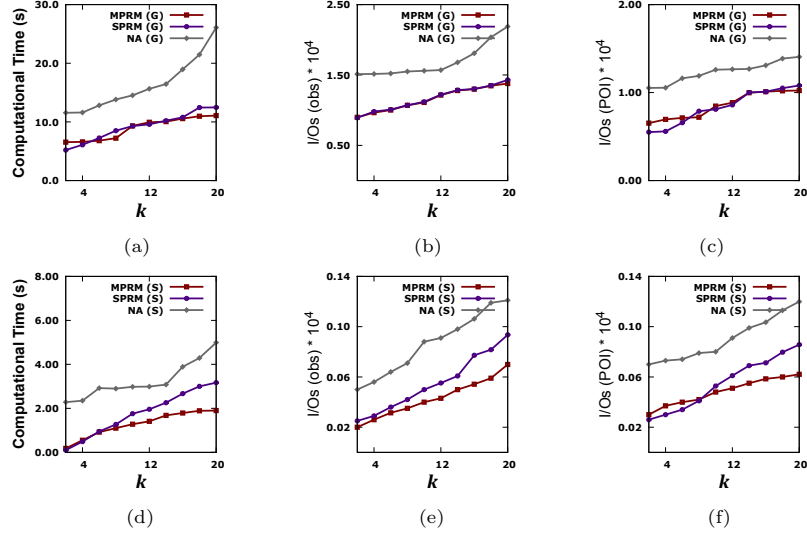
5.3 MPRM

Since SPRM expands the POI search space (i.e., ellipse) with respect to fixed locations l_f and d , some retrieved POIs may never become part of the CO k D answer with respect to the updated location l_c and d . To avoid the retrieval of those additional POIs, MPRM retrieves new POIs with respect to l_c and d instead of l_f and d . Similar to SPRM, MPRM does not start the search from the root of the POI R -tree node and reuses the already traversed nodes of the POI R -tree. However, MPRM incurs additional processing overhead for computing the minimum Euclidean detour distances with respect to l_c and d for the nodes/POIs stored in Q_p .

MPRM sorts the already retrieved POIs according to the obstructed detour distance with respect to l_c and d . Then MPRM resorts the elements in Q_p based on their Euclidean detour distances with respect to l_c and d . The algorithm continues to retrieve the next Euclidean nearest detour POI p with respect to l_c and d from Q_p as long as the Euclidean detour distance of p with respect to l_c and d is smaller than the current $(k+x)^{th}$ smallest obstructed detour distance of l_c and d based on already retrieved POIs.

6 Experiments

We present the performance of our safe region based approach using both SPRM and MPRM and compare them with a naive approach (NA) that independently finds k obstructed nearest detour POIs for every location update of a moving user using the O k D algorithm proposed in [16] (please see Section 3) for details. We use both real and synthetic data sets. The total space is normalized into $10,000 \times 10,000$ square units. The real dataset of Germany consists of 36334 Minimum Bounding Rectangles (MBRs) of railway lines (rrlines) and 76999 MBRs of hypsography data (hypos). In this dataset, end points of hypos represent POIs, and rrlines are the obstacles. Though we use MBRs to represent obstacles, our approach is applicable for obstacles of any shape. We also use the real datasets of rivers and lakes in Greece as obstacles, and generate synthetic POIs using uniform random distribution. We denote the synthetic dataset (Greece dataset) by ‘S’ and Germany dataset by ‘G’.



■ **Figure 6** Effect of the number of required POIs k .

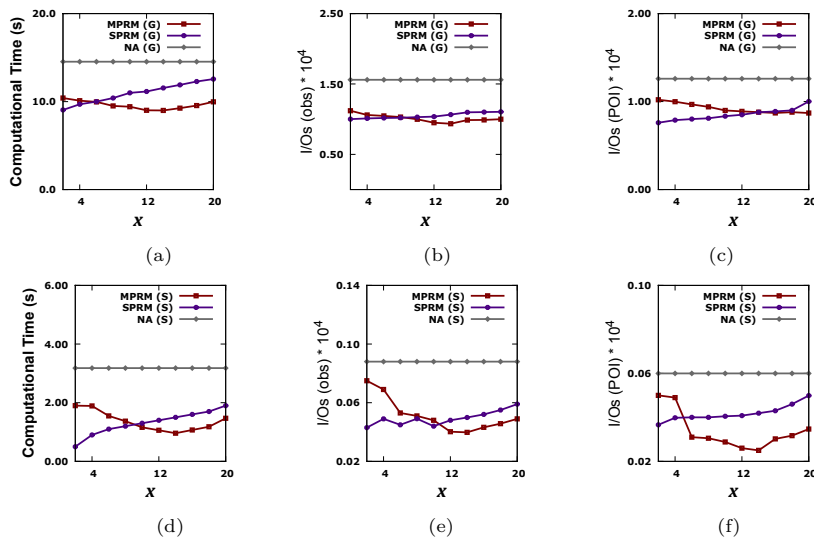
We use a 2.4 GHz Intel i5 CPU and 16 GB main memory. Table 2 shows the range and default values of our experiment parameters. To observe the effect of a parameter in an experiment, we set other parameters to their default values.

For every experiment, we consider 200 sample $COkD$ queries and takes the average performance in terms of the computational time and I/O costs for retrieving POIs and obstacles from the database. For every $COkD$ query sample, we randomly generate l_f and d according to the specified range in the experiment. Then we randomly generate l_c s in the following way: a user moves towards the direction of d but the followed path may not be the shortest one for arriving at d . Though the distance between two l_c s is kept fixed, the number of l_c s may vary for two paths having l_f and d in the same query range (e.g., 3000 units). Therefore, we show the average computational time and I/Os required per l_c for a path as the cost of a $COkD$ query sample.

6.1 Effect of the Number of Required POIs k

Figure 6 shows that the required computational time and I/Os are higher for the naive approach than those for our safe region based approach for varying k . From Figures 6(a) and 6(d), we observe that the computational time increases rapidly for the naive approach than our safe region based approach for higher values of k . This is because with the increase of k , for both SPRM and MPRM, the safe regions become larger and the probability for l_c to remain inside OISR increases, which avoids the re-computation of $COkD$ answer. On the other hand, the naive approach requires to evaluate the obstructed nearest detour POIs for every update of l_c and the time required for the evaluation increases for the higher values of k .

Figures 6(b), 6(e), 6(c) and 6(f) show that the I/O cost for both POIs and obstacles increases with the increase of k , which is expected because the number of POIs and obstacles retrieved from the database increase with the increase of k .



■ **Figure 7** Effect of the number of auxiliary POIs x .

6.2 Effect of the Number of Auxiliary POIs x

Figure 7 shows that the computational time and I/O cost for the naive approach is higher than our approach but remain same irrespective of values of x because the naive approach does not retrieve auxiliary POIs. On the other hand, we observe that for MPRM the performance improves with the increase of x upto a certain threshold then again degrades. The reason is as follows. With the increase of x , the area of safe region becomes larger and the query processing overhead decreases, but after certain threshold with the increase of x , the cost for computing the non visible regions diminishes the gain achieved from the large safe regions. For SPRM, we observe that the performance degrades with the increase of x . This is because for SPRM, POIs are always retrieved with respect to l_f , and the retrieval of POIs that are not required increases with the increase of x .

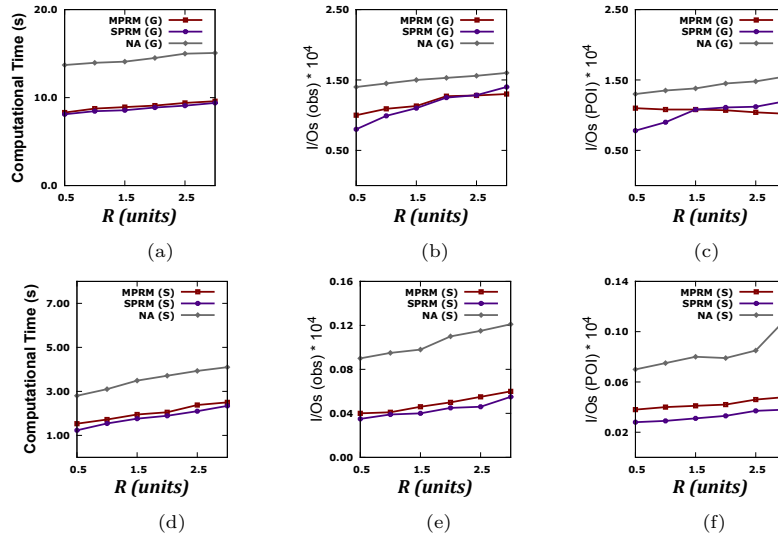
6.3 Effect of the Query Range R

In this experiment, we vary R from 500 meter to 3000 meter by considering the typical travelling distance of a pedestrian. Figure 8 shows that SPRM performs better than MPRM, which can be explained from the underlying structure of SPRM and MPRM. It is expected that set of nearest detour POIs remain same for several timestamps, and the number of POIs and obstacles retrieved with respect to l_c and d is small. On the other hand, MPRM needs to compute obstructed detour distances for every element in Q_p with respect to l_c . Therefore, SPRM performs better than MPRM.

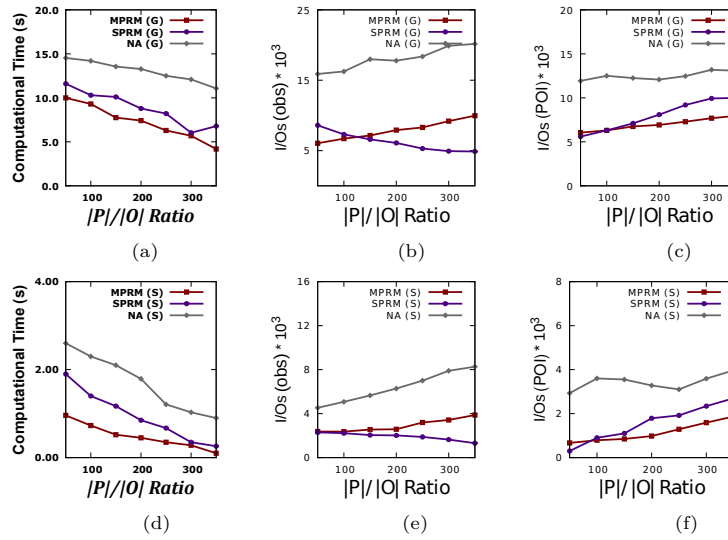
The performance of both naive and safe region based approaches degrades with the increase of R . Since the distance between consecutive l_c s increases with the increase of R , the probability that l_c falls outside the safe region also increases and more POIs and obstacles need to be retrieved from the database.

6.4 Effect of POI-Obstacle Ratio $|P|/|O|$

Figure 9 shows the comparative performance between the naive approach and the safe region based approach for varying the ratio of the number of POIs and the number of obstacles $|P|/|O|$. Increase of $|P|/|O|$ ratio means that the sample space contains more POIs than



■ **Figure 8** Effect of the query range R .



■ **Figure 9** Effect of POI-obstacle ratio $|P|/|O|$.

obstacles. With the increase of $|P|/|O|$, the I/O cost for POIs increases for both SPRM and MPRM, which is expected. For SPRM, the I/O cost of obstacles decreases because less number of obstacles are retrieved with respect to fixed locations l_f and d . However, for MPRM, the I/O cost of obstacles increases because detour obstructed distances of POIs are computed with respect to different locations.

7 Conclusion

We have introduced and formulated $COkD$ queries. We have proposed the first approach based on safe regions for efficient processing of $COkD$ queries. We have further improved the efficiency of our approach by developing two POI retrieval algorithms: SPRM and MPRM.

We have performed experiments using both real and synthetic datasets. The results show that our approach for CO k D queries with SPRM requires on average 67.3% less processing time, 62% less I/Os for obstacles and 72.6% less I/Os for POIs than the naive approach that applies the existing OkD query processing algorithm to evaluate for CO k D queries. On the other hand, our approach with MPRM requires on average 69.2% less processing time, 67% less I/Os for obstacles and 72% less I/Os for POIs than the naive approach.

References

- 1 Anika Anwar and Tanzima Hashem. Optimal obstructed sequenced route queries in spatial databases. In *EDBT*, pages 522–525, 2017.
- 2 Takao Asano, Tetsuo Asano, Leonidas J. Guibas, John Hershberger, and Hiroshi Imai. Visibility of disjoint polygons. *Algorithmica*, 1(1):49–63, 1986.
- 3 Ugur Demiryurek, Farnoush Banaei-Kashani, and Cyrus Shahabi. Efficient continuous nearest neighbor query in spatial networks using euclidean restriction. In *SSTD*, pages 25–43, 2009.
- 4 Yunjun Gao, Jiacheng Yang, Gang Chen, Baihua Zheng, and Chun Chen. On efficient obstructed reverse nearest neighbor query processing. In *SIGSPATIAL GIS*, pages 191–200, 2011.
- 5 Yunjun Gao and Baihua Zheng. Continuous obstructed nearest neighbor queries in spatial databases. In *SIGMOD*, pages 577–590, 2009.
- 6 Yu Gu, Ge Yu, and Xiaonan Yu. An efficient method for k nearest neighbor searching in obstructed spatial databases. *J. Inf. Sci. Eng.*, pages 1569–1583, 2014.
- 7 Antonin Guttman. R-trees: a dynamic index structure for spatial searching. In *SIGMOD*, pages 47–57, 1984.
- 8 Tanzima Hashem, Lars Kulik, and Rui Zhang. Countering overlapping rectangle privacy attack for moving knn queries. *Inf. Syst.*, 38(3):430–453, 2013.
- 9 Paul J. Heffernan and Joseph S. B. Mitchell. An optimal algorithm for computing visibility in the plane. *SIAM J. Comput.*, 24(1):184–201, 1995.
- 10 Chuanwen Li, Yu Gu, Jianzhong Qi, Rui Zhang, and Ge Yu. A safe region based approach to moving knn queries in obstructed space. *KAIS*, 45:417–451, 2015.
- 11 Kyriakos Mouratidis, Man Lung Yiu, Dimitris Papadias, and Nikos Mamoulis. Continuous nearest neighbor monitoring in road networks. In *VLDB*, pages 43–54, 2006.
- 12 Sarana Nutanong, Egemen Tanin, Jie Shao, Rui Zhang, and Ramamohanarao Kotagiri. Continuous detour queries in spatial networks. *IEEE TKDE*, 24:1201–1215, 2012.
- 13 Sarana Nutanong, Rui Zhang, Egemen Tanin, and Lars Kulik. The v*-diagram: a query-dependent approach to moving KNN queries. *PVLDB*, 1(1):1095–1106, 2008.
- 14 Shuo Shang, Ke Deng, and Kexin Xie. Best point detour query in road networks. In *SIGSPATIAL GIS*, pages 71–80, 2010.
- 15 Nusrat Sultana, Tanzima Hashem, and Lars Kulik. Group nearest neighbor queries in the presence of obstacles. In *SIGSPATIAL GIS*, pages 481–484, 2014.
- 16 Nusrat Sultana, Tanzima Hashem, and Lars Kulik. Group meetup in the presence of obstacles. *Inf. Syst.*, 61:24–39, 2016.
- 17 Chenyi Xia, David Hsu, and Anthony KH Tung. A fast filter for obstructed nearest neighbor queries. In *BICOD*, pages 203–215, 2004.
- 18 Jin Soung Yoo and Shashi Shekhar. In-route nearest neighbor queries. *GeoInformatica*, 9(2):117–137, 2005.
- 19 Jun Zhang, Dimitris Papadias, Kyriakos Mouratidis, and Manli Zhu. Spatial queries in the presence of obstacles. In *EDBT*, pages 366–384, 2004.

14:16 Continuous Obstructed Detour Queries

- 20 Huaijie Zhu, Xiaochun Yang, Bin Wang, and Wang-Chien Lee. Range-based obstructed nearest neighbor queries. In *SIGMOD*, pages 2053–2068, 2016.

Enhanced Multi Criteria Decision Analysis for Planning Power Transmission Lines

Joram Schito

ETH Zurich, Institute of Cartography and Geoinformation, Zurich, Switzerland
jschito@ethz.ch

Ulrike Wissen Hayek

ETH Zurich, Planning of Landscape and Urban Systems, Zurich, Switzerland
wissen@nsl.ethz.ch

Martin Raubal

ETH Zurich, Institute of Cartography and Geoinformation, Zurich, Switzerland
mraubal@ethz.ch

Abstract

The energy transition towards alternative energy sources requires new power transmission lines to connect these additional energy production plants with electricity distribution centers. For this reason, Multi Criteria Decision Analysis (MCDA) offers a useful approach to determine the optimal path of future transmission lines with minimum impact on the environment, on the landscape, and on affected citizens. As objections could deteriorate such a project and in turn increase costs, transparent communication regarding the planning procedure is required that fosters citizens' acceptance. In this context, GIS-based information on the criteria taken into account and for modeling possible power transmission lines is essential. However, planners often forget that the underlying multi criteria decision model and the used data might lead to biased results. Therefore, this study empirically investigates the effect of various MCDA parameters by applying a sensitivity analysis on a multi criteria decision model. The output of this analysis is evaluated combining a *Cluster Analysis*, a *Principal Component Analysis*, and a *Multivariate Analysis of Variance*. Our results indicate that the variability of different corridor alternatives can be increased by using different MCDA parameter combinations. In particular, we found that applying continuous boundary models on areas leads to more distinct corridor alternatives than using a sharp-edged model, and better reflects actual planning practice for protecting areas against transmission lines. Comparing the results of two study areas, we conclude that our decision model behaved similarly across both sites and, hence, that the proposed procedure for enhancing the decision model is applicable to other study areas with comparable topographies. These results can help decision-makers and transmission line planners in simplifying and improving their decision models in order to increase credibility, legitimacy, and thus practical applicability.

2012 ACM Subject Classification Information systems → Decision support systems

Keywords and phrases Geographic Information Systems, Transmission Line Planning, Multi-Criteria Decision Analysis, Sensitivity Analysis, Cluster Analysis

Digital Object Identifier 10.4230/LIPICs.GIScience.2018.15

Funding This research is financially supported by the Swiss Federal Office of Energy SFOE and by the grid operators Swissgrid, BKW Energie AG, and Austrian Power Grid. Furthermore, this research is part of the activities of the Swiss Competence Center for Energy Research on the Future Swiss Electrical Infrastructure (SCCER-FURIES), which is financially supported by the Swiss Innovation Agency (Innosuisse-SCCER program).



© Joram Schito, Ulrike Wissen Hayek, and Martin Raubal;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 15; pp. 15:1–15:16

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Multi Criteria Decision Analysis (MCDA) has been successfully applied in a large number of research projects to identify the optimal solution across a variety of conflicting criteria [12]. Regardless whether the underlying problem is spatial or not, the principle is the same, as different alternatives are compared by their utility to solve the given problem. Therefore, a decision-maker assigns each factor that contributes to the decision a value describing the utility to solve the underlying problem. Each factor is then weighted according to the decision-maker's preferences and summed up to the total utility by applying a set of decision rules [11]. Ideally, these decision rules should be based on consensus among all decision-makers to minimize the potential for post-decision regret [2].

When applying prescriptive MCDA on spatial problems, Geographic Information Systems (GIS) can be used as Decision Support Systems (DSS) to support decision-makers in identifying the best decision to take [19]. In particular, a large variety of visualization techniques has been successfully applied to support decision-making either when comparing sensitivities on maps or charts [15], or when determining pareto-optimal solutions [5, 20, 25]. Spatial decisions are taken, for example, for allocating an object to the optimal location, for evaluating the land use suitability, or for assessing a phenomenon's impact on the environment [19]. One field that strongly considers location-based factors is the planning of energy systems. The ongoing energy transition towards alternative energy sources incites national governments and companies to build new renewable energy power plants for various reasons, i.e., reliability of supply, providing cheap energy, reducing dependency, and reducing environmental impacts [24]. Consequently, the grid must be extended to connect a growing number of electricity producers with the consumers [16].

However, public acceptance of grid expansion projects is generally low [16], as transmission lines evoke opposition particularly when they are sited in rural landscapes [17]. Furthermore, land owners fear depreciation of their land value [4]. This low acceptance leads to high social resistance, which in turn raises objections, causes delays, and increases costs – all of them barriers against necessary grid expansions [1]. In order to increase acceptance, various methods have been applied or proposed so far. First, involving citizens in the decision-making process is known to foster acceptance [7]. Second, a transparent dialogue between grid operators and affected citizens can be enhanced by supporting communication with immersive virtual reality [21]. Both approaches move in the same direction, as acceptance might be increased through greater degrees of transparency in communicating the planning process to citizens. Moreover, the use of realistic virtual reality environments can support decision-makers in imagining how a transmission line could be blended into the landscape.

In this context GIS can support transparent communication and there are various examples of GIS-based DSS for determining the optimal path for transmission lines [3, 14]. The approach mostly used hereby is explained in section 2.3, which uses *spatial costs* to determine how feasible an area is for building a power line on its surface. However, the suggested corridors and paths resulting from such a DSS might be biased, as the underlying data or decision model limits the number of possible solutions and what the solutions actually reflect. With regard to transmission line planning particularly the spatial resistance against the construction of transmission lines (according to the law, etc.) and distances to spatially protected areas (e.g., nature protected areas or certain settlement zones) need to be reflected adequately. Therefore, we developed a 3D DSS and modified a standard MCDA model in a way that these aspects are taken into account. Moreover, a sensitivity analysis was conducted to proof the quality of our MCDA model.

As the effects of raster-based MCDA have been explored in prior work, we specifically investigated if a sensitivity analysis shows whether our modified MCDA model causes a systematic trend in computing the resulting suitability maps. By identifying such a trend, the corresponding parameters or parameter levels could be considered to be grouped to simplify the decision model. We further focused on the extent to which the single parameter levels contribute to the typical characteristics of a suitability map. In this respect, we assumed that in an initial procedural step decision-makers might appreciate to compare route alternatives that are clearly distinguishable. Therefore, we wanted to determine the most influential parameter levels that contribute to a wide variability of the resulting suitability maps. By doing so, stakeholders can focus their discussions on factors that essentially contribute to a specific alternative. To this end, we explore the utility of a *Cluster Analysis* in combination with a *Principal Component Analysis* and a *Multivariate Analysis of Variance* (MANOVA) for improving a decision model.

In this paper, we present the results of the sensitivity analysis and discuss how this approach supports simplifying and improving the MCDA model. Overall, we contribute to calibrating MCDA models so that they can actually assist in real world spatial planning processes to make transmission line planning faster, more reliable, and more accepted by affected citizens.

2 Method

2.1 Study areas

In accordance with our project partners *Swissgrid* and *Austrian Power Grid* we focused on the two study areas *Innertkirchen – Mettlen* in central Switzerland and *Kärnten* in southern Austria. Both areas have a similar topography, as the main settlement areas are located on a flatland on approx. 500 meters above sea level, each partially surrounded by Alpine foothills and crossed by rivers and lakes. In these areas, the legal requirements outlined in [9] oblige to successively reduce the area of interest for transmission lines. Therefore, we decided to use a general decision modeling approach similar to [14], which narrows down the area of interest in four steps: 1) from a large-scale planning area to 2) a corridor with a width of a few hundreds of meters to 3) a path and finally, to 4) the exact pylons' positions. The geodata were then represented in an interactive, online 3D Decision Support System (3D DSS).

2.2 Data preparation

In order to build a decision model, we analyzed the criteria that must be considered by law [9] and identified 33 spatially explicit factors with a legal influence against the construction of a transmission line (see tab. 1). These factors were grouped into the three categories *environmental protection*, *urban planning*, and *technical implementability*. Each of the 33 factors used in our decision model was assigned a *main objective* [11] based on the importance of the underlying legal source [8] (see tab. 1).

Based on this decision model, we collected the appropriate data from publicly accessible data portals and stored them in a database. In case a dataset was represented by point or line features, a buffer distance was assigned according to the legal requirements or expert's opinion. We further integrated two factors that foster building of new paths in areas already characterized by transmission lines, highways, or railway lines. These factors allow a decision-maker to assess bundling with existing linear infrastructure as more or less important.

■ **Table 1** Factors used in the decision model, sorted by category and main objective.

Category	Influencing factor	Main objective with code Ω
Environmental protection	Biosphere reserve Dry grassland Flood plains: high importance Inventory of protected landscapes Mire landscapes Mires Bird protection area	Preserve ecosystems: primary Ω_5
	Flood plains: low importance Forest Natural reserves Protection areas according to hunting laws	Preserve ecosystems: secondary Ω_6
	National parks UNESCO World Heritage Site	Preserve landscape: primary Ω_1
	Geotopes	Preserve landscape: secondary Ω_2
Technical implementability	Natural hazard areas	Decrease risks Ω_{10}
	Groundwater zone Inappropriate relief Water bodies	Ensure implementability Ω_8
	Infrastructure facilities	Avoid infrastructure facilities Ω_7
Urban planning	Airports	Decrease risks Ω_{10}
	Urban sprawl caused due to the grid Urban sprawl caused due to traffic routes	Increase bundling Ω_9
	Arable land	Preserve landscape: secondary Ω_2
	Areas within noise threshold of 40 dBA Residential / work / mixed areas Residential areas Cultural heritage: high importance	Preserve living space: primary Ω_3
	Cultural heritage: low importance Historic areas Historic traffic routes Public areas Recreational areas Tourism areas	Preserve living space: secondary Ω_4

Moreover, we extended our decision model with a factor that includes all areas unsuitable for constructing a transmission line. In particular, the results of a preliminary study showed that construction costs for a transmission line strongly increase for areas over 1300 meters and for areas with a slope greater than 55° .

2.3 Representing spatial resistances adequately

In collaboration with our project partners, we defined an MCDA model to compute the cost surface. In general, the corridor suitability maps and the transmission line paths were computed by combining MCDA with a *Least Cost Path* (LCP) analysis [10]. First, MCDA was applied on overlapping raster lattices with the same direction, origin, and cell size of 100 meters to obtain a *cost surface* [19]. Based on this cost surface, the LCP algorithm determined suitable corridors and the optimal transmission line path.

Further, decision-makers were deemed capable of making decisions about **resistances** and **weights** to distinguish between an interest-based assessment and the relative importance of a decision. Whereas the former represents a factor's friction against constructing a transmission line on top of the corresponding area, the latter represents the subjective importance the decision-maker assigns to this decision. Decision-makers used the direct rating method [11] to define a resistance on a Likert 5-point acceptability scale and a weight on a Likert 3-point priority scale [23]. In collaboration with the legal departments of various federal authorities we then restricted the resistance range of all factors that must comply with the hierarchy of laws [8]. For example, as wetlands are protected by the Swiss constitution, the range of possible resistances was restricted to 'unacceptable' and 'totally unacceptable'. By this, we expected to comply with factual premises in order to obtain realistic results.

In general, the total resistance t_x can be calculated for each location x by multiplying the resistance with the weight, as shown in the following equation:

$$t_x = \sum_{i=1}^n r_{i,x} \cdot w_i \quad (1)$$

where $r_{i,x}$ represents the resistance of factor i at location x and w_i the weight of factor i . However, this equation required modification for lack of consideration of special characteristics concerning the meaning of the resistance, the weight's effect on the total resistance, the behavior of overlapping pixels, and the influence of the boundary model. As such, these four modifications are explained subsequently.

Modification 1: utility function First, decision-makers might not perceive the differences between the levels of a given Likert scale equally. Strictly speaking, 'totally unacceptable' does not necessarily translate to 'twice as bad as unacceptable', even though the relative difference between the levels on the Likert scale are equal. In practice, the utility function is determined by applying different techniques when interviewing a decision-maker [11]. Therefore, we empirically defined four distinct utility functions for stretching or narrowing the relative distances between the levels on the Likert scale. By doing so, we expected the highest probability to determine whether different curve shapes, thus, utility functions have a significant effect on the result or not. Therefore, the modified resistance $u_{c,i,x}$ resulting from applying the subsequent utility functions replaces $r_{i,x}$ of eq. 1 and is defined as follows for the range from 1 to 5:

$$\forall [5 \geq r \geq 1] \rightarrow u_{1,i,x}(r_{i,x}) = r_{i,x} \quad (2)$$

$$\forall [5 \geq r \geq 1] \rightarrow u_{2,i,x}(r_{i,x}) = \frac{0.575}{\sqrt{|r_{i,x} - 3| + 1}} \cdot 3(r_{i,x} - 3) + 3 \quad (3)$$

$$\forall [5 \geq r \geq 1] \rightarrow u_{3,i,x}(r_{i,x}) = \sqrt{6 \cdot r_{i,x} - 5} \quad (4)$$

$$\forall [5 \geq r \geq 1] \rightarrow u_{4,i,x}(r_{i,x}) = \frac{r_{i,x}^2}{6} - \frac{5}{6} \quad (5)$$

The utility function described by eq. 2 is linear and does not apply any corrections on the chosen resistance. In contrast, eq. 3 enhances the effect of the resistances the more they differ from the mid neutral value. Finally, eq. 4 applies a logarithmic correction whereas eq. 5 uses an exponential correction for increasing aversion against constructing a transmission line. All utility functions are shown in fig. 1.

Modification 2: weighting model Due to its unipolar character, the application of eq. 1 leads to higher total resistances the higher the weights are. As decision-makers assessed the suitability of a factor on a bipolar range from ‘totally acceptable’ to ‘totally unacceptable’, they would expect lower total costs when applying a high weight on a low resistance instead of a low weight. Consequently, three weighting models were defined that enhance the effect of the chosen resistance r the higher the weight is. Additionally, we defined that our models must not overlap that is, a weight of 1 on the most extreme resistance (either 1 or 5) always leads to a more pronounced total value than applying a higher weight on a less pronounced resistance. Furthermore, we specified that the effect of the weighting model should, on the one hand, not be too extreme and, on the other hand, balanced between accepting and dismissing resistances. Thus, the modified weight $h_{b,i}$ resulting from applying the subsequent empirically defined weighting models, replaces w_i of eq. 1:

$$\forall [5 \geq r \geq 3] \rightarrow h_{1,i}(w_i) = \sqrt[3]{w_i} \quad \text{and} \quad \forall [3 > r \geq 1] \rightarrow h_{1,i}(w_i) = \sqrt{\frac{1}{w_i}} \quad (6)$$

$$\forall [5 \geq r \geq 3] \rightarrow h_{2,i}(w_i) = \sqrt[10]{w_i} \quad \text{and} \quad \forall [3 > r \geq 1] \rightarrow h_{2,i}(w_i) = \sqrt{\frac{1}{w_i}} \quad (7)$$

$$\forall [5 \geq r \geq 1] \rightarrow h_{3,i}(r_{i,x}, w_i) = r + \frac{\text{sgn}(r) \cdot (w_i - 1)}{4} \quad (8)$$

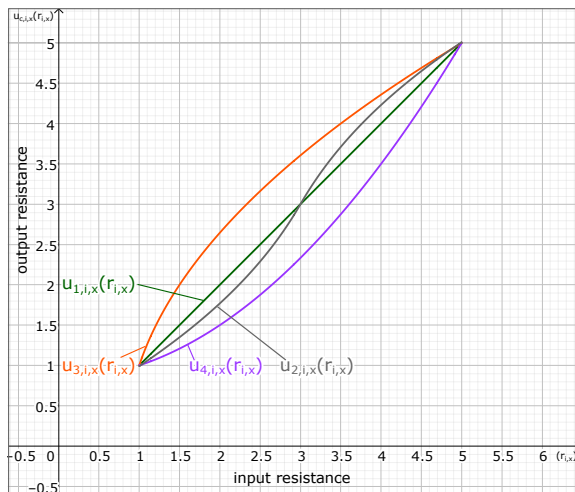
The weighting models of eq. 6 and eq. 7 are similar because they only differ in the chosen order of the root. Since the chosen weights must equally affect the decision of supporting or avoiding the construction of a transmission line, it follows that they had to be defined differently for negative and for positive resistances. In contrast, eq. 8 simply adds or subtracts 0.25 or 0.5 to or from the resistance, depending on the resistance’s sign and on the weight.

Modification 3: MCDA method The situation may arise that an area A defined in one dataset partially or completely overlaps with an area B of another dataset. A reason for this could be that A or parts of it may be listed in different protection inventories. As inventories are often based on different laws, it becomes more difficult to construct a transmission line in an area that is part of different inventories, as it is protected by various laws. From this perspective, the question arises whether the increase in difficulty should be considered to be linear and depend on the number of according protection inventories or not. Hence, the modified resistance $u_{c,i,x}$ and the modified weight $h_{b,i}$ were included in eq. 1 and therefore defined the three MCDA methods $t_{a,x}$ in terms of the way overlapping pixels should be treated by using the following equations:

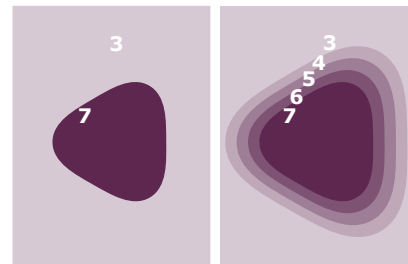
$$t_{1,x} = \sum_{i=1}^n u_{c,i,x} \cdot h_{b,i} \quad (9)$$

$$t_{2,x} = \frac{\sum_{i=1}^n u_{c,i,x} \cdot h_{b,i}}{\ln p_x + 1} \quad \forall p_x \geq 1 \quad (10)$$

$$t_{3,x} = \max_{i \in \{1, \dots, n\}} (u_{c,i,x} \cdot h_{b,i}) \quad (11)$$



■ **Figure 1** The four utility functions used to modify the resistances.



■ **Figure 2** The sharp-edged (left) and the continuous (right) boundary model.

where p_x is the number of overlapping pixels at location x . The approach used in eq. 9 is defined as *Simple Additive Weighting* [6] as it simply weights the factors and sums them up to a total resistance. In contrast, eq. 10 is an adaption of eq. 9 as it diminishes the effect of overlapping pixels by applying a logarithmic correction, aiming at reducing a potential overrating of overlapping pixels. Last, the *Maximum Value Method* described by eq. 11 chooses the maximum value of all overlapping pixels, as it is supposed to represent the strictest protection law.

Modification 4: boundary model Malczewski’s theory of fuzzy sets [19] states that fuzzy values define the grade of membership to a specific factor, leading to fuzzy boundaries. If we take Tobler’s First Law of Geography [22] into account and assume that the effect of a factor is not uniquely defined over distance, we recognize a similarity to the fuzzy sets explained above. Because protective effects do not often end at a protection area’s border, we used an approach that protects an area beyond its borders by continuously decreasing the cell resistance with increasing distance from the cell center (see the right panel of fig. 2). As an effect, the borders become fuzzy and adjacent borders may overlap (which might be corrected for instance by applying eq. 10). Consequently, protective effects are increased because the extended protection area presses – figuratively speaking – the transmission line away from the protection area. This approach complies with the current legal understanding, as greater levels of protection should be afforded to valuable locations. Furthermore, it is directly applicable to human perception, as [13] demonstrated that the visual impact of a transmission tower mainly depends on distance.

Consequently, we wanted to identify the distances that experts assign to each factor for protecting the corresponding areas according to the continuous boundary model. For this, we conducted three preliminary studies with a total of 28 participants, consisting of transmission line planning experts (n=18), representatives of federal authorities (n=7), and NGO representatives (n=3). For each of the decision model’s 33 influencing factors, experts defined the distance over which protective effects should influence the result. Furthermore, they could decide if the decreased shape should be defined linearly, logarithmically, or exponentially. This was followed by a statistical evaluation of the results and setting of the median as additional protective distance for the continuous boundary model. For each factor, we chose the linear decreasing curve, as it was always the most frequently chosen.

2.4 Sensitivity analysis

Contrary to the common approach to sensitivity analysis, in which the input factors' uncertainties are used to model the output variability, we set up a full factorial design to analyze the effect of all possible combinations between the different factor levels. Thus, the overall model consists of the **2 boundary models** (fig. 2), **3 MCDA methods** (eq. 9-11), **4 utility functions** (eq. 2-5), and **3 weighting models** (eq. 6-8), which results in 72 possible combinations. For computational reasons, the subsequent simplifications had to be applied. First, we aggregated the geometries of the decision model's 33 influencing factors according to their main objective set in the decision model. By doing this, we reduced the model's complexity to 10 factors, each representing areas with the same main objective. Moreover, we decreased complexity by limiting the number of Likert scale levels to 1 (low) and 3 (high) – for resistances as well as for weights.

According to the main objectives set in the decision model (see tab. 1), we only chose reasonable combinations by omitting combinations in which the resistance of the secondary protection objective was higher than the primary protection objective. If the resistances were equal, we only chose combinations in which the primary objective's weight was at least as high as those of the secondary objective. Similar to the approach chosen by [18], we then computed the following output files for every possible remaining combination for further analysis:

- corridor suitability maps, including the optimal path (see fig. 3)
- length over which a specific objective is violated (see tab. 4)

To compute the data, we used 48 CPUs on an Intel® Xeon® CPU E5-2680 v4 @ 2.40GHz server with 132 GB RAM by using Python's multiprocessing library. Generating the maps of all possible and reasonable settings took between 1 to 3 seconds for each map. This equated to approx. 8 days of computing time with a storage volume of approx. 4.0 TB per study area. Running the simulation for the study area in Innertkirchen – Mettlen generated $n=3'871'389$ records, while $n=3'190'344$ valid results could be generated for the study area Kärnten.

2.5 How the results were evaluated

The output parameters listed in section 2.4 including the rasters emerging from the simulation process were then sorted and statistically evaluated according to one of the 72 MCDA parameter combinations. Next, a moving average algorithm computed the mean of all rasters with the same parametrization. These 72 averaged maps were then compared to each other by determining Pearson's correlation coefficient R . The resulting correlation matrix was used to categorize the 72 parameter combinations into clusters of similar maps. For this, the *Partitioning Around Medoids* (PAM) method was applied because it defines differences by real Euclidean distances. This is similar to the model used to compute the maps, as location-based differences are represented by distances.

In order to support the evaluation, we determined the effect and the significance of the MCDA parameters' decomposed factor levels by conducting a *Multivariate Analysis of Variance* (MANOVA). For this, we first decomposed the 72 compound parameter combinations into 22 basic factor levels (see regressors in tab. 2). Since these represent explanatory variables, we used them as regressors for building the MANOVA regression model. As the variation in the suitability maps results from different parameter settings, we determined the model's principal components by applying *Principal Component Analysis* (PCA) on 3 items with orthogonal rotation. Although we determined that in both study areas eight components had eigenvalues over Kaiser's criterion of 1, we decided to use 3 principal components

■ **Table 2** Regressors used in MANOVA that represent the decomposed parameter settings in order to determine the influence of the underlying factor levels.

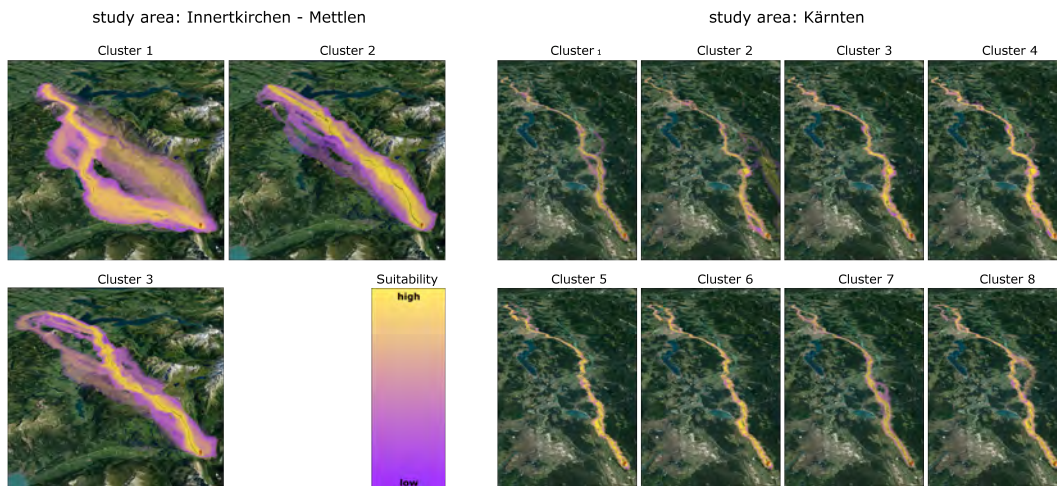
Regressor	Refers to	What the decomposed parameter might affect
β_1	fig. 2	Does the MCDA model have an influence?
β_2	fig. 2 left	Does the sharp-edged boundary model have an influence?
β_3	fig. 2 right	Does the continuous boundary model have an influence?
β_4	eq. 9-11	Does the MCDA method have an influence in general?
$\beta_5/\beta_6/\beta_7$	eq. 9/10/11	Does the MCDA method 1/2/3 have an influence?
β_8	eq. 2-5	Does the utility function have an influence in general?
$\beta_9/\beta_{10}/\beta_{11}/\beta_{12}$	eq. 2/3/4/5	Does the utility function 1/2/3/4 have an influence?
β_{13}	eq. 6-8	Does the weighting model have an influence in general?
$\beta_{14}/\beta_{15}/\beta_{16}$	eq. 6/7/8	Does the weighting model 1/2/3 have an influence?
β_{17}	interaction	Do the boundary model and the MCDA method interact?
β_{18}	interaction	Do the boundary model and the utility function interact?
β_{19}	interaction	Do the boundary model and the weighting model interact?
β_{20}	interaction	Do the MCDA method and the utility function interact?
β_{21}	interaction	Do the MCDA method and the weighting model interact?
β_{22}	interaction	Do the utility function and the weighting model interact?

in our multivariate model because inflexions on the scree plot indicated that the highest decrease of the principal components' eigenvalues occur at the 4th component. The 3 principal components explained 93.8% (Innertkirchen – Mettlen) and 88.9% (Kärnten) of the variance. Furthermore, Bartlett's test of sphericity, χ^2 (2556, N = 72) = 35341.61, $p < .001$ (Innertkirchen – Mettlen) and χ^2 (2556, N = 72) = 31764.79, $p < .001$ (Kärnten), indicated that correlations between items were sufficiently large for PCA. We therefore defined the factor loadings of the principal components as dependent variables, which should be predicted by the regressors. After conducting the MANOVA, we used the resulting *Pillai's trace* as a metric for evaluating the parameters' effect on the suitability maps.

3 Results

Surprisingly, the cluster analysis revealed a similar decision pattern in both study areas, as shown in the dendrograms in fig. 4. However, the dendrogram of the study area Innertkirchen – Mettlen was higher than the one of Kärnten, thus, the used parametrization model leads to more distinct patterns when used in Innertkirchen – Mettlen. This is also supported by analyzing the results of the PCA, as the two primary components explain 90.3% of the factor loading variability in Innertkirchen – Mettlen, whereas only 77.8% of the factor loading variability could be explained in Kärnten. By applying PAM, the k-medoids algorithm proposed as a means of grouping the suitability maps of Innertkirchen – Mettlen into 3 clusters, whereas 8 clusters were proposed for grouping the suitability maps of Kärnten (see fig. 3).

Our results reveal that the relative importance of the underlying parameters used for computing the corridor suitability maps is structured hierarchically. By ranking the regressors based on the averaged Pillai's traces among both study areas – as listed in tab. 3 – we could determine that the selection of the boundary model is most important, followed by the MCDA method, the weighting model, and last, the utility function. We will therefore detail the results using the same order.



■ **Figure 3** Suitability maps (opacity: 20%) of both study areas showing the optimal corridors for a new transmission line. According to the dendrograms of fig. 4 and read from left to right, the results are grouped into the clusters proposed by the k-medoids algorithm. Visualized with Google Earth. Yellow areas are suitable for constructing a transmission line, whereas purple areas are less suitable.

In general, the suitability maps in the study area Kärnten demonstrate higher average Pillai's traces and one significant regressor more than in Innertkirchen – Mettlen. This is because the effect of contributing to a diversification of the resulting maps must be higher the more clusters are suggested for this study area. Factors entailing the boundary model contribute most to the explanation of the model's principal components, as Pillai's traces lie between 67.6% and 99.3% with $p < .001$. Indeed, the application of different boundary models affects different solutions on a large scale. Furthermore, the dendrograms demonstrate that the choice between the sharp-edged and the continuous boundary model is most important, as this decision branched the dendrogram at the maximum height of approx. 37 for Innertkirchen – Mettlen and 17 for Kärnten.

Second, the MCDA methods contribute to the explanation of the principal components with a Pillai's trace between 47.0% and 96.2%. However, methods 1 (β_5) and 2 (β_6) explain the outcome of the resulting corridor alternatives better than method 3 (β_7). A reason for this might be that method 3 does not account for overlapping resistances, which in turn, results in less diversified corridor alternatives as the cost surface is flattened out. Moreover, the dendrograms illustrate a branching of MCDA method 3 between a relatively large height of 7 to 16. They also reveal that distinct clusters can be created when MCDA method 3 is applied on a continuous shape model. In contrast, the use of MCDA method 3 on a sharp-edged model branches the dendrogram at height 6, which does not necessarily affect separate clusters. Branching between MCDA methods 1 and 2 occurs at a very low height around 1 to 2 and is thus not relevant.

Third, the distinction between the different weighting models explains the model's principal components with a Pillai's trace between 22.0% and 98.5%. Certainly, the general distinction between the models (β_{13}) seems to be important as the corresponding Pillai's trace is very high. However, the variance among the weighting models is large, as β_{14} has a Pillai's trace of 49.4% to 82.5%, whereas β_{15} has 22.0% and β_{16} was insignificant. Generally, if MCDA methods 1 (eq. 9) or 2 (eq. 10) are used, the weighting model leads to a clear branching, although on a low height around 2. In contrast, the weighting model had no branching effect when it was applied on the *Maximum Value Method* (eq. 11), as it neglects the influence of overlapping factors.

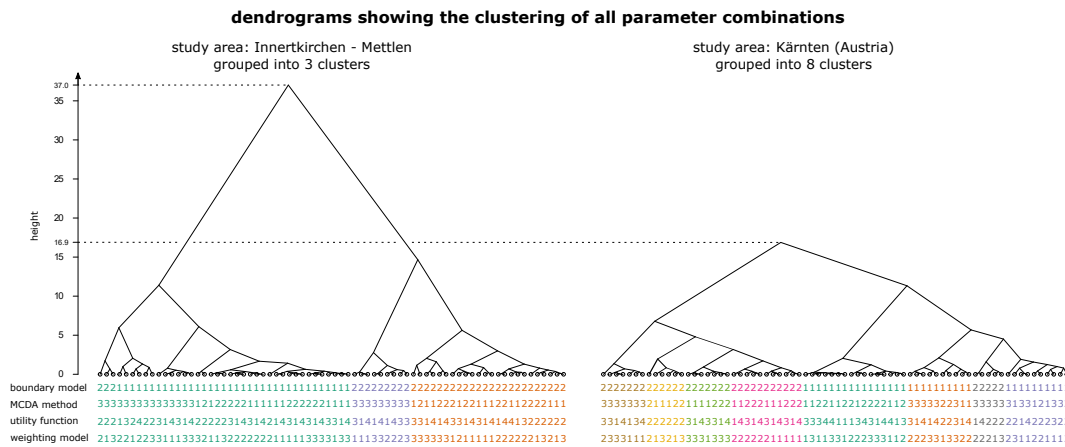


Figure 4 Clustering dendrogram of the study areas Innertkirchen – Mettlen (left) and Kärnten (right), normalized to the same absolute height. The numbers 1-4 represent the factor levels listed in tab. 2.

Fourth, the variation of the utility functions had the weakest effect with a Pillai’s trace between 18.2% and 90.4%. β_{10} and β_{12} modeled the principal components best with average Pillai’s traces of 81.0% and 79.5%. However, β_9 and β_{11} ranked lower and could explain the underlying principal components only to 49.7% and 28.2%. A distinct branching could only be determined for β_{10} , however, on a very low dendrogram height of approximately 1.

However, the corresponding regressor β_9 and even β_{10} were not determined to be significant by applying the MANOVA. In contrast, utility functions $u_{1,i,x}(r_{i,x})$ (β_8) and $u_{4,i,x}(r_{i,x})$ (β_{11}) were significant with a Pillai’s trace of 56.3% and 44.2% (both $p < .001$). The general result of distinguishing between the utility functions used, as shown by β_8 , had an effect on explaining the model by 22.3%. However, we could not determine any significant interaction between the boundary model, the MCDA method, the utility function, and the weighting model, as β_{17} to β_{22} were insignificant.

Another method to compare the goodness of the data model is to calculate to what extent the main objectives of the decision model have been violated. As shown in tab. 4, the primary objectives (Ω_1 , Ω_3 , and Ω_5) have been respected, which resulted in a low violation whereas areas corresponding to a secondary objective have been crossed more often.

4 Discussion

We set out to investigate the utility of a cluster analysis for improving a decision model. We therefore discuss in the following subsections, how our results are applicable in practice in order to simplify and improve a given decision model.

4.1 How the results help to simplify the decision model

Given that the considered principal components explain the variance of a defined model sufficiently, a MANOVA yields the strength of underlying factors that contribute to the explanation of the principal components. Thus, insignificant results indicate factors that can be excluded from the decision model. If the decision model aims at being universally applicable to different study areas, only factors significant across all study areas should be considered. In this study, only weighting model 1 (eq. 6, represented by β_{14}), could be used

■ **Table 3** Effect of all significant regressors used in the MANOVA, split by study area. The right panel lists the averaged Pillai’s traces and the according ranks of each regressor (see tab. 2).

Innertkirchen – Mettlen			Kärnten			Averaged Results		
Regressor	Pillai	Sig.	Regressor	Pillai	Sig.	Rank	Regressor	Pillai
β_2	.967	p <.001	β_2	.993	p <.001	1	β_2	.980
β_{13}	.925	p <.001	β_{13}	.985	p <.001	2	β_{13}	.955
β_3	.915	p <.001	β_3	.977	p <.001	3	β_3	.946
β_5	.825	p <.001	β_5	.962	p <.001	4	β_5	.894
β_{10}	.716	p <.001	β_{12}	.929	p <.001	5	β_6	.817
β_6	.712	p <.001	β_1	.924	p <.001	6	β_{10}	.810
β_1	.676	p <.001	β_6	.921	p <.001	7	β_1	.800
β_{12}	.662	p <.001	β_{10}	.904	p <.001	8	β_{12}	.795
β_{14}	.494	p <.001	β_{14}	.825	p <.001	9	β_{14}	.660
β_7	.484	p <.001	β_9	.746	p <.001	10	β_9	.497
β_8	.263	p <.001	β_7	.470	p <.001	11	β_7	.477
β_9	.247	p <.01	β_{11}	.425	p <.001	12	β_{11}	.282
β_{11}	.140	p <.05	β_{15}	.220	p <.01	13	β_8	.223
			β_8	.182	p <.05	14	β_{15}	.220

as β_{15} and β_{16} were insignificant across both study areas. It is further questionable whether factors with a small Pillai’s trace should be considered in the decision model. However, this would beg the question, from which value on a contribution should be specified to be sufficient. Thus, this question could be a line of interesting future research.

Although decision-makers might expect different outcomes based on every chosen parameterization, our results indicate that the solution space is limited. Even if solutions may differ slightly, it is still desirable for transmission line planners to obtain corridor alternatives that are clearly different from each other. For this, the applied procedure could help to determine the factors with the highest effect on the resulting corridor. The importance of these factors can be discussed within a group of decision-makers in order to improve the decision model based on a conjoint solution. Being able to explain which factors contribute most and adapting them in a participatory approach might lead to a fostering of transparency, which in turn will increase the acceptance of the model.

Especially when considering the MCDA methods used, the results concerning the weighting model would probably have been more distinct if we refused using MCDA model 3, as its results were categorized into a separate cluster. In addition, even though MCDA model 3 leads to more direct connections between start and end point, it intersects more protected areas when compared to the application of the remaining MCDA methods. As the branching between MCDA methods 1 and 2 occurs at a low height of around 1 to 2, we conclude that this distinction is not of high importance. Thus, *Simple Additive Weighting* as described in eq. 9 would be the easiest and most accessible solution to conduct an MCDA.

4.2 How the results help to improve the decision model

The statistical evaluation performed indicates that the factors contained by the decision model are structured hierarchically. Thus, factors contribute differently to the variability of the suitability maps. By knowing the Pillai’s trace, the decision model could be improved by multiplying each factor with a value that inverts its effect on explaining the model. In this

■ **Table 4** Percent of the average path length over which the according objective (Ω_i that correspond to tab. 1) does not comply with. The values were averaged across both study areas.

Parameter	Level	Ω_1	Ω_2	Ω_3	Ω_4	Ω_5	Ω_6	Ω_7	Ω_8	Ω_9	Ω_{10}
Boundary model	1	.00	.46	.11	.16	.08	.37	.01	.09	.50	.00
	2	.00	.37	.08	.13	.09	.42	.00	.17	.61	.00
MCDA method	1	.00	.41	.09	.14	.08	.40	.01	.13	.56	.00
	2	.00	.41	.09	.13	.08	.40	.01	.13	.56	.00
	3	.00	.43	.11	.18	.10	.41	.00	.15	.59	.00
Utility function	1	.00	.42	.10	.15	.09	.40	.01	.13	.56	.00
	2	.00	.40	.11	.14	.10	.39	.01	.15	.57	.00
	3	.00	.43	.09	.16	.08	.40	.01	.12	.56	.00
	4	.00	.42	.10	.15	.09	.40	.01	.13	.56	.00
Weighting model	1	.00	.41	.10	.15	.09	.40	.01	.13	.57	.00
	2	.00	.41	.10	.14	.09	.39	.01	.14	.56	.00
	3	.00	.43	.09	.15	.08	.40	.01	.12	.57	.00

way, the weight of factors with a low contribution could be increased and vice versa. If we took the only significant weighting model eq. 6 and aimed at standardizing the effect of all factors, the weighting model might be extended by the subsequent equation, where i is the total number of factors and p_i the factor’s Pillai’s trace, which is used as a *swing weight* [2]:

$$\forall x \geq 0 \rightarrow h_{1,i}(w_i, p_i) = \frac{\sqrt[3]{w_i}}{i \cdot p_i} \quad \text{and} \quad \forall x < 0 \rightarrow h_{1,i}(w_i, p_i) = \frac{1}{\sqrt{w_i} \cdot i \cdot p_i} \quad (12)$$

Furthermore, we could not detect any significant interactions between the factor levels used; neither by increasing the number of considered principal components to 8, as considered by using Kaiser’s criterion. Thus, we conclude that the factor levels used are independent, which emphasizes the unbiased nature of the decision model. In turn, this unbiased decision model may support decision-making, as decision-makers can independently choose a parametrization without accounting for the effect that a factor might have on another.

Another point that helps to improve the model can be deduced from the dendrograms. As large branching heights result in distinct clusters, the ideal choice of distinct factors might improve outcome variability. However, as the rules applied to generate the maps remained unchanged across both study areas, we assume that the underlying data model influences the amount of variability. Thus, decision-makers should pay attention when carefully deciding, which data model represents the reality best. The results listed in tab. 4 point in the same direction, as large and continuous areas were crossed more often than small and dispersed areas. We therefore propose that both the size and the spatial distribution of the underlying geodata should also be considered when defining the data model. A reflected setting of the data model might thus help to improve the quality of the subsequent analysis.

5 Conclusion

This study investigated to what extent a multi criteria decision model leads to biased results when determining the suitability for constructing new transmission lines at a specific place. We first defined a decision model consisting of 33 spatially explicit factors, each representing an area that emits a resistance against constructing a transmission line on it. Besides these factors, we modified a standard MCDA model by defining four modeling parameters that might alter the location and the course of the resulting transmission line corridor and

path. We then followed this by conducting a sensitivity analysis by computing all suitability maps resulting from combining all parameter levels with each other. Then, we averaged the resulting corridors by the 72 possible parameter settings. A cluster analysis was subsequently conducted to determine mutual corridor courses, thus, the decision model's bias. Finally, we applied a MANOVA to identify the parameters' influence for explaining the decision model based on its principal components.

Our results demonstrate that the decision, whether a sharp-edged or a continuous boundary model should be applied, is of highest importance, as the resulting corridors significantly differ from each other. Concerning the MCDA method chosen, *Simple Additive Weighting* and the *Maximum Value Model* led to the highest diversity, whereas the latter should be handled with caution, as the model considered the spatial structure of the given data worst. Our analysis further revealed that a logarithmic weighting model and a utility function enhancing the effects of low and high resistances led to more distinct corridor alternatives than using linear models. Moreover, our proposed procedure for enhancing the decision model led to similar results across both investigated study areas. Consequently, it also might be applicable to other study areas to simplify and to improve other MCDA models.

Contrary to prior work that commonly used AHP/ANP, MAUT/MAVT, or PROMETHEE for determining the factors' weights, we propose to adapt them based on the results obtained by statistically evaluating the results of a sensitivity analysis using the described analysis method. The proposed method aims at adjusting the subjectively assigned weight by including an additional swing weight for each factor. As the swing weights represent the statistically determined influence of the corresponding factors, the bias given by the data and decision model can be diminished, thus, enlarging the solution space for other corridor alternatives. We assume that acceptance can be increased by first knowing the DSS's behavior in generating alternative suitability maps and then improving it based on the results obtained by the proposed approach. Future work could, for example, explore whether the proposed weighting adaption effectively results in a higher diversity of generated alternatives, also by performing a sensitivity analysis with continuous, normally distributed weights around an expected value. Moreover, settings of resistances and weights pursuing the same objective could be combined to scenarios, which in turn could be integrated into an analysis approach to determine the combined effect of the geodata, the scenarios, and the MCDA parameters. It remains to be further investigated how planning experts assess the goodness, usability, and practicability of the proposed approach.

References

- 1 Antonella Battaglini, Nadejda Komendantova, Patricia Brtnik, and Anthony Patt. Perception of barriers for expansion of electricity grids in the European Union. *Energy Policy*, 47:254–259, 2012. doi:10.1016/j.enpol.2012.04.065.
- 2 Valerie Belton and Theodor Stewart. *Multiple Criteria Decision Analysis: An Integrated Approach*. Springer Science & Business Media, Dordrecht, Netherlands, 2002.
- 3 Kjetil Bevanger, Gundula Bartzke, Henrik Brøseth, Espen Lie Dahl, Jan Ove Gjershaug, Frank Hanssen, Karl-Otto Jacobsen, Oddmund Kleven, Pål Kvaløy, Roel May, Roger Meås, Torgeir Nygård, Steinar Refsnæs, Sigbjørn Stokke, and Jørn Thomassen. Optimal design and routing of power lines; ecological, technical and economic perspectives (OPTIPOL). Final Report 1012, Norwegian Institute for Nature Research, Trondheim, Norway, 2014.
- 4 Nicholas L. Cain and Hal T. Nelson. What drives opposition to high-voltage transmission lines? *Land Use Policy*, 33:204–213, 2013.

- 5 Shahar Chen, David Amid, Ofer M. Shir, Lior Limonad, David Boaz, Ateret Anaby-Tavor, and Tobias Schreck. Self-organizing maps for multi-objective pareto frontiers. In *2013 IEEE Pacific Visualization Symposium*, pages 153–160, 2013.
- 6 Charles W. Churchman, Russell L. Ackoff, and Nicolas M. Smith. An approximate measure of value. *Journal of the Operations Research Society of America*, 2(2):172–187, 1954.
- 7 Ana Roxana Ciupuliga and Eefje Cuppen. The role of dialogue in fostering acceptance of transmission lines: the case of a France–Spain interconnection project. *Energy Policy*, 60:224–233, sep 2013. doi:10.1016/j.enpol.2013.05.028.
- 8 Michael Clegg, Katherine Ellena, David Ennis, and Chad Vickery. *The Hierarchy of Laws: Understanding and Implementing the Legal Frameworks that Govern Election*. International Foundation for Electoral Systems, Arlington, USA, 2016.
- 9 DETEC. Sectoral Plan for Transmission Lines (SÜL), dec 2001.
- 10 David H. Douglas. Least-cost Path in GIS Using an Accumulated Cost Surface and Slope-lines. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 31(3):37–51, 1994. 00117. doi:10.3138/D327-0323-2JUT-016M.
- 11 Franz Eisenführ, Martin Weber, and Thomas Langer. *Rational Decision Making*. Springer, Berlin, 2010.
- 12 José Figueira, Salvatore Greco, and Matthias Ehrgott. *Multiple Criteria Decision Analysis: State of the Art Surveys*. Springer Science & Business Media, 2005.
- 13 Stefano Grassi, Roman Friedli, Michel Grangier, and Martin Raubal. A GIS-Based Process for Calculating Visibility Impact from Buildings During Transmission Line Routing. In Joaquín Huerta, Sven Schade, and Carlos Granell, editors, *Connecting a Digital Europe Through Location and Place*, Lecture Notes in Geoinformation and Cartography, pages 383–402. Springer International Publishing, jan 2014.
- 14 Gayle Houston and Christy Johnson. EPRI-GTC Overhead Electric Transmission Line Siting Methodology. Technical Report 1013080, Electric Power Research Institute and Georgia Transmission Corporation, Palo Alto (CA) and Tucker (GA), USA, 2006.
- 15 Piotr Jankowski, Natalia Andrienko, and Gennady Andrienko. Map-centred exploratory approach to multiple criteria spatial decision making. *International Journal of Geographical Information Science*, 15(2):101–127, 2001. doi:10.1080/13658810010005525.
- 16 Joshu Jullier. More acceptance for power lines in Switzerland: An evaluation of the acceptance increasing factors for transmission lines in Switzerland. Master’s thesis, ETH Zurich, Zurich, Switzerland, 2016. doi:10.3929/ethz-b-000240496.
- 17 Pascal Lienert, Bernadette Sütterlin, and Michael Siegrist. The influence of high-voltage power lines on the feelings evoked by different Swiss surroundings. *Energy Research & Social Science*, 23(Supplement C):46–59, jan 2017. doi:10.1016/j.erss.2016.11.010.
- 18 Arika Ligmann-Zielinska and Piotr Jankowski. Spatially-explicit integrated uncertainty and sensitivity analysis of criteria weights in multicriteria land suitability evaluation. *Environmental Modelling & Software*, 57:235–247, 2014.
- 19 Jacek Malczewski and Claus Rinner. *Multicriteria Decision Analysis in Geographic Information Science*. Advances in Geographic Information Science. Springer, Berlin, 2015.
- 20 Stephan Pajer, Marc Streit, Thomas Torsney-Weir, Florian Spechtenhauser, Torsten Möller, and Harald Piringer. Weightlifter: Visual weight space exploration for multi-criteria decision making. *IEEE transactions on visualization and computer graphics*, 23(1):611–620, 2017. doi:10.1109/TVCG.2016.2598589.
- 21 Arne Spieker. Stakeholder Dialogues and Virtual Reality for the German Energiewende. *Journal of Dispute Resolution*, 2018(1), jan 2018.
- 22 Waldo R. Tobler. A computer movie simulating urban growth in the Detroit region. *Economic geography*, pages 234–240, 1970.
- 23 Wade M. Vagias. Likert-type Scale Response Anchors, 2006.

15:16 Enhanced Multi Criteria Decision Analysis for Planning Power Transmission Lines

- 24 Geert Verbong and Frank Geels. The ongoing energy transition: Lessons from a socio-technical, multi-level analysis of the Dutch electricity system (1960–2004). *Energy Policy*, 35(2):1025–1037, feb 2007. doi:10.1016/j.enpol.2006.02.010.
- 25 Xun Zhao, Yanhong Wu, Weiwei Cui, Xinnan Du, Yuan Chen, Yong Wang, Dik Lun Lee, and Huamin Qu. SkyLens: Visual Analysis of Skyline on Multi-Dimensional Data. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):246–255, 2018.

FUTURES-AMR: Towards an Adaptive Mesh Refinement Framework for Geosimulations

Ashwin Shashidharan

Department of Computer Science, North Carolina State University, Raleigh, USA
ashdharan@ncsu.edu

Ranga Raju Vatsavai

Department of Computer Science, North Carolina State University, Raleigh, USA
rrvatsav@ncsu.edu

Derek B. Van Berkel

Center for Geospatial Analytics, North Carolina State University, Raleigh, USA
dbvanber@ncsu.edu

Ross K. Meentemeyer

Center for Geospatial Analytics, North Carolina State University, Raleigh, USA
rkmeente@ncsu.edu

Abstract

Adaptive Mesh Refinement (AMR) is a computational technique used to reduce the amount of computation and memory required in scientific simulations. Geosimulations are scientific simulations using geographic data, routinely used to predict outcomes of urbanization in urban studies. However, the lack of support for AMR techniques with geosimulations limits exploring prediction outcomes at multiple resolutions. In this paper, we propose an adaptive mesh refinement framework FUTURES-AMR, based on static user-defined policies to enable multi-resolution geosimulations. We develop a prototype for the cellular automaton based urban growth simulation FUTURES by exploiting static and dynamic mesh refinement techniques in conjunction with the Patch Growing Algorithm (PGA). While, the static refinement technique supports a statically defined fixed resolution mesh simulation at a location, the dynamic refinement technique supports dynamically refining the resolution based on simulation outcomes at runtime. Further, we develop two approaches - asynchronous AMR and synchronous AMR, suitable for parallel execution in a distributed computing environment with varying support for solution integration of the multi-resolution results. Finally, using the FUTURES-AMR framework with different policies in an urban study, we demonstrate reduced execution time, and low memory overhead for a multi-resolution simulation.

2012 ACM Subject Classification Computing methodologies → Distributed simulation, Computing methodologies → Multiscale systems, Applied computing → Environmental sciences

Keywords and phrases Adaptive mesh refinement, Geosimulation, Distributed system, Multi-resolution, Urban geography

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.16

1 Introduction

Over the past decade, advancements in remote sensing technologies and classification techniques have increased the availability of high-resolution datasets relevant to urban simulation. High resolution LiDAR derived DEMs, land cover classifications, and the increasing amount of vector-based spatial layers promise to deliver a better understanding of urbanization for forecasting urban development. However, in practice, computational constraints impact



© Ashwin Shashidharan, Ranga Raju Vatsavai, Derek B. Van Berkel, and Ross K. Meentemeyer; licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 16; pp. 16:1–16:15

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

the resolution of input data, or the extent of the study region used in an urban simulation. Particularly, memory and I/O constraints limit studies leveraging high-resolution data to small study extents, while study of large extents are often only possible with low-resolution data. Although an urban simulation may require high-resolution data only in a small region of the study (as shown in Fig. 1b), current urban simulation frameworks do not support selectively varying the resolution of a simulation for different regions (as shown in Fig. 1c and Fig. 1d) at runtime. Further, if new urbanization is highly likely only on a small portion of the study extent, modifying the urban growth simulation to use high-resolution data over the complete study extent is highly inefficient.

Adaptive mesh refinement (AMR) is a technique that can support multi-resolution simulations using high-resolution data in regions where it is necessary. For urban growth simulations in large study extents, adaptive mesh refinement at runtime would allow focusing computational resources for simulating emerging urban patterns in regions of interest (ROIs). An AMR approach using high-resolution data would account for more prominent local effects like topographic features and land cover classes to simulate accurate urbanization patterns. Additionally, using low-resolution data for simulation in regions of less importance would reduce memory overhead and enable faster simulation. In effect, such an approach would eliminate the computational overhead of a high-resolution simulation over the global extent of a study region, while generating fine spatial patterns where necessary.

Although an AMR approach promises significant computational savings, AMR techniques developed thus far only support refinement and coarsening criteria for solving partial differential equations (PDEs) in a scientific simulation. In particular, these are not applicable to geosimulations which use cellular automaton (CA) based models to generate urbanization outcomes. Thus, the first challenge is the development of new refinement and coarsening criteria to support AMR with geosimulations like urban growth. In particular, geosimulations require a *mesh placement strategy* that specifies the location, extent and spacing of a mesh (resolution), and a *mesh generation strategy* for use with different datatypes in the simulation. In turn, the choice of a mesh generation strategy impacts the integration strategy for synchronizing the results generated at different resolutions.

In this paper, we address this research gap and develop a distributed AMR framework, FUTURES-AMR that supports multi-resolution geosimulations. Specifically, the framework supports *refinement* and *coarsening* requests using multi-resolution data in regions of interest (ROIs) for two scenarios: (i) *static refinement* in ROIs specified by an end user; (ii) *dynamic refinement* based on a combination of static policies and the simulation outcomes at runtime. For both scenarios, we allow end users to specify static policies that define refinement and coarsening criteria for the AMR simulation. Finally, we develop two approaches - asynchronous AMR and synchronous AMR with different load balancing and solution integration strategies in a master-worker style distributed system architecture.

The rest of the paper is organized as follows: in Sect. 2, we summarize existing research for AMR simulation. In Sect. 3, we provide an overview of Adaptive Mesh Refinement as used in numerical analysis. In Sect. 4, we describe our distributed system architecture for AMR in the asynchronous and synchronous AMR approaches, and how we adapt the FUTURES geosimulation in our AMR framework. In Sect. 5, we describe our experimental setup and present results from executing FUTURES-AMR in two different geographic regions with user-defined policies. Finally, we conclude in Sect. 6, with future work in Sect. 7.

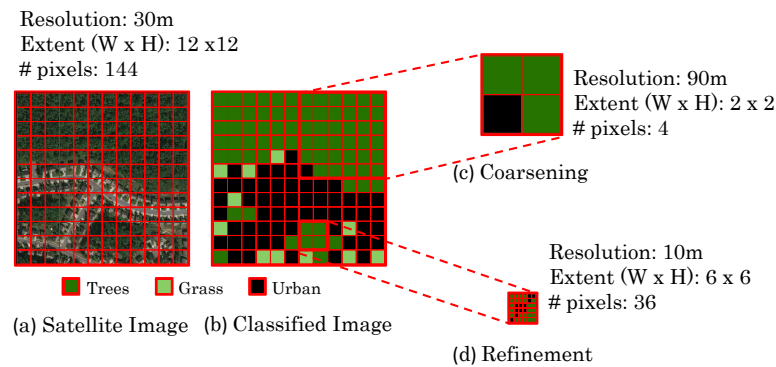


Figure 1 Illustration of the proposed FUTURES-AMR framework – The study extent shown in the classified image has 60% non-urban pixels. Refinement using 10m resolution data, and coarsening using 90m resolution data are requested on 5% and 40% of the total non-urban pixels in the study area, respectively. The default simulation at 30m resolution executes on the remaining 55% of non-urban pixels.

2 Related Work

Adaptive mesh refinement (AMR) is a technique that can be used with both structured and unstructured meshes. AMR techniques support dynamically adjusting the cell spacing on a mesh to achieve an accurate numerical solution. Structured adaptive mesh refinement (SAMR) was first proposed by Berger et al. [4] to solve partial differential equations (PDE) in shock hydrodynamics. This technique which relies on partitioning the problem space into different regions with varying spatial resolutions is achieved by imposing varying resolution grids in space. Further, each region is assumed to be rectangular in shape with a grid hierarchy to represent the relationships between different regions. As the solution progresses, nested grids or new grids are generated refining the problem in these regions. In case of time-dependent equations, these refinements can be applied to compute solutions at finer temporal resolutions as well.

Initially developed to solve simulations using hyperbolic conservation laws [4, 3], AMR approaches have since been extended to solve parabolic and elliptic equations. These numerical solvers find widespread use across various domains such as Computational Fluid Dynamics [4, 3], Astrophysics [9] and Climate Modeling [18]. General-purpose AMR frameworks have also been developed that support developing applications not specifically designed for a domain. BoxLib [2], Chombo [6] and SAMRAI [24] are examples of such frameworks with numerical solvers and APIs for developing codes for new applications. A comprehensive listing of the different frameworks for AMR refinement and their applications can be found in a survey by Dubey et al. [8].

AMR frameworks typically define a grid hierarchy management scheme to handle the coarse and fine regions. Block representation schemes have been devised which represent regions as grids using lower and upper coordinates of a bounding box [4, 2, 6, 24]. Similarly, tree representations exist that define coarse and fine regions in terms of parent-child relations and their splitting criteria [10, 20, 12]. These representations have implications on the number of cells to refine, the data distribution strategy for parallel computation, memory requirements and storage overhead. Exclusive computational geometry libraries also exist which support creation of structured and unstructured meshes for scientific simulations. CGAL [15], Silo [19], PARAMESH [12] are examples of libraries with geometry algorithms

and mesh generation and management routines for use in scientific applications. However, these libraries lack support for numerical solvers and AMR refinement operations. PETSc [1] and Hypr [14] are libraries with parallel numerical solvers. Even so, combining individual libraries to port existing serial code and develop a parallel AMR application requires parallel programming expertise and significant rework.

Owing to the numerical complexity of solving partial differential equations (PDEs), most AMR related research has focused on developing data structures and algorithms to support parallel and distributed computation [7]. These frameworks are designed with one of the two popular load balancing strategies for AMR: *patch-based* and *domain-based* (or *tree-based*). In the patch-based approach [4, 23], load balancing distributes regions for refinement over a set of processors using a binning, greedy or round-robin technique. Although, a patch-based approach offers a simple load balancing strategy to balance overall computational work at a processor, data movement for synchronizing results across refinement levels is unavoidable and could lead to significant communication overhead. On the other hand, domain-based approaches attempt to optimize communication overhead by assigning coarsening or refinement operations for a sub-region to a processor where its parent region resides [10, 20, 12]. However, domain-based approaches suffer from scalability issues at higher levels of refinement as dynamic reconfiguration of the workload necessitates data migration to maintain the load and avoid synchronization between nested levels at each processor. A comprehensive comparison of the parallelization techniques for dynamic load balancing can be found in a survey by Rantakokko et al. [21]. The results of the survey indicate that no single partitioning scheme performs best across all types of applications and systems. Finally, AMR frameworks typically also define techniques to integrate results at the boundary of coarse-fine interfaces. Refluxing and circulation integration techniques, which combine results from interpolation of low resolution data at coarser levels and aggregation of data at finer levels are used to update PDE solutions at the boundaries.

General-purpose parallel AMR frameworks attempt to reduce the programming effort to develop parallel structured AMR applications. While most of these frameworks are distributed memory implementations [9, 6, 24], AMRCLAW [5] is a shared memory implementation. Parallel AMR frameworks facilitate development of parallel AMR applications by handling data organization and distribution, load balancing and data communication as part of the framework [17]. Along with numerical solvers for PDEs, these frameworks abstract the implementation details such as the data type, parallel communication patterns and data placement strategies from the user. AMR frameworks [16, 11] also exist that compute solutions in irregularly shaped regions of the sub-domain without assuming a logically rectangular structure. However, similar to structured AMR frameworks these are only suitable for scientific applications using partial differential equations. Finally, we are not aware of AMR frameworks developed to support geosimulations.

3 Adaptive Mesh Refinement

To compute a numerical solution for PDEs, an adaptive mesh refinement technique starts by imposing a coarse grid (or mesh) over the complete problem domain. The grid defines the cell spacing, or resolution for computation in the domain. Imposing a finer grid in the domain introduces more grid points while, a coarse grid presents fewer points at which, solutions for the equation must be calculated. Thus, the computational complexity to solve a PDE depends on the grid spacing of the domain. An adaptive mesh refinement technique superimposes fine grids only in certain sub-domains (or regions) of the problem (also known

as regridding). These are identified by estimating the accuracy or error of the computed solution. Finer grids are recursively imposed in the region till the error or accuracy of the computed solution is acceptable (i.e., below or above a threshold), or a maximum level of refinement is reached. Specifying a maximum level of refinement avoids infinite recursion in the regridding step. Thus, in an AMR based solution, a coarse grid is applied on the complete problem domain, but recursively refined in regions till a suitably accurate solution is obtained.

In regions superimposed with finer grids, AMR uses interpolation to resolve the initial values at the fine grid points from the coarse grid points. Subsequently, the solutions of the equations at the finer grids points are computed, and results at the fine grid points are aggregated to update the solution at the coarse grid points. Along the fine-coarse grain region boundaries, the AMR integration approach uses a flux conservation or circular integral control technique to update values at the coarse grid points. Thus, a solution at the initial coarse resolution (or default resolution) for the complete domain is obtained using AMR.

4 FUTURES-AMR

In our framework, we modify the Berger-Oliger-Collela approach [4] to support adaptive mesh refinement for a geosimulation. We make two major modifications in the four step Berger-Oliger-Collela approach. Firstly, we substitute the problem of solving PDEs at different intervals in a domain with an urban growth simulation using a Patch Growing Algorithm (PGA) [13] in a geographic region. Secondly, we modify the error-based AMR refinement criteria for PDE solvers with AMR refinement criteria based on user-defined policies for a region. Thus, in our FUTURES-AMR framework, the FUTURES urban simulation executes the PGA at different resolutions based on refinement criteria expressed in user-defined policies.

We also make a few assumptions about supported geosimulations in this framework. First, a geosimulation executing in this framework is assumed to be a cellular automaton consisting of a grid of cells with transition rules such as defined by a PGA. Second, each cell has a fixed spatial resolution representing a fixed area on the landscape. A geosimulation begins at this fixed resolution over the complete landscape. Third, the transition rules of the CA-based geosimulation for patch growth must be specified, or generalizable for use at different resolutions. Based on these assumptions and modifications, we define the FUTURES-AMR algorithm as follows:

- ▶ **Step 1.** Start a geosimulation with a coarse default resolution over the complete study extent.
- ▶ **Step 2.** Evaluate static policies as part of PGA to identify regions that need higher/lower resolution data.
- ▶ **Step 3.** Superimpose finer grids for refinement or coarser grids for coarsening in these regions. Subsequently, execute PGA till either the PGA halting criteria is met or higher resolution data is unavailable.
- ▶ **Step 4.** Integrate multi-resolution simulation outcomes from refinement and coarsening in different regions with the default resolution result in the global extent.

We design two simulation approaches namely, asynchronous AMR and synchronous AMR that vary in their implementation of Step 4. We describe these approaches and their varying support for policies in Section 4.3 and 4.4. We begin with a brief description of the PGA for the proposed framework.

4.1 Patch Growing Algorithm (PGA)

In the simulation of an urban landscape, new urban patches are developed by executing a Patch Growing Algorithm at suitable development sites in the landscape. One standard method [13] is to determine a suitable seed and execute a neighbor discovery process to determine new cells for urban patch growth. The PGA generates new patches that characterize the spatial changes due to urbanization in terms of *patch shape* and *patch size* starting at the seed location. However, the algorithm depends on a fine grid to capture these patterns at a fine granularity. In general, wider spacing of grid points results in lower data resolution representing the landscape and hence, higher inaccuracy in patterns of the generated patches. These solutions may be acceptable in certain regions of a landscape, e.g. in a sparsely populated remote rural region, but not in a dense urban region like a central business district (CBD). Thus, to support varying mesh spacing depending on the requirement in a region, we modify the PGA to generate refinement and coarsening requests at runtime.

4.1.1 Refinement/Coarsening

In FUTURES-AMR, a refinement or coarsening request is generated in response to user-defined policies in a region. These policies (see Section 4.2) define a refinement or coarsening criteria in a region for use during the simulation. A refinement criterion imposes a finer grid in a buffer region surrounding the seed site. In turn, the simulation executes the PGA using high-resolution data (resolution higher than the default resolution) in this region. Besides fine grids, coarse grids may also be specified for patch growth using PGA. In case of coarse grids, the simulation uses low-resolution data (resolution lower than the default resolution) in this region for the PGA. In case of both, fine and coarse grids, further refinement may be triggered to meet the PGA halting criteria until a higher resolution of data is unavailable. Thus, the simulation proceeds in discrete time-steps executing the PGA at default resolution, or by refining, or coarsening select regions in the geographic extent. The simulation result at the end of each time-step is a collection of coarsening results, refinement results and the simulation result at the default resolution. We formally define a refinement and coarsening request as follows:

$$X(L, E, r) \leftarrow P_1 \wedge P_2 \wedge \dots \wedge P_n \quad (1)$$

where X is either a refinement or coarsening request, L is the geolocation of the request, E is the extent to refine or coarsen from L , r is the resolution to use with the request, and $P_1 \dots P_n$ are user-defined policies in the extent E .

4.2 Policy Specification

In our AMR framework, we support user-defined static policies specified as input to the simulation. These policies serve as refinement and coarsening criteria for a simulation to perform *static* or *dynamic* refinement. If a geosimulation is unable to satisfy urbanization conditions using low-resolution data, refinement is triggered. Similarly, satisfying development conditions by coarsening with low-resolution is also supported. We formally define a policy as follows:

$$P \leftarrow A_1 \wedge A_2 \wedge \dots \wedge A_n \quad (2)$$

where, A_i is a spatial or non-spatial attribute, and P is a user-defined policy expressed as a conjunction of such attributes.

4.2.1 Static Refinement

Static refinement is a technique used to a priori superimpose meshes in regions of interest. In these regions, the mesh resolution is adjusted once, which is then maintained throughout the simulation. In case of geosimulations, as described in Section 4.1, coarsening may also be acceptable in certain regions of the landscape. Thus, in the FUTURES-AMR framework static policies specified a priori support both, refinement and coarsening criteria in a region. Static policies specify such regions where simulations with a different resolution must be carried out. For example:

- P1:** A *spatial refinement policy* specifies a polygon feature and resolution of data (1m/10m) to simulate patterns of urban development.
- P2:** A *spatial coarsening policy* specifies a polygon feature and resolution of data (90m/270m) to simulate patterns of urban development.

4.2.2 Dynamic Refinement

Dynamic refinement is carried out in response to conditions arising during a simulation. In case of dynamic refinement, fine meshes are superimposed in regions based on a combination of simulation outcomes and a refinement criteria satisfied at runtime. The same is applicable to coarsening as well. In the FUTURES-AMR framework, refinement criteria for patch growth is defined using static policies. For example:

- P1:** A *patch growth refinement policy* specifies high-resolution data (1m/10m) to develop urban patches smaller than a given size within a distance from a central business district.
- P2:** A *patch growth coarsening policy* specifies low-resolution data (90m/270m) to develop urban patches greater than a given size beyond a distance from a central business district.
- P3:** A *data-driven policy* specifies the resolution of data (1m/10m/90m/270m) to develop urban patches based on site development potential determined at runtime.

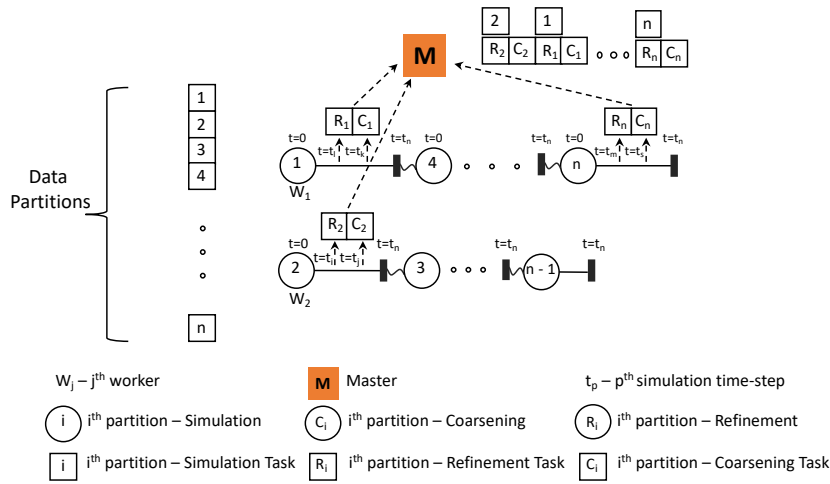
In the simulation, during PGA execution, these policies are evaluated at runtime to determine if dynamic refinement is necessary. A data-driven policy (e.g., P3) serves to resolve potential conflicts in case of multiple user-defined policies. If dynamic refinement is triggered, PGA iteratively refines the mesh to simulate urbanization till the refinement criteria is met. Thus, in dynamic refinement, the PGA adaptively adheres to the structure of the patch being developed at higher resolutions.

4.3 Asynchronous AMR

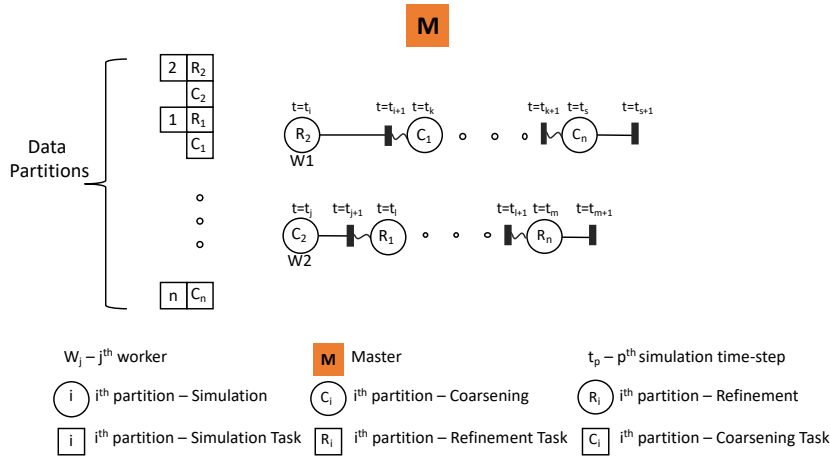
The asynchronous approach in our AMR framework is designed to support experimentation of policies at different resolutions in a study area. To be able to compare outcomes due to a user-defined policy, the approach executes the simulation, both in the presence and absence of policies, and generates results at different resolutions. In particular, the approach executes the PGA at multiple resolutions only in regions with user-defined policies, avoiding the execution overhead of a multi-resolution simulation over the complete study extent. Further, the emerging spatial structures in a time-step at different resolutions are retained as-is, eliminating additional I/O required to aggregate the results generated at different resolutions.

4.3.1 Solution Integration

In the asynchronous AMR approach, the results from adaptive mesh refinement are not integrated with the solution computed at the default resolution. Such an approach preserves



■ **Figure 2** Asynchronous AMR (Phase 1) – Each worker executes a simulation generating coarsening and refinement requests at each time-step of the simulation. The master receives and aggregates these requests from the workers at every time-step for processing in Phase 2.



■ **Figure 3** Asynchronous AMR (Phase 2) – The workers execute the refinement and coarsening requests from Phase 1 as assigned by the master.

the patch specificity obtained from multi-resolution simulations of different regions. GIS overlay techniques can be used to visualize the result layers from refinement and coarsening in different regions along with the global output raster.

4.3.2 Load Balancing

In the asynchronous approach, refinement and coarsening requests are processed independent of the simulation at the default resolution. Refinement and coarsening requests triggered by the simulation execute asynchronously without blocking the simulation. The approach ensures maximum resource utilization throughout the simulation.

We implement a master-worker approach for distributed asynchronous AMR simulations. In Phase 1 of this approach (Fig. 2), we begin by assigning each worker a geographic partition for simulation. Each worker executes a simulation on its partition generating new refinement

and coarsening requests. At the end of every time-step, the worker relays these requests to the master. Finally, once a worker completes all time-steps of the simulation on its assigned partition, the master schedules a new partition at the worker, if any.

Phase 2 (Fig. 3) begins when all partitions in the study have been processed. In Phase 2, the master schedules the refinement and coarsening requests received from the workers during Phase 1. Similar to Phase 1, each worker receives a refinement or coarsening request till all requests at the master have been processed. Finally, if further refinement becomes necessary while processing a request at a worker, it is executed at the same worker.

4.4 Synchronous AMR

The synchronous AMR approach propagates the effects of static policies in each time-step of the geosimulation to the subsequent time-step of the simulation. In this approach, spatial structures that emerge due to a policy at a particular time-step are input to the next time-step, i.e., the spatial effects of policies are temporally preserved as well. Specifically, the simulation outcomes from refinement and coarsening requests at different resolutions are integrated with the global solution for the region at every time-step. Thus, using the synchronous AMR approach, a user can explore long-term effects of static policies in a region.

4.4.1 Solution Integration

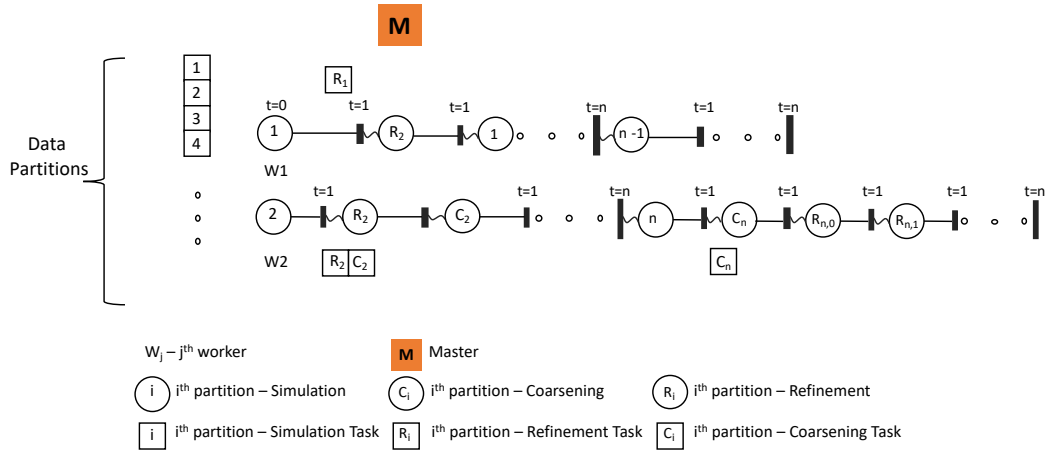
We devise a simple integration approach to merge solutions at the default resolution of the simulation for the global extent. In regions where coarsening occurs, we interpolate the low-resolution simulation result to the default resolution, and perform map algebra addition to combine it with the global output raster. Similarly, for refinement, we first aggregate the simulation result and perform map algebra addition on the global output raster. Thus, the refinement and coarsening results at different resolutions, in different regions, are integrated in every time-step at the default resolution of the global solution.

Effect of datatype on integration: In case of urbanization outcomes represented by a boolean datatype, we use average, mode or near resampling techniques to merge multi-resolution results at the default resolution (Fig. 9). In case of development pressure represented by a real datatype (e.g., in FUTURES [13]) we adopt one of the two approaches:

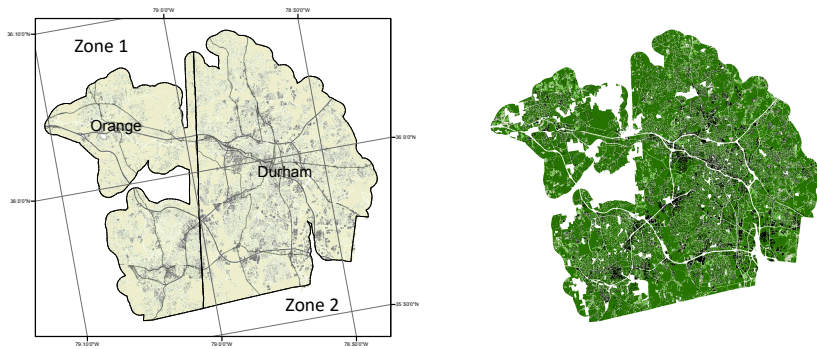
- (i) recalculate the development pressure over the complete study area after integrating the simulated urbanization results over the global extent or,
- (ii) use the result from the highest data resolution simulation in regions with multiple solutions.

4.4.2 Load Balancing

Once again, in the synchronous approach, the master begins by assigning different partitions for simulation at the workers. In every time-step, the workers build and maintain a list of coarsening and refinement requests. Subsequently, these requests are processed at the worker, i.e., a refinement or coarsening request is scheduled for execution at the same worker after the completion of a time-step. Any further refinement required is also carried out at the same worker. Additionally, as part of synchronization, integration of results (see Section 4.4.1) is carried out before the next time-step. Once the solution integration is complete, the worker resumes the simulation on its assigned partition at default resolution for the next time-step. This process is repeated in every time-step for all partitions the study area. Thus, in the synchronous AMR approach, all spatial and temporal interactions are preserved.



■ **Figure 4** Synchronous AMR – In each time-step, a worker aggregates the refinement and coarsening requests generated during the simulation on its partition. At the end of the time-step, the worker processes these requests and merges their solutions with the global output of the partition at the default resolution.



■ **Figure 5** The two zones used in the experiment (left) and the urbanization scene in 2010 (right).

5 Experimental Evaluation

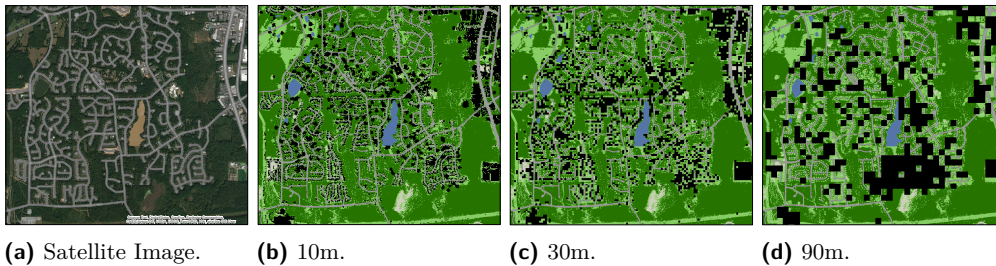
In this section, we describe the experimental setup of our proposed AMR framework. Figure 5 shows the two sub-county zones in the Raleigh-Durham (RDU) region used in our experiments. We carry out our experiments on a system with a hardware spec of 2.5 GHz Intel Core i7 processor and 16 GB memory, and software support for GDAL 2.0 and OpenMPI 1.10. Further, we setup our experiments to use three cores for MPI execution.

Experiment 1: Simulation overhead at different resolutions

In our first experiment, we measure the memory requirement and execution time for a simulation using a fixed input resolution. We setup the study area shown in Fig. 5 to execute 20 time-steps of the simulation in our experiment. We perform three simulation runs, varying the input resolution in each run to use 10m, 30m and 90m input resolution, respectively. Table 1 presents the simulation overhead and Figure 6 illustrates the output maps generated using different input resolution.

■ **Table 1** Simulation Execution Time and Memory Requirement at different input resolutions.

Execution Time (in seconds)			Memory Requirement (MB)		
10m	30m	90m	10m	30m	90m
91.7944	9.2878	1.1584	997	118	12



■ **Figure 6** Durham subdivision - Ridges of Parkwood. Fig. 6a is a satellite image from 2017. Fig. 6b, 6c, 6d illustrate urbanization in the year 2030 at 10m, 30m, 90m resolution data, respectively.

■ **Table 2** Simulation Execution Time and Memory Requirement with varying ROI extents.

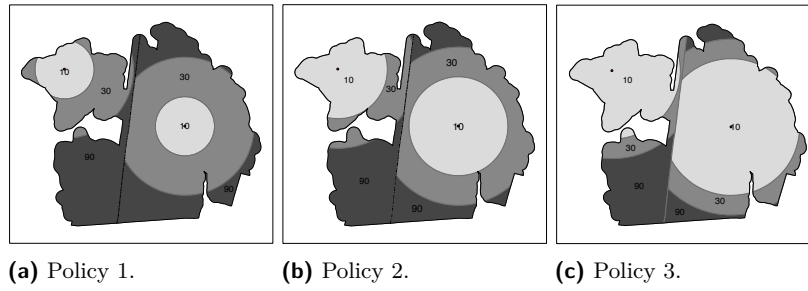
Extent of ROI (in 30m pixels)	Execution Time (in seconds)		Memory Requirement (MB)	
	10m	90m	10m	90m
30 x 30	0.12	0.12	16	15
60 x 60	0.17	0.16	25	21
200 x 200	0.53	0.41	63	39
300 x 300	0.84	0.63	96	48
400 x 400	1.78	1.14	128	55
500 x 500	3.13	1.67	193	61
600 x 600	4.3	2.42	248	64

We observe that both, the execution time and memory requirement increase with use of high-resolution data. Specifically, there is a 9-10x increase in both, the memory requirement and execution time, when the spatial resolution of the simulation is increased by a factor of 3. We use this as a baseline for comparison of the computational improvements in the synchronous and asynchronous approaches in our FUTURES-AMR framework.

Experiment 2: Static Refinement using static policies

In the FUTURES-AMR framework, static refinement supports superimposing finer or coarser meshes in particular regions of interest (ROIs). A static policy for refinement or coarsening defines the exact location and extent of these ROIs for high-resolution or low-resolution simulation. In our second experiment, we measure the overhead to execute refinement (10m) and coarsening (90m) requests with varying ROI extents to test how varying policies would impact computational efficiencies.

We observe that as the size of the ROI increases, execution time and memory requirement for executing a refinement and coarsening request increases. However, the refinement overhead is significantly lesser when compared to using high-resolution 10m data for the simulation over the complete study extent (shown in Table 1). Specifically, in the worst case, a refinement request by a default 30m resolution simulation, increases the the total execution time by 4.3 seconds and peak memory requirement of the simulation by 248MB. Thus, by using the FUTURES-AMR framework for processing refinement and coarsening requests, we incur significantly low computational costs for a multi-resolution simulation.



■ **Figure 7** The figure illustrates three policies with varying buffer zones based on two central business districts (Hillsborough in Zone 1 and Durham in Zone 2). In the inner zone, urban development using PGA triggers refinement requests (10m resolution) for patches with patchSize > 15 (patchSize is the total number of 30m pixels to simulate in an urban patch). In the outer zone, urban development using PGA triggers coarsening requests (90m resolution) for patches with patchSize > 30. In the middle zone, urban development using PGA always uses 30m resolution data.

■ **Table 3** Asynchronous AMR - Simulation Execution Time and Number of Requests.

Policy	Resolution		Number of Requests				Time (in s)
	10m	30m	Zone 1		Zone 2		
			Refinement	Coarsening	Refinement	Coarsening	
d2city	< 150	> 350	83	50	6	31	53.57
patchSize	> 15	> 30					
d2city	< 250	> 400	549	13	26	24	133.68
patchSize	> 15	> 30					
d2city	< 350	> 450	683	0	60	16	153.39
patchSize	> 15	> 30					

Experiment 3: Dynamic Refinement using static policies

In our third experiment, we use static policies as illustrated in Figure 7 for dynamic refinement. We run three experiments, where each experiment uses a different policy to simulate urban growth. We begin the simulation using coarse 30m resolution data, switching to high or low-resolution data for patch growth as determined by policy evaluation at runtime. The policies in our experiment specify two attributes for variable resolution simulation:

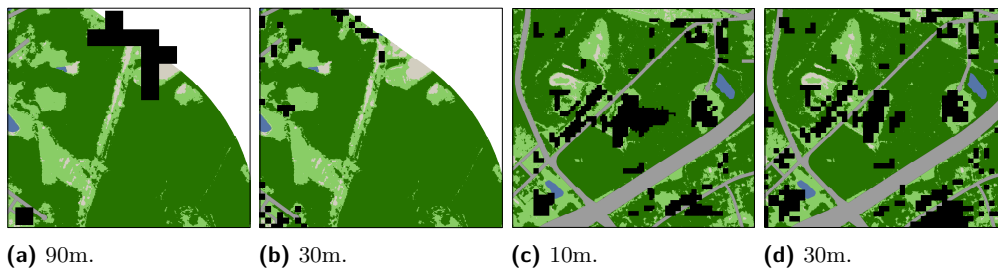
- (i) distance of the patch from a central business district (d2city);
- (ii) the size of the patch (patchSize).

The attributes are used to define threshold values for coarsening and refinement criteria. In dynamic refinement, the parameter values generated during the simulation are compared against these threshold values to trigger coarsening or refinement. Further, unlike static refinement, additional refinement is triggered if PGA halting criteria is not met. Table 3 and 4 present the measured execution times in the asynchronous and synchronous AMR approaches in our framework with the three policies. Both, d2city and patchSize in Tables 3 and 4 are expressed in terms of number of 30m pixels.

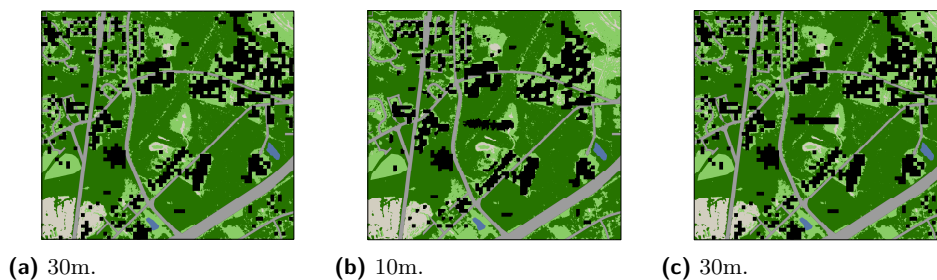
The results indicate that the execution time varies based on the number of requests, which are different between the approaches. Moreover, the execution time for processing different policies vary based on the number of requests. In particular, we observe that total execution time increases with increasing number of requests. Thus, user-defined policies must be carefully selected to limit the adverse impact on the total execution time. Nevertheless,

■ **Table 4** Synchronous AMR - Simulation Execution Time and Number of Requests.

Policy	Resolution		Number of Requests				Time (in s)
	10m	30m	Zone 1		Zone 2		
			Refinement	Coarsening	Refinement	Coarsening	
d2city	< 150	> 350	84	51	7	35	85.37
patchSize	> 15	> 30					
d2city	< 250	> 400	639	13	33	26	297.53
patchSize	> 15	> 30					
d2city	< 350	> 450	784	0	63	20	347.35
patchSize	> 15	> 30					



■ **Figure 8** Asynchronous AMR - Fig. 8a illustrates an output map of a coarsening request at 90m resolution. Fig. 8b illustrates the output map for the region in Fig. 8a at the default 30m resolution. Fig. 8c illustrates an output map of a refinement request at 10m resolution. Fig. 8d illustrates the output map for the region in Fig. 8c at the default 30m resolution.



■ **Figure 9** Synchronous AMR - Fig. 9a illustrates an output map at the default 30m resolution. Fig. 9b illustrates the output map for the region in Fig. 9a for a refinement request at a 10m resolution. Fig. 9c illustrates the composite output map generated by the simulation at the default 30m resolution by combining Fig. 9b and Fig. 9a in the synchronous approach.

the FUTURES-AMR multi-resolution framework demonstrates memory scalability, incurring a maximum additional memory overhead of 248MB as seen in Experiment 2. We also observe that total execution time in the synchronous AMR approach is higher than the asynchronous AMR approach. This increase in execution time is a result of the solution integration approach in the synchronous mode, where results from the multi-resolution simulations at different locations are merged into the final output raster of the study in every time-step. As the asynchronous AMR approach does not merge output results, it performs faster. Finally, in Fig. 8 and Fig. 9, using a few select regions from our study area, we illustrate the effects of user-defined policies on the simulation results generated in the two approaches.

6 Conclusion

FUTURES-AMR has been developed as a computing framework to support multi-resolution geosimulations for use in urban planning and development. In this paper, we described a generic framework for executing a distributed multi-resolution geosimulation and demonstrated its use with the FUTURES geosimulation. We developed static refinement and dynamic refinement techniques with support for expert defined and data-driven policies, along with two new approaches - synchronous and asynchronous AMR for distributed execution of a geosimulation. The results from evaluating the impact of three different user-defined policies on the quality and computational requirements demonstrate the framework's ability to execute a multi-resolution geosimulation with minimal execution time and memory overhead. Thus, in conclusion, the FUTURES-AMR framework, with its support for selective refinement in ROIs is suitable for urban studies using high-resolution data in large study extents.

7 Future Work

Urban development policies are designed in response to urbanization outcomes witnessed in previous years. They have a definitive timeframe associated with them, and often, success or failure of a policy leads to new or modified policies. However, currently, the FUTURES-AMR framework only supports static policies specified a priori. To support dynamic policies in different regions over time, our AMR framework can be integrated with computational steering features that support modification of simulation input at runtime. In future work, we propose to modify our computational steering framework, tFUTURES [22] to allow users to provide dynamic policies as steering input to the simulation.

References

- 1 Satish Balay, Kris Buschelman, William D Gropp, Dinesh Kaushik, Matt Knepley, L Curfman McInnes, Barry F Smith, and Hong Zhang. PETSc, the portable, extensible toolkit for scientific computation, 1998.
- 2 J Bell, A Almgren, V Beckner, M Day, M Lijewski, A Nonaka, and W Zhang. BoxLib user's guide. *github.com/BoxLib-Codes/BoxLib*, 2012.
- 3 Marsha J Berger and Phillip Colella. Local adaptive mesh refinement for shock hydrodynamics. *Journal of computational Physics*, 82(1):64–84, 1989.
- 4 Marsha J Berger and Joseph Oliger. Adaptive mesh refinement for hyperbolic partial differential equations. *Journal of computational Physics*, 53(3):484–512, 1984.
- 5 MJ Berger and R LeVeque. Adaptive mesh refinement for two-dimensional hyperbolic systems and the AMRCLAW software. *SIAM J. Numer. Anal.*, 35:2298–2316, 1998.
- 6 P Colella, DT Graves, TJ Ligoeki, DF Martin, D Modiano, DB Serafini, and B Van Straalen. Chombo software package for AMR applications-design document, 2000. URL: <http://seesar.lbl.gov/anag/chombo/ChomboDesign-3.1.pdf>.
- 7 Lori Freitag Diachin, Richard Hornung, Paul Plassmann, and Andy Wissink. Parallel adaptive mesh refinement. In *Parallel processing for scientific computing*, pages 143–162. SIAM, 2006.
- 8 Anshu Dubey et al. A survey of high level frameworks in block-structured adaptive mesh refinement packages. *Journal of Parallel and Distributed Computing*, 74(12):3217–3227, 2014. doi:10.1016/j.jpdc.2014.07.001.
- 9 Greg L Bryan et al. Enzo: An adaptive mesh refinement code for astrophysics. *The Astrophysical Journal Supplement Series*, 211(2):19, 2014.


- 10 Joseph E. Flaherty et al. Adaptive local refinement with octree load balancing for the parallel solution of three-dimensional conservation laws. *Journal of Parallel and Distributed Computing*, 47(2):139–152, 1997.
- 11 Orion S Lawlor et al. ParFUM: a parallel framework for unstructured meshes for scalable dynamic physics applications. *Engineering with Computers*, 22(3-4):215–235, 2006.
- 12 Peter MacNeice et al. PARAMESH: A parallel adaptive mesh refinement community toolkit. *Computer Physics Communications*, 126(3):330–354, 2000. doi:10.1016/S0010-4655(99)00501-9.
- 13 Ross K. Meenteemeyer et al. FUTURES: Multilevel Simulations of Emerging Urban-Rural Landscape Structure Using a Stochastic Patch-Growing Algorithm. *Annals of the Association of American Geographers*, 103(4):785–807, 2013.
- 14 Robert D Falgout and Ulrike Meier Yang. hypre: A library of high performance preconditioners. In *International Conference on Computational Science*, pages 632–641. Springer, 2002.
- 15 Efi Fogel and Monique Teillaud. The computational geometry algorithms library CGAL. *ACM Communications in Computer Algebra*, 47(3/4):85–87, 2014.
- 16 Daniel A Ibanez, E Seegyong Seol, Cameron W Smith, and Mark S Shephard. PUMI: Parallel unstructured mesh infrastructure. *ACM Transactions on Mathematical Software (TOMS)*, 42(3):17, 2016.
- 17 Scott R. Kohn and Scott B. Baden. Parallel software abstractions for structured adaptive mesh methods. *Journal of Parallel and Distributed Computing*, 61(6):713–736, 2001. doi:10.1006/jpdc.2001.1700.
- 18 John G Michalakes. RSL: A parallel runtime system library for regional atmospheric models with nesting. *IMA Volumes in Mathematics and Its Applications*, 117:59–74, 2000.
- 19 M Miller. Silo – a mesh and field I/O library and scientific database, 2018. URL: <https://wci.llnl.gov/simulation/computer-codes/silo>.
- 20 Manish Parashar and James C Browne. On partitioning dynamic adaptive grid hierarchies. In *System Sciences, 1996., Proceedings of the Twenty-Ninth Hawaii International Conference on.,* volume 1, pages 604–613. IEEE, 1996.
- 21 Jarmo Rantakokko and Michael Thuné. Parallel structured adaptive mesh refinement. *Parallel computing*, pages 147–173, 2009.
- 22 Ashwin Shashidharan, Ranga Raju Vatsavai, Abhinav Ashish, and Ross K. Meenteemeyer. tFUTURES: Computational steering for geosimulations. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL'17*, pages 27:1–27:10, New York, NY, USA, 2017. ACM. doi:10.1145/3139958.3140049.
- 23 John A Trangenstein. Adaptive mesh refinement for wave propagation in nonlinear solids. *SIAM Journal on Scientific Computing*, 16(4):819–839, 1995.
- 24 Andrew M. Wissink, Richard D. Hornung, Scott R. Kohn, Steve S. Smith, and Noah Elliott. Large scale parallel structured AMR calculations using the SAMRAI framework. In *Proceedings of the 2001 ACM/IEEE Conference on Supercomputing, SC '01*, pages 6–6, New York, NY, USA, 2001. ACM. doi:10.1145/582034.582040.

xNet+SC: Classifying Places Based on Images by Incorporating Spatial Contexts

Bo Yan

STKO Lab, University of California, Santa Barbara, USA

boyan@geog.ucsb.edu

 <https://orcid.org/0000-0002-4248-7203>

Krzysztof Janowicz

STKO Lab, University of California, Santa Barbara, USA

jano@geog.ucsb.edu

Gengchen Mai

STKO Lab, University of California, Santa Barbara, USA

gengchen@geog.ucsb.edu

Rui Zhu

STKO Lab, University of California, Santa Barbara, USA

ruizhu@geog.ucsb.edu

Abstract

With recent advancements in deep convolutional neural networks, researchers in geographic information science gained access to powerful models to address challenging problems such as extracting objects from satellite imagery. However, as the underlying techniques are essentially borrowed from other research fields, e.g., computer vision or machine translation, they are often not spatially explicit. In this paper, we demonstrate how utilizing the rich information embedded in spatial contexts (SC) can substantially improve the classification of place types from images of their facades and interiors. By experimenting with different types of spatial contexts, namely spatial relatedness, spatial co-location, and spatial sequence pattern, we improve the accuracy of state-of-the-art models such as ResNet – which are known to outperform humans on the ImageNet dataset – by over 40%. Our study raises awareness for leveraging spatial contexts and domain knowledge in general in advancing deep learning models, thereby also demonstrating that theory-driven and data-driven approaches are mutually beneficial.

2012 ACM Subject Classification Computing methodologies → Computer vision tasks, Computing methodologies → Neural networks, Theory of computation → Bayesian analysis

Keywords and phrases Spatial context, Image classification, Place types, Convolutional neural network, Recurrent neural network

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.17

1 Introduction

Recent advancements in computer vision models and algorithms have quickly permeated many research domains including GIScience. In remote sensing, computer vision methods facilitate researchers to utilize satellite images to detect geographic features and classify land use [5, 26]. In urban planning, researchers collect Google Street View images and apply computer vision algorithms to study urban change [22]. In cartography, pixel-wise segmentation has been adopted to extract lane boundary from satellite imagery [32] and deep convolutional neural network (CNN) has been utilized to recognize multi-digit house numbers from Google Street View images [10]. These recent breakthroughs in computer



© Bo Yan, Krzysztof Janowicz, Gengchen Mai, and Rui Zhu;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 17; pp. 17:1–17:15

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

vision are achieved, in equal parts, due to advances in deep neural networks as well as the ever-increasing availability of extensive training datasets. For example, the classification error in the latest image classification challenge using the ImageNet dataset is down to about 0.023.¹

However, such impressive results do not imply that these models have reached a level in which no further improvement is necessary or meaningful. On the contrary, such deep learning models which primarily depend on visual signals are susceptible to error. In fact, studies have shown that deep (convolutional) neural networks suffer from a lack of robustness to adversarial examples and a tendency towards biases [25]. Researchers have discovered that, by incorporating adversarial perturbations of inputs that are indistinguishable by humans, the most advanced deep learning models which have achieved high accuracy on test sets can be easily fooled [6, 11, 28]. In addition, deep learning models are also vulnerable to biased patterns learned from the available data and these biases usually resemble many unpleasant human behaviors in our society. For instance, modern neural information processing systems such as neural network language models and deep convolutional neural networks have been criticized for amplifying racial and gender biases [3, 4, 25, 33]. Such biases, which can be attributed to a discrepancy between the distribution of prototypical examples and the distribution of more complex real world systems [16], have already caused some public debates. To give a provocative example, almost three years after users revealed that Google erroneously labeled photos of black people as “gorillas”, no robust solutions have been established besides simply removing such labels for now.²

The above-mentioned drawbacks are being addressed by improvements to the available training data as well as the used methods [23, 3]. In our work, we follow this line of thought to help improve image classification. In our case, these images depict the facades or interiors of different types of places, such as restaurants, hotels, and libraries. Classifying images by place types is a hard problem in that more often than not the training image data is inadequate to provide a full visual representation of different place types. Solely relying on visual signals, as most deep convolutional neural networks do, falls short in modeling the feature space as a result. To give an intuitive example, facades of restaurants may vary substantially based on the type of restaurant, the target customers, and the surrounding. Their facade may be partially occluded by trees or cars, may be photographed from different angles and at different times of the day, and the image may contain parts of other buildings. Put differently, the principle of spatial heterogeneity implies that there is considerable variation between places of the same type.

To address this problem and improve classification accuracy, we propose to go beyond visual stimuli by incorporating spatial contextual information to help offset the visual representational inadequacy. Although data availability is less of an issue nowadays, the biased pattern in the data poses a real challenge, especially as models such as deep convolutional neural networks take a very long time to train. Instead of fine-tuning the parameters (weights) by collecting and labeling more unbiased data, which are very resource-consuming, we take advantage of external information, namely spatial context. There are many different ways one can model such context; in this work, we focus on the types of nearby places. We explore and compare the value of three different kinds of spatial context, namely spatial relatedness, spatial co-location, and spatial sequence pattern.

We combine these context models with state-of-the-art deep convolutional neural network models using search re-ranking algorithms and Bayesian methods. The result shows that,

¹ <http://image-net.org/challenges/LSVRC/2017/results#loc>

² <https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/>

by considering more complex spatial contexts, we can improve the classification accuracy for different place types. In fact, our results demonstrate that a *spatially explicit* model [9], i.e., taking nearby places into account when predicting the place type from an image, improves the accuracy of leading image classification models by at least 40%. Aside from this substantial increase in accuracy, we believe that our work also contributes to the broader and ongoing discussion about the role of and need for theory, i.e., domain knowledge, in machine learning. Finally, and as indicated in the title, our spatial context (*SC*) models, can be added to any of the popular CNN-based computer vision models such as AlexNet, ResNet, and DenseNet – abbreviated to *xNet* here.

The remainder of this paper is organized as follows. Section 2 provides an overview of existing work on spatial context and methods for incorporating spatial information into image classification models. Section 3 presents the image classification tasks and provides information about the convolutional neural network models used in our study. Section 4 explains in detail three different levels of spatial context and ways to combine them in image classification models. Section 5 presents the results. Finally, Section 6 concludes the research and points to future directions.

2 Related Work

There is a large body of work that utilizes spatial context to improve existing methods and provide deeper insights into the rich semantics of contextual information more broadly. For instance, spatial context has been recognized as a complementary source of information in computational linguistics. By training word embeddings for different place types derived from OpenStreetMap (OSM) and Google Places, Cocos and Callison-Burch [7] suggested that spatial context provides useful information about semantic relatedness. In Points of Interest (POI) recommendation, spatial context has been used to provide latent representations of POI, to facilitate the prediction of future visitors [8], and to recommend similar places [34]. By implementing an information theoretic and distance-lagged augmented spatial context, Yan et al. [30] demonstrated that high-dimensional place type embeddings learned using spatial contexts can reproduce human-level similarity judgments with high accuracy. The study showed that such a spatially explicit Place2Vec model substantially outperforms Word2Vec-based models that utilize a linguistic-style of context. Liu et al. [21] used spatial contexts to measure traffic interactions in urban area. In object detection, Heitz and Koller [13] leveraged spatial contexts in a probabilistic model to improve detection result. Likewise, by embracing the idea that spatial context provides valuable extrinsic signals, our work analyzes different kinds of spatial contexts and tests their ability to improve image classification of place types.

Existing work on image classification has realized the importance of including a geographic component. One direction of research focused on enriching images with geospatial data. Baatz et al. [1] took advantage of digital elevation models to help geo-localize images in mountainous terrain. Lin et al. [20] made use of land cover survey data and learned the complex translation relationship between ground level images and overhead imagery to extend the reach of image geo-localization. Instead of estimating a precise geo-tag, Lee et al. [19] trained deep convolutional neural networks to enrich a photo with geographic attributes such as elevation and population density. Another direction of research (which is more similar to our study) focused on utilizing geographic information to facilitate image classification. In order to better understand scenes and improve object region recognition, Yu and Luo [31] exploited information from seasons and location proximity of images using a probabilistic graphical model. Berg et al. [2] combined one-vs-most image classifiers with spatiotemporal class priors to address the problem of distinguishing images of highly similar bird species.

Tang et al. [29] encoded geographic features extracted from GPS information of images into convolutional neural networks to improve classification results.

Our work differs from the existing work in that we explicitly exploit the distributional semantics found in spatial context [30] to improve image classification. Following the linguistic mantra that one *shall know a word by the company it keeps*, we argue that one can know a place type by its neighborhood’s types. This raises the interesting question of how such a neighborhood should be defined. We will demonstrate different ways in which spatial contextual signals and visual signals can be combined. We will assess to what extent different kinds of spatial context, namely spatial relatedness, spatial co-location, and spatial sequence pattern, can provide such neighborhood information to benefit image classification.

3 Image Classification

In this section, we first describe the image classification task and the data we use. The task is similar to scene classification but we are specifically interested in classifying different business venues as opposed to natural environment. Then we explain four different deep convolutional neural networks that solely leverages the visual signals of images. These convolutional neural network models are later used as baselines for our experiment.

3.1 Classification Task

Our task is to classify images into one of the several candidate place types. Because we want to utilize the spatial context in which the image was taken, we need to make sure each image has a geographic identifier, e.g. geographic coordinates, so that we are able to determine its neighboring place and their types. In order to classify place types of images, we consider the scene categories provided by Zhou et al. [35] as they also provide pretrained models (Places365-CNN) that we can directly use.³ Without losing generality, we select 15 place types as our candidate class labels. The full list of class labels and their alignment with the categories in Places365-CNN is shown in Table 1. For each candidate class, we selected 50 images taken in 8 states⁴ within the US by using Google Maps, Google Street View, and Yelp. These images include both indoor and outdoor views of each place type. Please note that classifying place types from facade and interior images is a hard problem and even the most sophisticated models only distinguish a relatively small number of place types so far which is nowhere near the approximately 420 types provided by sources such as Foursquare. Places365, for instance, offers 365 classes but many of these are scenes or landscape features, such as waves, and not POI type, such as cinemas, in the classical sense.

3.2 Convolutional Neural Network Models

To establish baselines for our study, we selected several state-of-the-art image classification models, namely deep convolutional neural networks. Unlike traditional image classification pipelines, CNNs extract features from images automatically based on the error messages that are backpropagated through the network, thus fewer heuristics and less manual labor are needed. Contrary to densely connected feedforward neural networks, CNN adopts parameter sharing to extract common patterns which help capture translation invariance and creates sparse connections which result in fewer parameters and being less prone to overfitting.

³ https://github.com/CSAILVision/places365/blob/master/categories_places365.txt

⁴ Arizona, Illinois, Nevada, North Carolina, Ohio, Pennsylvania, South Carolina, and Wisconsin

■ **Table 1** Class label alignment between Yelp and the Places365 model.

Class label	Places365-CNN category
Amusement Parks	amusement_park
Bakeries	bakery
Bookstores	bookstore
Churches	church
Cinema	movie_theater
Dance Clubs	discotheque
Drugstores	drugstore, pharmacy
Hospitals	hospital, hospital_room
Hotels	hotel, hotel_room
Jewelry	jewelry_shop
Libraries	library
Museums	museum, natural_history_museum, science_museum
Restaurants	fastfood_restaurant, restaurant, restaurant_kitchen, restaurant_patio
Shoe Stores	shoe_shop
Stadiums & Arenas	stadium

The architecture of CNNs has been revised numerous times and has become increasingly sophisticated since its first appearance about 30 years ago. These improvements in architecture have made CNN more powerful as can be seen in the ImageNet challenge. Some of the notable architectures include: LeNet [18], AlexNet [17], VGG [24], Inception [27], ResNet [12], and DenseNet [15]. We selected AlexNet, ResNet with 18 layers (ResNet18), ResNet with 50 layers (ResNet50), and DenseNet with 161 layers (DenseNet161). AlexNet is among the first deep neural networks that increased the classification accuracy on ImageNet by a significant amount compared with traditional classification approaches. By using skip connections to create residual blocks in the network, ResNet makes it easy to learn identity functions that help with the vanishing and exploding gradient problems when the network goes deeper. In DenseNet, a dense connectivity pattern is created by connecting every two layers so that the error signal can be directly propagated to earlier layers, parameter and computational efficiency can be increased, and low complexity features can be maintained [15]. These models were trained on 1.8 million images from the Places365-CNN dataset. We used the pretrained weights for these models.

4 Spatial Contextual Information

In this section, we introduce three different kinds of spatial contexts and explore ways in which we can combine them with the CNN models in order to improve image classification. The first type of spatial context is spatial relatedness, which measures the extend to which different place types relate with each other. The second type of spatial context is spatial co-location, which considers what place types tend to co-occur in space and the frequency they cluster with each other. The third type of spatial context is spatial sequence pattern which considers both spatial relatedness and spatial co-location. In addition, spatial sequence pattern considers the interaction between context place types and the inverse relationship between distance and contextual influence. We use POIs provided by Yelp as dataset.⁵

⁵ <https://www.yelp.com/dataset>

4.1 Spatial Relatedness

Since the output of CNN is the probability score for each class label, it is possible to interpret our task as a ranking problem: given an image, rank the candidate class labels based upon the visual signal and spatial context signal. For the visual signal, we can obtain the ranking scores (probability scores) from the CNN architectures mentioned in Section 3. Since the original CNN models has 365 labels, we renormalize the probability scores for each candidate place type by the sum of the 15 candidate ranking scores so that they sum up to 1. This renormalization procedure is also applied to the other two spatial context methods explained in Section 4.2 and Section 4.3. We will refer to the renormalized scores as CNN scores in this study. For the spatial context signal, the ranking scores are calculated using the place type embeddings proposed in [30]. These embeddings capture the semantics of different place types and can be used to measure their similarity and relatedness. In this regard, the task is equivalent to a re-ranking problem, which adjusts the initial ranking provided by the visual signal using auxiliary knowledge, namely the spatial context signal. Intuitively, the extent to which the visual signals from the images match with different place types and the level of relevance of the surrounding place types with respect to candidate place types jointly determine the final result.

Inspired by search re-ranking algorithms in information retrieval, we use a *Linear Bimodal Fusion* (LBF) method (here essentially a 2-component convex combination), which linearly combines the ranking scores provided by the CNN model and the spatial relatedness scores, as shown in Equation 1.

$$s_i = \omega^v s_i^v + \omega^r s_i^r \quad (1)$$

where s_i , s_i^v , and s_i^r are the LBF score, CNN score, and spatial relatedness score for place type i respectively, ω^v and ω^r are the weights for the CNN component and spatial relatedness component, and $\omega^v + \omega^r = 1$. The weights here are decided based on the relative performance of individual components. Specifically, the weight is determined using Equation 2.

$$\omega^v = \frac{acc^v}{acc^v + acc^r} \quad (2)$$

where acc^v and acc^r are the accuracies for CNN and spatial relatedness measurements for the image classification task. Intuitively, this means that we have higher confidence if the component performs better on its own and want to reflect such confidence using the weight in the LBF score.

In order to calculate the spatial relatedness scores, we use cosine similarity to measure the extend to which each candidate class embedding is related with the spatial context embedding of an image in a high dimensional geospatial semantic feature space. Following the suggestions in [30], we use a concatenated vector of 350 dimensions (i.e., 70D vectors for each of 5 distance bins) as the place type embeddings. The candidate class embeddings can be retrieved directly. Then we search for the nearest n POIs based on the image location, determine the place types of these n POIs, and calculate the average of these place type embeddings as the final spatial context embeddings for images. The cosine similarity score sm_i is calculated between the spatial context embedding of an image and the embedding of each candidate place type class i . Because sm_i ranges from -1 to 1, we use min-max normalization to scale the values to $[0, 1]$. Finally, we apply the same renormalization as for the CNN score to turn the normalized score sm_i' into probability score, i.e. spatial relatedness score s_i^r .

Combining these normalizations together with Equation 1 and Equation 2, we are able to derive that $0 \leq s_i \leq 1$ and $\sum_{i=1}^N s_i = 1$ where $N = 15$ in our case. This means that the LBF score s_i can be considered a probability score.

4.2 Spatial Co-location

The spatial relatedness approach follows the assumption that relatedness implies likelihood which is reasonable in cases where similar place types cluster together, such as restaurant, bar, and hotel. However, in cases of high spatial heterogeneity, this assumption will fall short of correctly capturing the true likelihood. An example would be places of dissimilar types that co-occur, e.g., grocery stores and gas stations. Moreover, the LBF method can only capture a linear relationship between the two signals.

Following Berg et al.[2], we also test a Bayesian approach in which we assume there is a complex latent distribution of the data that facilitates our classification task. Intuitively, the CNN score gives us the probability of each candidate class t given the image I , i.e., $P(t|I)$, and the spatial context informs us of the probability of each candidate class given its neighbors $c_1, c_2, c_3, \dots, c_n$, denoted as C , around the image location, i.e., $P(t|C)$. We would like to obtain the posterior probability of each candidate class given both the image and its spatial context, i.e., $P(t|I, C)$. Using Bayes' theorem, the posterior probability can be written as:

$$P(t|I, C) = \frac{P(I, C|t)P(t)}{P(I, C)} \quad (3)$$

For variables I , C , and t , we construct their dependencies using a simple probabilistic graphical model, i.e., Bayesian network, which assumes that both the image I and the spatial context C are dependent on the place type t , which intuitively makes sense in that different place types will result in different images and different place types of their neighbors. We know that given information about the image I we are able to update our beliefs, i.e., the probability distributions, about the place type t . In addition, the changes in our beliefs about the place type t can influence the probability distributions of the spatial context C . However, if place type t is observed, the influence cannot flow between I and C , thus we are able to derive the conditional independence of I and C given t . So Equation 3 can be rewritten as:

$$\begin{aligned} P(t|I, C) &= \frac{P(I|t)P(C|t)P(t)}{P(I, C)} \\ &= \frac{P(t|I)P(I)}{P(t)} \frac{P(t|C)P(C)}{P(t)} \frac{P(t)}{P(I, C)} \\ &\propto \frac{P(t|I)}{P(t)} P(t|C) \end{aligned} \quad (4)$$

in which we have dropped all the factors that are not dependent on t as they can be considered as normalizing constants for our probabilities. It follows that the posterior probability $P(t|I, C)$ can be computed using the CNN probability score $P(t|I)$, the spatial context prior $P(t|C)$, and the candidate class prior $P(t)$. Instead of estimating the distribution of spatial context priors, we take advantage of the spatial co-location patterns and calculate the prior probabilities using the Yelp POI data directly. As mentioned earlier, the spatial context C is composed of multiple individual context neighbors $c_1, c_2, c_3, \dots, c_n$; hence, we need to calculate $P(t|c_1, c_2, c_3, \dots, c_n)$. In order to simplify our calculation, we impose a bag-of-words assumption as well as a Naive Bayes assumption in the spatial co-location patterns. The bag-of-words assumption simplifies the model by assuming that the position (or the order) in

which different context POIs occur does not play a role. The Naive Bayes assumption implies that the only relationship is the pair-wise interaction between the candidate place type t and an individual neighbor's place type c_i and there is no interaction between neighboring places wrt. their types, i.e. $(c_i \perp\!\!\!\perp c_j | t)$ for all c_i, c_j . Using spatial co-location, we are able to calculate the conditional probability using place type co-location counts $P(c_i | t) = \frac{\text{count}(c_i, t)}{\text{count}(t)}$ where $\text{count}(c_i, t)$ is the frequency that neighbor type c_i and candidate type t co-locate within a certain distance limit and $\text{count}(t)$ is the frequency of candidate type t in the study area. Combining all these components, we can derive:

$$\begin{aligned} P(t|C) &= P(t|c_1, c_2, \dots, c_n) \\ &= \frac{P(t) \prod_{i=1}^n P(c_i|t)}{P(c_1, c_2, c_3, \dots, c_n)} \\ &= \frac{P(t)}{P(c_1, c_2, c_3, \dots, c_n)} \frac{\prod_{i=1}^n \text{count}(c_i, t)}{\text{count}(t)^n} \end{aligned} \quad (5)$$

Using Equation 4 and Equation 5, we can derive the final formula for calculating $P(t|I, C)$ shown in Equation 6. For the sake of numerical stability, we calculate the log probability $\log P(t|I, C)$ using the natural logarithm. Since the natural logarithm is a monotonically increasing function, it will not affect the final ranking of the classification results.

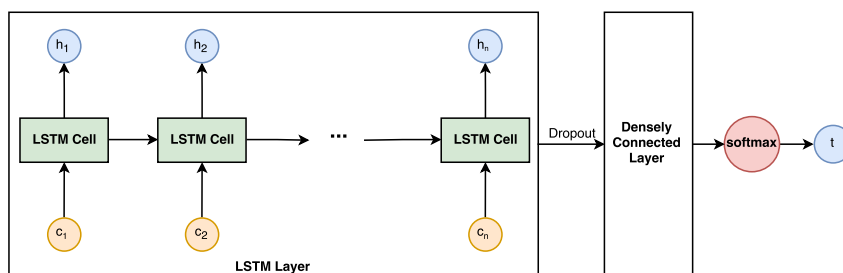
$$\begin{aligned} \log P(t|I, C) &\propto \log \left(\frac{P(t|I)}{P(t)} P(t|C) \right) \\ &= \log \left(\frac{P(t|I)}{P(c_1, c_2, c_3, \dots, c_n)} \frac{\prod_{i=1}^n \text{count}(c_i, t)}{\text{count}(t)^n} \right) \\ &\propto \log P(t|I) + \sum_{i=1}^n \log(\text{count}(c_i, t)) - n \log(\text{count}(t)) \end{aligned} \quad (6)$$

where we also drop $P(c_1, c_2, c_3, \dots, c_n)$ as it does not depend on t , so it will not affect the result ranking. The log posterior probability is then used to generate the final ranking of candidate place types and produce the classification results.

4.3 Spatial Sequence Pattern

The spatial co-location approach follows the bag-of-words assumption that the position of spatial context POIs does not matter and the Naive Bayes assumption that the context neighbors are independent of each other. However, in many cases this assumption is too strong. In fact, numerous methods, such as Kriging and multiple-point geostatistics, have been devised to model geospatial proximity patterns and complex spatial interaction patterns. However, incorporating these complex spatial patterns in a multidimensional space would adversely affect the model complexity and make the distribution in Section 4.2 intractable. In order to strike the right balance between the complexity of model and the integrity of spatial context pattern, we propose to capture the spatial sequence pattern in our model by collapsing the 2D geographic space into a 1D sequence.

Specifically, we use the Long Short-Term Memory (LSTM) network model, a variant of recurrent neural network (RNN), in our study. Recurrent neural networks are frequently used models to capture the patterns in sequence or time series data. In theory, the naive recurrent neural networks can capture long term dependencies in the sequence, however, due to the vanishing and exploding gradient problem, they fail to do so in practice. LSTM is explicitly designed to solve the problem by maintaining a cell state and controlling the



■ **Figure 1** Structure of the LSTM.

input and output flow using forget gate, input gate, and output gate [14]. We use LSTM as a generative model in order to capture the latent distribution of place types using the spatial sequence pattern. In the training stage, the input is a sequence of context place types $c_1, c_2, c_3, \dots, c_n$ and the output is the place type t of the POI from which the context is created. The input sequence is ordered in a way so that the previous one is further away from the output than the next one in the collapsed 1D space. Image one would drive around a neighborhood before reaching a destination. For each of the POIs encountered during the route, one would update the beliefs about the neighborhood by considering the current POI and all previously seen POIs. Upon arriving at the destination, one would have a reasonable chance of guessing this final POI's type. The structure of the LSTM model is shown in Figure 1. We apply a dropout after the LSTM layer to avoid overfitting. After training the LSTM model on Yelp's POI dataset, we are able to obtain the spatial context prior $P(t|c_1, c_2, c_3, \dots, c_n)$ based on the spatial sequence pattern around the image locations in our test data. We specifically removed the image locations and their context in the training data. Similar to the spatial co-location approach, we use Bayesian inference and log probability to calculate the final result:

$$\begin{aligned} \log P(t|I, C) &\propto \log \left(\frac{P(t|I)}{P(t)} P(t|C) \right) \\ &= \log P(t|I) + \log P(t|c_1, c_2, c_3, \dots, c_n) - \log P(t) \end{aligned} \quad (7)$$

where the candidate class prior $P(t)$ can be computed using the Yelp data. Since we use LSTM as a generative model, in the prediction phase, sampling strategies, such as greedy search, beam search, and random sampling, can be applied based on the distribution provided by the output of the LSTM prediction. However, we only generate the next prediction instead of a sequence, so we do not apply these sampling strategies. Instead, we make use of the hyperparameter *temperature* τ to adjust the probability scores returned by the LSTM model before combining them with the CNN model in a Bayesian manner. Including the hyperparameter τ , the softmax function in the LSTM model can be written as:

$$P(t_i|C) = \frac{\exp(\frac{\text{logit}_i}{\tau})}{\sum_{j=1}^N \exp(\frac{\text{logit}_j}{\tau})} \quad (8)$$

where logit_i is the logit output provided by LSTM before applying the softmax function and $N = 15$ in our case. Intuitively, when the temperature τ is high, i.e., $\tau \rightarrow \infty$, the probability distribution will become diffuse and $P(t_i|C)$ will have almost the same value for different t_i ; when τ is low, i.e., $\tau \rightarrow 0^+$, the distribution becomes peaky and the largest logit_i stands out to have a probability close to 1. This idea is closely related to the exploration and exploitation trade-off in many machine learning problems. The value of τ will affect the probability scores $P(t_i|C)$ but not the ranking of these probabilities.

In this study, we propose two ways to model the 2D geographic space as a 1D sequence. The first one is a distance-based ordering approach. For any given POI, we search for nearby POIs within a certain distance from it, choose the closest n POIs, and rearrange them by distance with descending order, thereby forming a 1D array. This distance-based method is isotropic in that it does not differentiate between directions while creating the sequence. The second method is a space filling curve-based approach. We utilize *Morton order* here which is also used in geohashing to encode coordinates into an indexing string that can preserve the locality of spatial locations. We use Morton order to encode the geographic locations of every POI and order them in a sequence based upon their encodings, i.e., indexing sequence. After obtaining the sequence, for each POI, we use the previous n POI in the sequence as the context sequence. Other space filling curves could be used in future work.

Because each POI can have multiple place types associated with it, e.g., restaurant and beer garden, the sequence of place types is usually not unique for the same sequence of POIs. As our LSTM input is a sequence of place *types*, we compute the Cartesian product of all POI type sets in the sequence of nearby places:

$$T_{c_1} \times T_{c_2} \times T_{c_3} \times \dots \times T_{c_n} = \{(t_{c_1}, t_{c_2}, t_{c_3}, \dots, t_{c_n}) | \forall i = 1, 2, 3, \dots, n, t_{c_i} \in T_{c_i}\} \quad (9)$$

where T_{c_i} is the set of place types associated with POI c_i in the context sequence. In practice, however, we randomly sample a fixed number of place type sequences from each of the Cartesian product for the POI context sequence as the potential combinations grow exponentially with increasing context size.

5 Experiment and Result

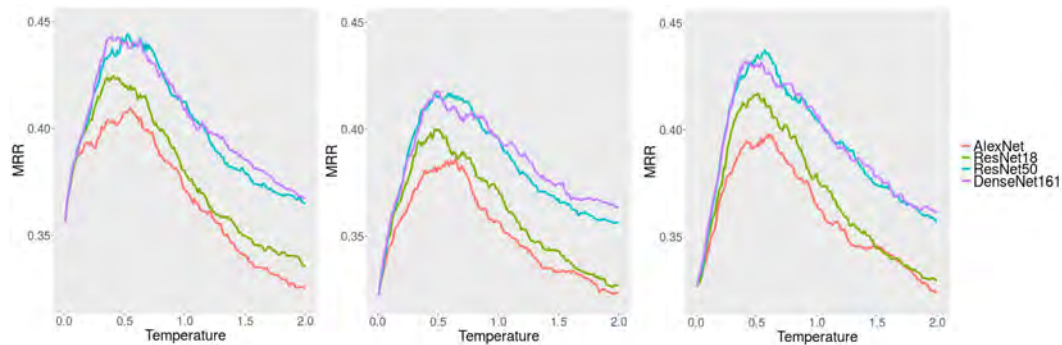
In this section, we explain our experimental setup for the models described above, describe the metrics used to compare the model performance for place type image classification, and present the results and findings.

5.1 Implementation Details

For all three types of spatial context, we use 10 as the maximum number of context POIs and a distance limit of 1000m for the context POI search. For the spatial sequence pattern approach, we use a fixed sample size of 50 to sample from the Cartesian product of all POI type sets in the sequence.⁶ We use a one-layer LSTM with 64 hidden units. We train our LSTM model using the recommended Root Mean Square Propagation (RMSProp) optimizer with a learning rate of 0.005. A dropout ratio of 0.2 is applied in the LSTM and we run 100 epochs. The same settings are used for all LSTM trainings in our experiment. The total number of POI in the dataset is 115,532, yielding more than 5 million unique training sequences.

For evaluation, we use three different metrics, namely Mean Reciprocal Rank (MRR), Accuracy@1, and Accuracy@5. Another common metric for image classification would also be Mean Average Precision (MAP), but since there is only one true label per type in our task, we use MRR instead.

⁶ The median for types per place in Yelp is 3.



■ **Figure 2** From left to right, MRR result using distance-based sequence, random sequence, and Morton code-based sequence with varying temperatures

5.2 Results

We run the 750 test images we collected, i.e., 50 images per each of 15 types, on the four CNN baseline models (AlexNet, ResNet18, ResNet50, and DenseNet161) as well as the combined models using our three different types of spatial context.⁷ In addition to the two methods for converting geographic space into 1D sequences in the spatial sequence pattern approach, we also test one model using random sequences with the same context count and distance limits. We did so to study whether results obtained using the LSTM would benefit from distance-based spatial contexts. A higher result for the spatial sequence based LSTM over the random LSTM would indicate that the network indeed picked up on the distance signal.

The hyperparameter τ can be adjusted; a value of 0.5 has been proposed as a good choice before. In order to test this and find the optimal temperature value, we run the combined model using spatial sequence patterns with three types of sequencing approaches, namely random sequence, distance-based sequence, and Morton order-based sequence.

We test temperature values ranging from 0.01 to 2 with a step of 0.01. We combine the spatial sequence pattern models with all CNN models. The MRR result with respect to temperature are shown in Figure 2. Although there are a slight variations, the MRR curves all reach their peaks around a τ value of 0.5. This confirms the suggestion from the literature. Figure 3 shows selected example predictions. The results for MRR, Accuracy@1, and Accuracy@5 using the baseline models as well as our proposed, spatially explicit models are shown in Table 2, Table 3, and Table 4.⁸

As we can see, by incorporating spatial context in the image classification model, we are able to improve the classification result in general. However, integrating spatial relatedness using the LBF method does not seem to affect the result. This essentially confirms our aforementioned assumption that relatedness does not always imply likelihood. The benefit of incorporating spatial relatedness in cases of spatial homogeneity are likely to be offset by cases of high spatial heterogeneity in which spatial relatedness may have a negative effect as dissimilar places co-occur.

⁷ Transfer learning could be applied to fine tune the CNN models first, but we only have limited images and our hypothesis is that spatial context can be used as a powerful complement or alternative to the visual component for image classification.

⁸ The baseline models are not comparable with a random classifier which would yield an expected accuracy of 1/15 in this case, because the baseline CNN models have 365 unique labels and we choose 15 labels in our experiment.



■ **Figure 3** From left to right, images of a restaurant, a hotel, and a museum from Yelp, Google Street View, and Google Maps respectively. The first image is incorrectly classified as library using all 4 CNN models and it is correctly classified as restaurant using the spatial sequence pattern (distance) models. The second image is classified as hospital and library by the original CNN models and is classified as hotel by the spatial sequence pattern (distance) models. For the third image the correct label museum is in the third position in the label rankings of all 4 CNN models while, using the spatial sequence pattern (distance) models, ResNet18 and ResNet50 can correctly label it and in the label rankings of AlexNet and DenseNet161 museum is in the second position.

■ **Table 2** MRR result using baseline models and proposed combination models using different types of spatial context and sequences

MRR	AlexNet	ResNet18	ResNet50	DenseNet161
Baseline	0.27	0.28	0.31	0.31
Relatedness	0.27	0.28	0.31	0.32
Co-location	0.30	0.31	0.31	0.32
Sequence Pattern (Random)	0.38	0.40	0.42	0.42
Sequence Pattern (Distance)	0.41	0.42	0.44	0.44
Sequence Pattern (Morton order)	0.39	0.42	0.43	0.43

The Accuracy@1 measurement is improved by incorporating spatial co-location component in the models. This confirms our previous reasoning that considering the external signal, namely spatial contexts, and assuming a complex latent distribution of the data in a Bayesian manner improve image classification. However, for MRR the improvement is marginal and for Accuracy@5 there even is a decrease after incorporating the spatial co-location component because this type of spatial context falls short of taking into account the intricate *interactions* of different context neighbors. This shortcoming is not clear when only looking at the first few results in the ranking returned by the combined models, but it becomes clearer in later results in the ranking output, thus resulting in a decrease for Accuracy@5 and only a slight increase in the MRR measurement.

The Bayesian combination model using spatial sequence patterns shows better overall results compared with the baseline models, the spatial relatedness model, and the spatial co-location model. This is because the spatial sequence patterns capture spatial interactions between the neighboring POIs that are neglected by the other models. From the result we can see that using a distance-based sequence is better than using a random sequence. To prevent confusion and to understand why the random model still performs relatively well, it is important to remember that this model utilizes spatial context. However, it does not utilize the distance signal within this context but merely the presence of neighboring POI. The results show that a richer spatially explicit context, one that comes with a notion of *distance decay*, indeed improves classification results. Interestingly, the sequence using Morton order, which is widely used in geohashing techniques, does not further improve the result compared to the distance-based sequence. There may be multiple reasons for this. First, we may have reached a ceiling of possible improvements by incorporating spatial contexts. Second, our Morton order implementation takes the 10 places that precede the target place in the index.

■ **Table 3** Accuracy@1 result using baseline models and proposed combination models using different types of spatial context and sequences

Accuracy@1	AlexNet	ResNet18	ResNet50	DenseNet161
Baseline	0.07	0.07	0.09	0.09
Relatedness	0.07	0.07	0.09	0.09
Co-location	0.15	0.17	0.17	0.17
Sequence Pattern (Random)	0.18	0.18	0.19	0.20
Sequence Pattern (Distance)	0.20	0.20	0.22	0.22
Sequence Pattern (Morton order)	0.19	0.20	0.22	0.22

■ **Table 4** Accuracy@5 result using baseline models and proposed combination models using different types of spatial context and sequences

Accuracy@5	AlexNet	ResNet18	ResNet50	DenseNet161
Baseline	0.50	0.56	0.59	0.60
Relatedness	0.52	0.56	0.58	0.59
Co-location	0.42	0.44	0.45	0.44
Sequence Pattern (Random)	0.65	0.69	0.73	0.73
Sequence Pattern (Distance)	0.67	0.70	0.73	0.75
Sequence Pattern (Morton order)	0.65	0.70	0.72	0.71

This may result in directional effects. Finally, all space filling curves essentially introduce different ways to preserve local neighborhoods; utilizing another technique such as Hilbert curves may yield different results. Given that the Morton order-based sequence in many cases yield results of equal quality to the distance-based sequences, further work is needed to test the aforementioned ideas.

Summing up, the results demonstrate that incorporating a (distance-based) spatial context improves the MRR of state-of-the-art image classification systems by over **40%**. The results for Accuracy@1 are more than **doubled** which is of particular importance for humans as this measure only considers the first ranked result.

6 Conclusion and Future Work

In this work, we demonstrated that utilizing spatial contexts for classifying places based on images of their facades and interiors leads to substantial improvements, e.g., increasing MRR by over 40% and doubling Accuracy@1, compared to applying state-of-the-art computer vision models such as ResNet50 and DenseNet161 alone. These advances are especially significant as the classification of places based on their images remains a hard problem. One could argue that our proposal requires additional information, namely about the types of nearby places. However, such data are readily available for POI, and only a few nearby places are needed. Secondly, and as a task for future work, one could also modify our methods to work in a *drive-by-typing* mode in which previously seen places are classified, and these classification results together with their associated classification uncertainty are used to improve estimation of the currently seen place, thereby relaxing the need for POI datasets. In the future, we would like to apply transfer learning and experiment with other ways to encode spatial contexts, e.g., by testing different space-filling curves. We plan to develop models to directly capture 2D spatial patterns rather than using a 1D sequence as a proxy and test whether spatial contexts also aid in recognizing objects beyond places and their facades.

References

- 1 Georges Baatz, Olivier Saurer, Kevin Köser, and Marc Pollefeys. Large scale visual geolocalization of images in mountainous terrain. In *Computer Vision–ECCV 2012*, pages 517–530. Springer, 2012.
- 2 Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2019–2026. IEEE, 2014.
- 3 Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357, 2016.
- 4 Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.
- 5 Marco Castelluccio, Giovanni Poggi, Carlo Sansone, and Luisa Verdoliva. Land use classification in remote sensing images by convolutional neural networks. *arXiv preprint arXiv:1508.00092*, 2015.
- 6 Moustapha Cisse, Yossi Adi, Natalia Neverova, and Joseph Keshet. Houdini: Fooling deep structured prediction models. *arXiv preprint arXiv:1707.05373*, 2017.
- 7 Anne Cocos and Chris Callison-Burch. The language of place: Semantic value from geospatial context. In *15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2, pages 99–104, 2017.
- 8 Shanshan Feng, Gao Cong, Bo An, and Yeow Meng Chee. Poi2vec: Geographical latent representation for predicting future visitors. In *AAAI*, pages 102–108, 2017.
- 9 Michael F Goodchild and Donald G Janelle. Thinking spatially in the social sciences. *Spatially integrated social science*, pages 3–22, 2004.
- 10 Ian J Goodfellow, Yaroslav Bulatov, Julian Ibarz, Sacha Arnoud, and Vinay Shet. Multi-digit number recognition from street view imagery using deep convolutional neural networks. *arXiv preprint arXiv:1312.6082*, 2013.
- 11 Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- 12 Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- 13 Jeremy Heitz and Daphne Koller. Learning spatial context: Using stuff to find things. In *European conference on computer vision*, pages 30–43. Springer, 2008.
- 14 Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- 15 Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2261–2269, 2017.
- 16 Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems*, pages 2280–2288, 2016.
- 17 Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- 18 Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

- 19 Stefan Lee, Haipeng Zhang, and David J Crandall. Predicting geo-informative attributes in large-scale image collections using convolutional neural networks. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pages 550–557. IEEE, 2015.
- 20 Tsung-Yi Lin, Serge Belongie, and James Hays. Cross-view image geolocalization. In *Computer Vision and Pattern Recognition*, pages 891–898. IEEE, 2013.
- 21 Kang Liu, Song Gao, Peiyuan Qiu, Xiliang Liu, Bo Yan, and Feng Lu. Road2vec: Measuring traffic interactions in urban road system from massive travel routes. *ISPRS International Journal of Geo-Information*, 6(11):321, 2017.
- 22 Nikhil Naik, Scott Duke Kominers, Ramesh Raskar, Edward L Glaeser, and César A Hidalgo. Computer vision uncovers predictors of physical urban change. *Proceedings of the National Academy of Sciences*, 114(29):7571–7576, 2017.
- 23 Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pages 582–597. IEEE, 2016.
- 24 Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- 25 Pierre Stock and Moustapha Cisse. Convnets and imagenet beyond accuracy: Explanations, bias detection, adversarial examples and model criticism. *arXiv:1711.11443*, 2017.
- 26 Wanxiao Sun, Volker Heidt, Peng Gong, and Gang Xu. Information fusion for rural land-use classification with high-resolution satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 41(4):883–890, 2003.
- 27 Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, et al. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015.
- 28 Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
- 29 Kevin Tang, Manohar Paluri, Li Fei-Fei, Rob Fergus, and Lubomir Bourdev. Improving image classification with location context. In *Proceedings of the IEEE international conference on computer vision*, pages 1008–1016, 2015.
- 30 Bo Yan, Krzysztof Janowicz, Gengchen Mai, and Song Gao. From itdl to place2vec—reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. *Proceedings of SIGSPATIAL*, 17:7–10, 2017.
- 31 Jie Yu and Jiebo Luo. Leveraging probabilistic season and location context models for scene understanding. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 169–178. ACM, 2008.
- 32 Andi Zang, Runsheng Xu, Zichen Li, and David Doria. Lane boundary extraction from satellite imagery. In *Proceedings of the 1st ACM SIGSPATIAL Workshop on High-Precision Maps and Intelligent Applications for Autonomous Vehicles*, page 1. ACM, 2017.
- 33 Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.
- 34 Shenglin Zhao, Tong Zhao, Irwin King, and Michael R Lyu. Geo-teaser: Geo-temporal sequential embedding rank for point-of-interest recommendation. In *Proceedings of the 26th international conference on world wide web companion*, pages 153–162. International World Wide Web Conferences Steering Committee, 2017.
- 35 Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

A Critical Look at Cryptogovernance of the Real World: Challenges for Spatial Representation and Uncertainty on the Blockchain

Benjamin Adams

Department of Geography, University of Canterbury, New Zealand
benjamin.adams@canterbury.ac.nz

Martin Tomko

Department of Infrastructure Engineering, University of Melbourne, Australia
tomkom@unimelb.edu.au

Abstract

Innovation in distributed ledger technologies—blockchains and smart contracts—has been lauded as a game-changer for environmental governance and transparency. Here we critically consider how problems related to spatial representation and uncertainty complicate the picture, focusing on two cases. The first regards the impact of uncertainty on the transfer of spatial assets, and the second regards its impact on smart contract code that relies on software oracles that report sensor measurements of the physical world. Cryptogovernance of the environment will require substantial research on both these fronts if it is to become a reality.

2012 ACM Subject Classification Information systems → Spatial-temporal systems, Applied computing → Environmental sciences, Social and professional topics → Socio-technical systems

Keywords and phrases spatial information, spatial uncertainty, blockchain, smart contract, environmental management

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.18

Category Short Paper

1 Introduction

Distributed ledger technologies, such as blockchains, have generated tremendous interest of late, because of their ability to support peer-to-peer transactions of digital assets. The first and still most notable public blockchain is the distributed ledger of Bitcoin transactions [10]. Yet, the discussion of distributed ledgers needs to go beyond this particular example. Blockchain technology is based on a distributed consensus algorithm—such as proof of work—which ensures that the ledgers cannot be corrupted by bad actors. As a consequence, the transactions in such distributed ledgers are *trustless*, meaning that the system works to verify transactions between participants who might not trust or even know each other.

Chapron, in his Utopian vision of cryptogovernance [3], makes a number of strong claims about the benefits of distributed ledgers with respect to “wins” for ownership, traceability, incentives, and governance of the environment. Despite sharing enthusiasm for technological advancement, we take a more skeptical view toward benefits of distributed ledgers and environmental cryptogovernance. In this paper, we initiate the discussion of the potential pitfalls of automated smart contracts supported by distributed ledgers relating to the physical environment. Our spatial perspective can take at least two aspects – through *spatial assets* being the subject of transactions, or the *spatial context* (of one or multiple transaction parties) acting as the enabler of the transaction. We highlight why distributed ledgers cannot be



© Benjamin Adams and Martin Tomko;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 18; pp. 18:1–18:6

Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

decoupled from the particular characteristics of environmental and land assets, the limitations of the technology that is used to sense the environment, and societal needs and degree of digital literacy.

2 A Brief Introduction to Distributed Ledgers and Smart Contracts

The consensus algorithms underpinning transactions in blockchain are the foundation of the technology. They need to be highly robust so as not to be easily corruptible [7]. For example, in the case of Bitcoin, no one has yet successfully corrupted the public ledger of transactions. This removes the need for a third-party notary to mediate transactions and enables the distributed characteristic of the ledger. For a standard blockchain, the kinds of transactions supported are fixed to a particular type. For example, for Bitcoin the ledger records the transfer of Bitcoin currency from one account to another.

The innovation of *smart contracts* has expanded the potential of blockchains by introducing a method of encoding scripts or programmable code onto distributed ledgers [13, 2]. These scripts must execute only once certain conditions are met. Any arbitrarily complex combination of computable rules can be defined to test that certain conditions are met.

Once the conditions are met, the contract will automatically execute the transaction. The main limitation is that whatever is being transferred must be tokenizable in digital form. For assets that are easily digitized and tokenized, such as money, the potential of the technology is clear. For rules, regulations, and laws that can be formalized into unambiguous algorithms a smart contract can, in theory, fully automate complex chained management and transaction of assets, thus replacing the need for third-party actors or escrow to complete the process.

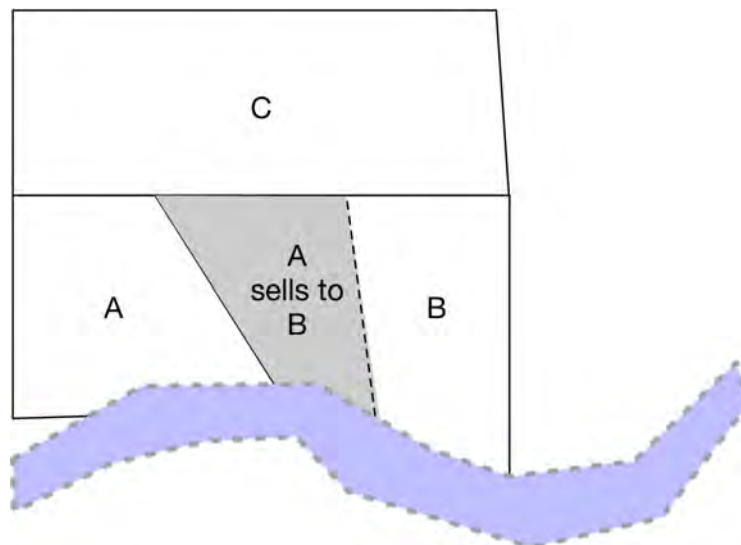
There has been of late an emergence of proposals to apply blockchain and smart contract technology to problems that require digital representations of the spatial attributes of real world objects, either to support the transaction of physical and environmental assets or to detect spatio-temporal events that trigger execution of smart contract code. Recently proposed spatial applications of blockchain technology include the internet of things [4], transport networks and smart cities [17, 11], land registration and administration [1], governance of the environment [3], and timber supply chain tracking [5].

When the transfer concerns only *digital representations* of physical assets (such as land parcels), however, a number complications arise. Unsurprisingly, questions around uncertainty in the spatial representation of the physical objects arise—a fact that has been long-studied in geographic information science [6, 12]. This is the first aspect of distributed ledgers discussed in this paper and one that – we believe – has not been considered critically enough.

Recall the conditions that have to be met for a contract to be executed. Some smart contract conditions can be assured through so-called *oracles*, linking the virtual world of the ledger to the physical world through sensors. This is the second aspect where space may come into play. Consider spatial (co-)presence as a condition (a catalyzer) of a transaction to occur, ascertained by e.g., GPS sensing. Imagine that two parties *have to* physically meet at a certain location as a condition for a transfer to occur.

It is noteworthy that up until now, nearly all of these proposed applications of smart contracts are still at the conceptual stage. As a result, supporters have been able to largely gloss over detailed discussion of spatial representation and uncertainty. A few start-up companies are currently working on proof-of-location systems, but these remain at the early stages¹. In this paper we focus on two cases in order to probe further into these issues. In

¹ Cf. <https://www.foam.space>, <https://platin.io>.



■ **Figure 1** An illustration of the scenario described in the text of Farmer A selling land to Farmer B.

the first case, we critique claims made about the use of blockchains and smart contracts to programmatically enact environmental policy and land transfers (Section 4). In the second case, we explore the idea of blockchain oracles and the role that spatial representation and uncertainty plays in how they might operate (Section 5).

3 Land transaction scenario

Let's imagine *Farmer A* who is willing to sell a piece of his paddock to *Farmer B* using a distributed ledger, sensing technology and an automated legal framework. The farmers meet to agree on the boundary of the piece of land transferred from *A* to *B* and to be merged with *B*'s current land (Fig. 1). They walk along the boundary, identify and measure the position of the new corners (metes) of their shared boundary with a GPS on their smartphones. The title to this land is then automatically transferred to *B* using a smart contract. Funds are transferred electronically from *B* to *A*, and a state land tax is automatically levied, in proportion to the areas of land transferred.

The following complications arise:

1. **Ownership problem.** The neighbor *C* of *A* and *B* questions the position of one of the metes, claiming it infringes on his land and shifts the current boundary.
2. **Traceability problem.** The areas of *A*'s and *B*'s lands do not add up after land transfer, due to measurement uncertainty and consequently the digital representation of the physical asset. The taxable land area of both (actually, all three farmers) has changed, and moreover, *A* and *B* now *digitally encroach* on the protected buffer zone around the waterway on their southern boundary.
3. **Error propagation.** The uncertain numeric representation of the new boundary triggers an automated response from the titling database, and stops the legal transfer due to the computational, automated interpretation of the legal code and regulations.
4. **Incentives problem.** In the absence of a trusted third party, it may be problematic to assure that the transfer occurred under mutual agreement, without coercion. This is particularly true for subdivisible assets (such as land), where a new identity and demarcation must be established. A chartered surveyor or similar professional currently assures this function.

5. **Digital divide** An advanced technology that relies on the promise of decentralized, ad-hoc information repositories requires an extensive investment of trust from the users. The lack of a physical artifact issued by a central authority and endowed by legitimacy may undermine this trust, in particular in societies affected by the digital divide [8] and with low digital literacy. Paradoxically, these may be the ones that would profit most from the decentralized system removed from governmental control.
6. **Governance problem.** The ability to ensure common good and protection for areas of special value must be preserved. Limiting what authorities must do may be an appealing argument in some situations, but the question is whether reform, rather than replacement, is not more desirable. The tragedy of the commons may well ensue in situations where the majority of people in a certain area have individual interests (e.g., logging) that are in conflict with a common good.

4 Is environmental cryptogovernance desirable?

The above scenario describes a common set of issues that land surveyors and public notaries help to resolve routinely. Currently, the land transfer system in most countries relies on a centralized authority, certified workforce of highly regulated surveyors, and a certain leeway in the interpretation of the digital representation of the physical asset.

We now review the consequences of a purely digital, decentralized ledger system for transferring sensitive physical assets with fiat vs bona-fide boundaries [12].

4.1 Distributed ledgers and contracts about land

A distributed ledger is a record-keeping system for transactions where a centralized registrar (authority) is not necessary, and the authority certifying the enduring nature of the transfer of ownership is assured by peers and a *consensus algorithm*.

For the transfer of a physical asset, the asset itself must be well identifiable (by a unique identifier) and distinguishable (from other assets). This may apply to transfer of diamonds or timber logs, but is more complex when it comes to land transfer, in particular when it comes to the ability to distinguish the extent of the asset. Administrative boundaries, as well as many parcel boundaries are social constructs, demarcated by agreement or authority (*fiat boundaries*). Yet, the number of legal cases and conflicts over fictitious lines demarcating property worldwide attests to the problems with the distinguishable property of these assets.

Many land assets and protected areas are bound by bona-fide boundaries, such as natural coastlines or rivers. Similarly, wetlands change in extent between the dry and wet season. These bona-fide boundaries may be highly indeterminate and a purely peer-based contract and title transfer system may result in increasing legal uncertainty with respect to land use rights and restrictions, or the inability to ensure protection of natural areas of national importance.

4.2 Crowdsourced sensing and spatial demarcation

The demarcation of boundaries, as well as the measurement of the location of the spatial context is always impacted by uncertainty [6]. This demarcation of the asset has often been left to protected professions (i.e., chartered surveyors, or notaries), thus controlling for adequate training and certifying that proper methodological approaches and equipment is used to demarcate the boundaries, and assuring the legal status of the professional as a trusted third party (further ascertaining that both parties are present and agreed about the identity and the demarcation of the asset to be transferred).

With recent improvements in consumer-grade sensing and their ubiquity (GPS sensors in smartphones), these professions have been touted obsolete despite concerns about low quality sensing [9]. Indeed, if no disagreement ensues, two parties could very well agree on their shared boundaries (Farmers *A* and *B*), and identify them by GPS coordinates. Yet, a third party may often be impacted by such decisions, if the topological correctness of the partition is to be assured (the boundary of *C* must follow those of *A* and *B*).

5 Blockchain oracles and spatial uncertainty

We now return to the second aspect of how space becomes an important consideration in smart contracts. Blockchain oracles are software or hardware services that are external to the blockchain and which are queried by smart contract code to test whether certain conditions in the real world have occurred [16]. This may include temperature sensors, rain gauges, proximity sensors, or GPS sensors. Consider a web service that provides access to real-time environmental sensor network measurements. This service may be monitored by smart contract code designed to regulate and impose fines to polluters. Another example would be an oracle that relies on sensors to verify the location of physical objects in space in order to verify the movement of goods or autonomous transport vehicles – a payment may be conditioned on certain goods reaching the client.

Issues regarding spatial uncertainty and representation remain largely unexamined in the initial discussions around blockchain oracles. Decentralized ledgers may still need to operate in an environment where certain conditions are mandated – such as both the buyer and the seller walking the boundary of the sold land parcel together, or at least meeting at the same location. How to assure that such conditions are met in a legally indisputable manner remains a concern, especially as the sensor information from consumer-grade devices could be easily questioned in legal proceedings. The need for blockchain oracles therefore will lead to a number of difficult research challenges to which the GIScience community – with its foundational work on spatial uncertainty and representation as well as environmental sensor networks and spatial change detection – can readily contribute [15].

6 Conclusion

We have chosen a land transfer scenario to illustrate that blockchain technology itself is not a *panacea* for problems of environmental governance, and will lead to unanticipated collateral effects. While distributed ledgers and smart contracts may create new possibilities for the management of digital assets, their applicability is also limited by aspects of environmental governance that deal with concepts that can not be simply tokenized and reduced to unambiguous digital representations. In addition to the spatial representation of land assets, biodiversity and non-point source pollution are two complex areas of environmental regulation which can not be interpreted through a simple automated set of rules with binary outcomes [14].

We therefore urge for strong caution and do not believe that “*time is ripe for ‘cryptogovernance’*,” [3] at least in the foreseeable future. In particular when it comes to the relationship of people with their living environment, their land ownership and the conservation of common-good natural assets, strong institutional frameworks, legal certainty and awareness of the fluid relationship between land and people are necessary. Research to develop a stronger understanding of the relationship between spatial representation and the workings of distributed ledger technologies is warranted, and a necessary prerequisite (among others) to any widespread adoption of environmental cryptogovernance.

References

- 1 Aanchal Anand, Matthew McKibbin, and Frank Pichel. Colored coins: Bitcoin, blockchain, and land administration. In *Annual World Bank Conference on Land and Poverty*, 2016. URL: <http://cadasta.org/resources/white-papers/bitcoin-blockchain-land/>.
- 2 Vitalik Buterin. A next-generation smart contract and decentralized application platform. 2014. URL: https://www.weusecoins.com/assets/pdf/library/Ethereum_white_paper-a_next_generation_smart_contract_and_decentralized_application_platform-vitalik-buterin.pdf.
- 3 Guillaume Chapron. The environment needs cryptogovernance. *Nature*, 545(7655):403, 2017.
- 4 Konstantinos Christidis and Michael Devetsikiotis. Blockchains and smart contracts for the internet of things. *IEEE Access*, 4:2292–2303, 2016.
- 5 Boris Döder and Omri Ross. Timber tracking: Reducing complexity of due diligence by using blockchain technology. *SSRN*, 2017. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3015219.
- 6 Peter F Fisher. Models of uncertainty in spatial data. In P. A. Longley, M. Goodchild, D. J. Maguire, and D. W. Rhind, editors, *Geographical Information Systems*, volume 1, book section 13, pages 191–205. Longman, Essex, UK, 2nd edition, 1999.
- 7 Arthur Gervais, Ghassan O Karame, Karl Wüst, Vasileios Glykantzis, Hubert Ritzdorf, and Srdjan Capkun. On the security and performance of proof of work blockchains. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 3–16. ACM, 2016.
- 8 Mark Graham, Scott Hale, and Monica Stephens. Featured graphic: Digital divide: the geography of internet access. *Environment and Planning A*, 44(5):1009–1010, 2012.
- 9 Alastair Lewis and Peter Edwards. Validate personal air-pollution sensors: Alastair lewis and peter edwards call on researchers to test the accuracy of low-cost monitoring devices before regulators are flooded with questionable air-quality data. *Nature*, 535(7610):29–32, 2016.
- 10 Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. 2008. URL: <https://bitcoin.org/bitcoin.pdf>.
- 11 Pradip Kumar Sharma, Seo Yeon Moon, and Jong Hyuk Park. Block-vn: A distributed blockchain based vehicular network architecture in smart city. *Journal of Information Processing Systems*, 13(1):84, 2017.
- 12 Barry Smith and Achille C Varzi. Fiat and bona fide boundaries. *Philosophical and phenomenological research*, pages 401–420, 2000.
- 13 Nick Szabo. Formalizing and securing relationships on public networks. *First Monday*, 2(9), 1997.
- 14 Susan Walker, Ann L Brower, RT Stephens, and William G Lee. Why bartering biodiversity fails. *Conservation Letters*, 2(4):149–157, 2009.
- 15 Mike Worboys and Matt Duckham. Monitoring qualitative spatiotemporal change for geosensor networks. *International Journal of Geographical Information Science*, 20(10):1087–1108, 2006.
- 16 Xiwei Xu, Cesare Pautasso, Liming Zhu, Vincent Gramoli, Alexander Ponomarev, An Binh Tran, and Shiping Chen. The blockchain as a software connector. In *Software Architecture (WICSA), 2016 13th Working IEEE/IFIP Conference on*, pages 182–191. IEEE, 2016.
- 17 Yong Yuan and Fei-Yue Wang. Towards blockchain-based intelligent transportation systems. In *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*, pages 2663–2668. IEEE, 2016.

Towards Optimal Deployment of a Sensor Network in a 3D Indoor Environment for the Mobility of People with Disabilities

Ali Afghantoloe

Center for Research in Geomatics, Laval University, Quebec City, Canada,
Ali.afghantoloee.1@ulaval.ca

Mir Abolfazl Mostafavi

Center for Research in Geomatics, Laval University, Quebec City, Canada
Center for Interdisciplinary Research in Rehabilitation and Social Integration, Laval University,
Quebec City, Canada
Mir-Abolfazl.Mostafavi@scg.ulaval.ca

Abstract

Mobility of people with disabilities is one of the most important challenges for their social integration. There have been significant effort to develop assistive technologies to guide the PWD during their mobility in recent years. However, these technologies have limitations when it comes to the navigation and guidance of these people through accessible routes. This is specifically problematic in indoor environments where detection, location and tracking of people, and other dynamic objects that may limit the mobility of these people, are very challenging. Thus, many researches have leveraged the use of sensors to track users and dynamic objects in indoor environments. However, in most of the described methods, the sensors are manually deployed. Due to the complexity of indoor environments, the diversity of sensors and their sensing models, as well as the diversity of the profiles of people with disabilities and their needs during their mobility, the optimal deployment of a sensor network is a challenging task. There exist several optimization methods to maximize coverage and minimize the number of sensors while maintaining the minimum connectivity between the sensor nodes in a network. Most of the current sensor network optimization methods oversimplify the environment and do not consider the complexity of 3D indoor environments. In this paper, we propose a novel 3D local optimization algorithm based on a geometric spatial data structure that takes into account some of these complexities for the purpose of helping PWD in their mobility in 3D indoor environments such as shopping centers, museums and other public buildings.

2012 ACM Subject Classification Hardware → Sensors and actuators

Keywords and phrases 3D indoor navigation, Sensor network deployment, People with disabilities

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.19

Category Short Paper

1 Introduction

Social participation of people with disabilities (PWD) is one of the challenging problems in our society. According the United Nation's convention for PWD "persons with disabilities may include those who have long-term physical, mental, intellectual or sensory impairments which in interaction with various barriers may hinder their full and effective participation in



© Ali Afghantoloe and Mir Abolfazl Mostafavi;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 19; pp. 19:1–19:6

Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

society on an equal basis with others” [6]. According to a recent publication by Statistics Canada (2013), 13.7% of the population aged over 15 years live with a type of disability.

Based on the International Classification of Functioning (ICF) and the Disability Creation Process (DCP) model [5], social participation of PWD results from the interactions between their personal characteristics and the physical and social environmental factors. Most of the urban infrastructures and services are designed for people without any disability and do not consider the specific needs of PWD. This significantly limits the mobility of PWD and their social participation (e.g., going to work, the market, the museum, etc.). Mobility is a life habit that significantly influences other human life habits [5], and depending on the context, mobility may include movements such as postural transfers (e.g., from a chair to a bed) or moving from a point to another during diverse daily activities (walking, working or playing, driving a car, and using public transportation).

With the expansion of urban development and the construction of complex city infrastructures such as road networks, public buildings, shopping malls, airports, and museums, there is an increasing need for assistive navigation technologies to help PWD in their mobility. Efficient navigation in such environments require accurate and up-to-date information on the accessibility of those environments including information on possible obstacles and facilitators for the mobility of PWD. For this purpose, sensor networks provide interesting potentials to locate and track the dynamics of indoor environments and provide timely information to PWD during their navigation.

In recent years, a variety of sensor types has been developed and used for monitoring and measuring dynamic environments. For instance, in a mobility context, the majority of sensors have been used for positioning and tracking of people and moving objects. Tracking sensors are generally embedded in the environment and constitute a sensor network. These sensors must be deployed in the environment and have the best configuration to maximize the coverage and guarantee their connectivity and minimize the cost (optimal number of sensors and their types). There exist several optimization methods to maximize coverage and minimize the number of sensors while maintain the minimum connectivity between the sensor nodes in a network. Most of the current sensor network optimization methods oversimplify the environment and do not consider the complexity of 3D indoor environments. In this paper, we propose a novel 3D local optimization algorithm based on a geometric spatial data structure that takes into account some of these complexities for the purpose of helping PWD in their mobility in 3D indoor environments such as shopping centers, museums and other public buildings.

The remainder of this paper is organized as follows: Section 2 presents a brief literature review on sensor network deployment in indoor environments for mobility purposes and highlights their strengths and limitations. In section 3, the methodology of the proposed local deployment approach will be elaborated with consideration of indoor complex environment models and mobility applications. Then in section 4, an experiment will be conducted in an indoor environment. Finally, the results will be discussed in the last section.

2 Related works

Optimal deployment of a sensor network in a complex indoor environment is a challenging task. This complexity becomes even more challenging if we consider the diversity of sensor types and their sensing models as well as the specificity of the requirements for each application. With network deployment optimization methods, we try to maximize the coverage of the network and minimize the cost of the network and energy consumption for each node while maintaining a minimum connectivity between nodes in a wireless sensor network (WSN) [2].

The WSN coverage problem has been studied intensively in the last decade. A sensor coverage can be either target-based or area-based. In some WSN applications, detecting target points such as buildings, doors, flags and boxes are desired, while in area-based coverage, the aim is to detect mobile targets such as intruders in a given area [7]. Covering target points, instead of the whole area, is addressed in the target-based coverage problem, whose purpose is to cover the maximum number of target points. In the area-based coverage problem, which is used in this research, the objective is to obtain the maximum region covered by sensors, which is usually evaluated as the ratio of the covered area to the whole area [8].

Several methods have been proposed for optimal deployment of sensor networks based on the maximum coverage criteria [3]. These methods are either global or local and can be deterministic or stochastic. Particle Swarm Optimization (PSO) algorithms [9], and Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [2] are among global approaches for sensor network deployment optimization. These methods apply a global objective function that is optimized for the whole network. In local algorithms such as Virtual force-based methods [11], and Voronoi algorithms [10], the optimization is done locally by changing the position of sensors with respect to the local context and the configuration of the neighboring algorithms. Both global and local algorithms can be considered as stochastic or deterministic depending on the definition of the sensing model of the sensors.

Most of the sensor network optimization methods use 2D raster representations of the environment [2] or voxel representation for 3D environments [4], which limit their precision and efficiency. This is because raster and voxel representations need a regular partition of the whole space even for homogeneous areas (i.e. the unoccupied pixels or voxels). Moreover, the raster-based models cannot be used to represent precisely indoor environments as they are constrained by their resolutions.

Voronoi based algorithms have attracted much attention in the research community interested in optimal sensor networks deployment, specially for its interesting spatial and topological properties for defining and managing sensor networks. For instance, [3] have proposed a local context-aware sensor network deployment algorithm based on 2D Voronoi diagrams for urban environments. In the latter work, Voronoi diagram is also used to define a movement strategy for sensors to heal the coverage holes of a sensor network where the environment was represented using a 2.5D digital surface model (DSM). In [1], a sensor coverage estimation method has been proposed based on precise 3D vector representation of the environment. Here in this paper, we propose to take advantage of 3D Voronoi diagrams and the vector-based representation of the indoor environment to develop a local sensor network optimization algorithm for indoor environments in order to support the navigation of PWD.

2.1 Methodology

For the deployment of a sensor network in an indoor environment, we propose a local context aware optimization algorithm based on 3D Voronoi diagrams. For this purpose, we assume that sensors can be deployed mainly on the walls and ceilings. Building floors are considered as target areas to be covered where navigation activities are expected. As mentioned previously, the sensing model (binary or probabilistic), sensor orientation (omni-directional or directional) and other sensor characteristics such as observation angle and distance ranges, should also be defined. In this paper, we consider an omni-directional sensor model for our sensor network.

In addition to sensor characteristics, the 3D indoor environment needs to be represented in details for optimal sensor network deployment. We also need to consider the presence

of other objects embedded in the indoor environment that may affect coverage information (e.g., presence of a column or other permanent obstacles in the environment). Hence, we need a data structure that supports precise representation of the indoor environment and allows semantic specification of all its components. For modeling 3D indoor environments, we consider to benefit from the potentials of 3D IndoorGML for the representation of such environments.

3D IndoorGML is an extension of CityGML (Level of details (LoD) 4) that provides semantical, topological, and spatial information of objects and services. Like CityGML LoD4, IndoorGML is an open standardized data model of interior space of 3D buildings that includes core modules, appearance modules, and thematic modules. The main structure of IndoorGML divides the indoor space into multi-spaces called cells, and the intersected area of two neighboring cells is called boundary surface. IndoorGML uses two related spaces to model indoor environments: (1) primal space is the geometrical representation of cells and boundary surfaces, (2) dual space is the Node Relationship Graph representation of cells and boundary surfaces, which respectively corresponds to nodes and edges. Generally, IndoorGML contains connection spaces (e.g., doors), anchor spaces (e.g., building exits), general spaces (e.g., rooms) and transition spaces (e.g., passages). In contrast, CityGML includes boundary surfaces, rooms, openings, and closure surfaces (e.g., the space between the kitchen and the living room is a virtual surface called closure surface).

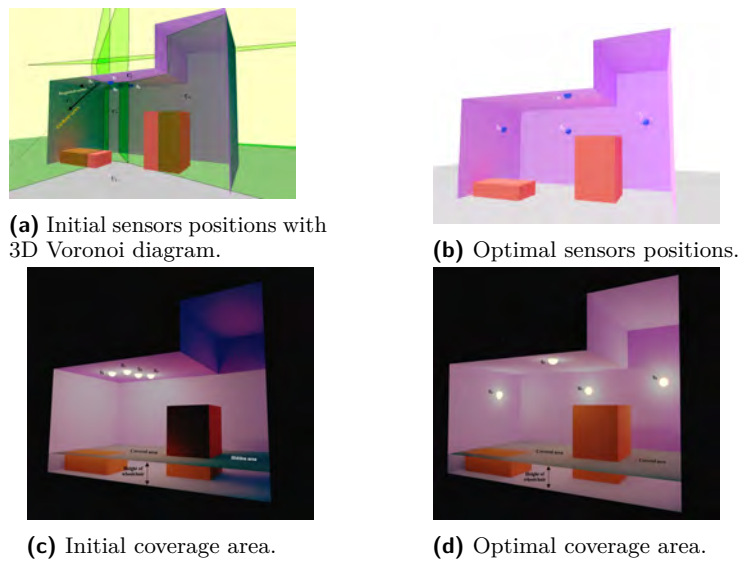
Algorithm 1: 3D Voronoi deployment algorithm.

input : n omni-directional cameras $S_i(x_i, y_i, z_i)$
output : (X_i, Y_i, Z_i) optimal solutions
objective: Maximizing the coverage of cameras network

Initialize: Random distribution of the cameras on deployment planes (walls/ceilings) Compute initial sensor network coverage ;

while *stop_criterion* **do**
 3D_Voronoi(S_1, \dots, S_n);
 for $i \leftarrow 1$ **to** n **do**
 Movement strategy(S_i);
 {
 1- choose the farthest vertex in the same direction of path segments;
 2- project the movement vector on sensor deployed plane;
 3- if movement vector has intersection with obstacle, keep a given distance between sensors and obstacle;
 }
 Update sensor network coverage (S_i);
 {
 1- choose the movement amount based on the coverage improvement
 }
 end
end

The objective of sensors deployment in such environment is the maximization of the covered areas of path segments that include floors with the height of a typical pedestrian who navigates in the indoor environment. Our aim with placing sensors in such environment is to inform the PWD of the dynamics of those environments and also to guide them safely towards their final destination.



■ **Figure 1** Deployment of 4 cameras in an indoor environment with obstacles.

The proposed algorithm for deployment of sensors (the cameras in this research) is inspired from a local 2D Voronoi approach presented in [10]. This method uses a Voronoi diagram for the representation of a sensor network and the relations between sensors. We extend that method to 3D space and use a 3D Voronoi diagram for the representation of sensors and their topological relations in the sensor network. Thus, in the proposed algorithm, we first create the 3D Voronoi structure using sensors as the generators of the 3D cells in algorithm 1. In each iteration, we move the sensors towards the farthest vertex of their Voronoi cell to reduce the overlapping coverages and to better cover the target areas. It should be noted that the motion of each sensor needs to be done on the wall or ceiling. Therefore, the motion vector of each sensor is projected on the sensor position plane and the sensor is moved in this direction towards its new position. In the case of the presence of a permanent obstacle in the moving direction we need to keep the sensor away from the obstacle with a given distance so that its sensing field is maximized.

3 Experiment and results

In this experiment, we assume that four cameras are deployed in a semi-complex indoor environment where the ceiling is composed of two sections with different heights. We assume that the cameras have a spherical sensing model with a defined range of view. Each camera has an initial position (x, y, z) and is located on the ceiling or walls.

The environment model represents a semi-complex three-dimensional indoor environment and contains a few static obstacles (Figure 1a). The 3D indoor model is stored using IndoorGML and the geometric information can be easily extracted and analyzed if needed. In our case study, the indoor environment model consists of 8 segments (faces) and includes two obstacles. The goal of this experiment is to reach the maximum coverage of the floor that can be used as a part of path for the mobility of a PWD (e.g., a person using a wheelchair) from an initial configuration of cameras (Figure 1c). Then, the objective function is defined in a way that the floor is covered with a height corresponding to the height of a person using a wheelchair for her/his mobility (Figures 1b and 1d).

4 Conclusions

Navigation of PWD is a complex task in indoor environments. These people need assistive technologies to help them in their mobility and to guide them through their path by providing them directions and information on the accessibility of their path. Wireless sensor networks provide interesting opportunities to help these people with their navigation in indoor environments. However, optimal deployment of a sensor network in a 3D indoor environment is a very challenging problem given the complexity of the indoor environments and the presence of diverse obstacles as well as the diversity of sensors and their sensing models. Here in this paper, we have presented a new local optimization algorithm integrating 3D Voronoi diagrams for sensor network representation and 3D IndoorGML for the representation of the 3D indoor environments. We have defined an iterative algorithm for sensors movement that allows the improvement of the overall coverage of the sensor network. Finally we have presented a concept proving experiment with promising results. This work is part of an ongoing research project. We plan to carry out more comprehensive experiments in the near future to test and improve the proposed algorithm.

References

- 1 A Afghantoloe, S Doodman, F Karimipour, and M A Mostafavi. Coverage Estimation of Geosensor in 3d Vector Environments. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40(2):1, 2014.
- 2 Vahab Akbarzadeh, Christian Gagné, Marc Parizeau, Meysam Argany, and Mir Abolfazl Mostafavi. Probabilistic sensing model for sensor placement optimization based on line-of-sight coverage. *IEEE Transactions on Instrumentation and Measurement*, 62(2):293–303, 2013.
- 3 Meysam Argany. *Development of a GIS-Based Method for Sensor Network Deployment and Coverage Optimization*. PhD thesis, Université Laval, 2015.
- 4 Francois-Michel De Rainville, Christian Gagné, and Denis Laurendeau. Automatic Sensor Placement For Complex Three-dimensional Inspection and Exploration.
- 5 Patrick Fougeyrollas, René Cloutier, Hélène Bergeron, Ginette St-Michel, Jacques Côté, Marcel Côté, Normand Boucher, Kathia Roy, and Marie-Blanche Rémillard. *The Quebec classification: Disability creation process*. Québec RIPPH/SCCIDIH., 1998.
- 6 Melvyn Colin Freeman, Kavitha Kolappa, Jose Miguel Caldas de Almeida, Arthur Kleinman, Nino Makhashvili, Sifiso Phakathi, Benedetto Saraceno, and Graham Thornicroft. *Convention on the Rights of Persons with Disabilities*, 2015.
- 7 M Amac Guvensan and A Gokhan Yavuz. On coverage issues in directional sensor networks: A survey. *Ad Hoc Networks*, 9(7):1238–1255, 2011.
- 8 Chi-Fu Huang and Yu-Chee Tseng. The coverage problem in a wireless sensor network. *Mobile Networks and Applications*, 10(4):519–528, 2005.
- 9 Raghavendra V Kulkarni and Ganesh Kumar Venayagamoorthy. Particle swarm optimization in wireless-sensor networks: A brief survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 41(2):262–267, 2011.
- 10 Guiling Wang, Guohong Cao, and Thomas F La Porta. Movement-assisted sensor deployment. *IEEE Transactions on Mobile Computing*, 5(6):640–652, 2006.
- 11 Yao Zou and Krishnendu Chakrabarty. Sensor deployment and target localization based on virtual forces. In *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications*. IEEE Societies, volume 2, pages 1293–1303. IEEE, 2003.

Challenges in Creating an Annotated Set of Geospatial Natural Language Descriptions

Niloofer Aflaki

Massey University, Auckland, New Zealand
n.aflaki@massey.ac.nz

Shaun Russell

Massey University, Auckland, New Zealand
shaun@ensemblemusic.co.nz

Kristin Stock

Massey University, Auckland, New Zealand
k.stock@massey.ac.nz

Abstract

In order to extract and map location information from natural language descriptions, a first step is to identify different language elements within the descriptions. In this paper, we describe a method and discuss the challenges faced in creating an annotated set of geospatial natural language descriptions using manual tagging, with the purpose of supporting validation and machine learning approaches to annotation and text interpretation.

2012 ACM Subject Classification Applied computing → Annotation

Keywords and phrases Annotation challenges, spatial relations, spatial language

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.20

Category Short Paper

Funding This work is partly funded through a Ordnance Survey PhD scholarship.

1 Introduction and Background

To progress research on the interpretation of geospatial natural language, methods for automated tagging of spatial language are required [5, 9]. In this paper, we discuss the challenges that we encountered when trying to create manually tagged annotated data set that addresses the shortcomings of previous data sets, using two experiments. A number of researchers have addressed the problem of annotating geospatial natural language. For example, Stock and Yousaf [10] annotated a wide range of language elements, including adverb and parts of objects as well as relatum, locatum and spatial relation, mainly by extending POS tags in a rule-based approach. Kordjamshidi et al [5] restrict their attention to trajector, landmark and spatial prepositions, although they acknowledge that other parts of speech can be used to express spatial relations. GUM Space specifies a broad range of tags including locatum, relatum, spatial modality [3]. SpatialML uses mark-up language to tag elements [7] including places, coordinate, orientations, form of reference, direction, distance and frame. Work by Zwarts [12] and Kracht [6] address spatial prepositions, with a focus on directional prepositions and location. Much of the previous work is either limited to very simple elements [5]; adopts a complex tag structure [3] or assumes a particular syntactic (grammatical) structure [5, 10]. We propose an annotation scheme that addresses these limitations in that it focuses on semantics rather than syntax.



© Niloofer Aflaki, Shaun Russell, and Kristin Stock;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 20; pp. 20:1–20:6

Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

2 Methodology

We conduct our exploration of the challenges of creating an annotated data set using two experiments. The first one compares the tagging conducted by pairs of human annotators and discusses discrepancies and issues involved in manual tagging. The second one discusses variations between individual human respondents in matching natural language descriptions to spatial relations, highlighting the lack of consensus.

2.1 Experiment 1: Creating an Annotated Data Set

The selection of an annotation scheme was based on three criteria: 1. What must be individually identified in order to support effective geocoding of the text? This is difficult to evaluate conclusively, as it depends upon the geocoding approach, and some aspects of spatial language are still not well understood. This criterion influences not only which items we tag, but also which items we identify as separate elements. For example, it is not useful to separate *next to* into two separate tags, because the meaning depends on the combination of the words, and the meaning of *to* in particular is dependent on the presence of *next*. In contrast, adverbs like *right*, or *directly*, have their own meanings which are similar regardless of the preposition they appears with, although the meaning may be influenced by the latter. 2. Can some of the tags or their subcategories be reliably determined automatically? If a particular semantic tag can be reliably identified through an automated approach, then there is little point in annotating in manually. The reliability of an automated approach is a question of degree, but we use the yardstick that if the set of words of interest can be defined by a clear set of specific words, none of which are homonyms, then they might reliably be identified automatically. In practice this is rare, because for example, even though the set of prepositions is a closed word class, since we are interested in semantic tags rather than syntactic, and prepositions normally encode spatial relations, there are examples of spatial relations that are not prepositions (*e.g. in line with*). 3. What is practical to expect people to reliably annotate? This involves both volume and simplicity. A set of tags that is too complex will be difficult for manual annotators to deal with. The set of tags must be manageable in quantity, and simple enough to understand without specialist knowledge.

In Experiment 1, we develop a generic spatial annotation framework based on the semantic roles of tokens in a sentence. To this end, 1000 sentences were randomly selected from the combined set of three data sources: The Nottingham Corpus of Spatial Language[9], The Landcare Research National Soils Database ¹ and The Where Am I survey, in which natural language descriptions were elicited from human respondents, as described in [8]. Table 1 identifies, describes and explains the annotation scheme that was used. Four annotators were given an expanded version of Table 1 with a simple explanation of terms and examples. The purpose of the work was explained to them in simple terms, and they were given access to the tagging app. Each annotator was then asked to annotate 10 expressions using the tagging app, after which the authors examined the expressions and gave feedback on any issues, before the annotator began annotating in earnest. Each expression was tagged twice by two different annotators.

¹ <https://soils.landcareresearch.co.nz/index.php/soil-data/national-soils-data-repository-and-the-national-soils-database/>

■ **Table 1** Tag labels and descriptions.

Title	Explanation
Trajector	The object whose location is being described. The important role of the trajector in spatial language has been discussed by a number of researchers and is also known as locatum [3] or figure [11].
Landmark	The object that is used as a reference point in the description. The landmark also plays an important and well documented role in spatial language, and is similar to the relatum and ground identified by other researchers[11].
Spatial Relation	The word or words that indicate how two objects are positioned relative to other. The importance of spatial relations has also been well recognised, and they have been widely researched [1, 4, 12]. In syntactic terms, spatial relations are most often represented using prepositions, but not always.
Location and movement verb (lmv)	A verb that describes the manner in which one object is positioned relative to the other. The location and movement verb is a subset of the verb syntactic category[11]. <i>The road crosses behind the church.</i>
Spatial qualifier	A word or set of words that adds more information to the spatial relation and or the location and movement verb. Spatial qualifiers have not been widely recognized as an important carrier of spatial information as yet, and may be represented with a range of different parts of speech, including adverbs, adjectives and nouns. <i>The road goes right beside the church</i>
Spatial specifier	A word or set of words that describes particular subparts of a feature.E.g. <i>The north of the country.</i> Spatial specifiers have also not been widely studied in specific terms, with work instead focusing on general issues of mereology [2].

2.2 Experiment 2: Matching of Expressions to Spatial Relations

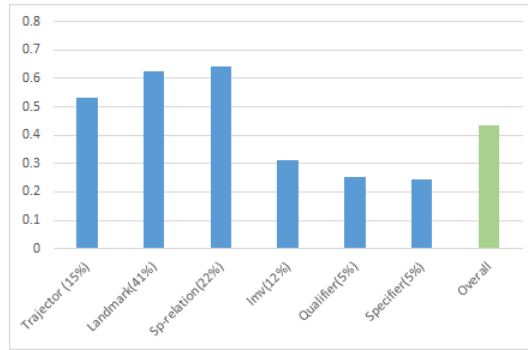
In the second experiment, we used data collected in earlier work [10]. In this work, respondents were shown expressions one at a time, and asked to match each expression to one of a series of diagrams that illustrated spatial relations. After viewing the expression and the set of available spatial relation diagrams, each annotator was asked to select values on a Likert scale that included only the positive side of the scale, to indicate his or her opinion about how closely each of the selected spatial relation diagrams matched the expression: *Strongly Agree, Agree, Agree Somewhat*. Only the positive half of the scale was used because users were invited to only select diagrams that they thought reflected the expressions (i.e. if they did not agree, the respective diagram would not be selected). Weights were allocated to each response for a given spatial relation diagram-expression pair, using 1, 0.75 and 0.5 for Strongly agree, Agree and Agree Somewhat respectively. The score for each expression and its geometric configuration was calculated using this formula:

$$GCOScore_{expression, diagram} = \sum_{k=0}^n (response_k weight_k) / n \quad (1)$$

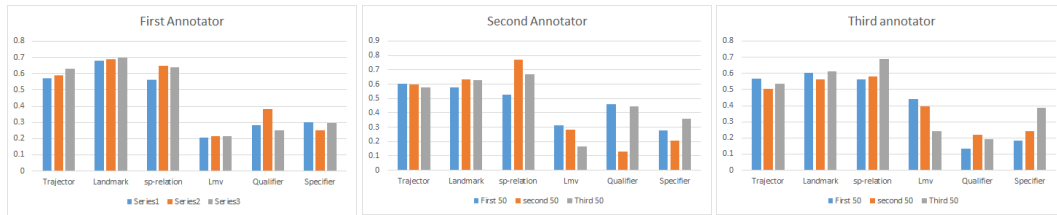
In which *response k* represents the number of responses with *weight k*, and *n* defines the total number of responses for expression k. Full details of the methodology can be found in[10].

3 Results

In order to evaluate the reliability of the manual annotation process in Experiment 1, we calculate inter-annotator agreement among the four annotators. Since expressions were randomly allocated to annotator, any combination of pairs of specific annotators may annotate a given expression. Inter-annotator agreement was calculated by comparing the words in



■ **Figure 1** Study 1. Mean inter-annotator agreement by tag type.



■ **Figure 2 a-c** Annotator performance through the time.

a given expression that were given a particular tag by each annotator. Since many of the expressions were complex and contained more than one of some tags, we calculate agreement by proportion of overlap between the words annotated with a particular tag by each user, rather than by a simple true/false agreement. Formula 2 expresses this measurement of agreement between annotators for a single expression: For a given tag, ME_k denotes the number of mutual elements (words or multi-word tagged values) that both annotators agree on, and max_k denotes the maximum number of elements that are tagged by either annotator. The total agreement score for the expression is then average of agreement across the populated tags. For example, if user 1 specifies Australia, New Zealand and Canada as landmarks and user 2 specifies Canada and USA as landmarks ME_k for the landmarks would be 1, because just Canada is mutual and the max_k would be three as the maximum number of landmarks by either annotator. The agreement score is calculated for all the tags in an expression, and the average is calculated to determine the agreement across the entire expression.

$$AgreementScore = Average(\sum(ME_k/max_k)) \quad (2)$$

Figure 1 shows the mean inter-annotator agreement for individual tags, as well as overall and also the percentage of tags of each type that were annotated in the 1000 expressions. We used this formula, to have an accurate calculation of each separate tag. We also explore the role of annotator experience in the manual tagging process, and evaluate whether annotator performance improves over time. For each annotator, we calculated inter-annotator agreement for the first, second and third 50 expressions tagged by three annotators through the time to see whether their performance changed by time or not. Only 3 annotators are shown because the remaining did not annotate sufficient expressions. Figures 2a to c show the results. We then calculated the inter-annotator agreement of different subsets of annotators, to determine whether some annotators were more successful than others in tagging, either overall or for specific tags. The results (Figure 3) show some inconsistency. It is, however, clear that Annotator 2's contribution is important, with her exclusion resulting in overall deterioration.

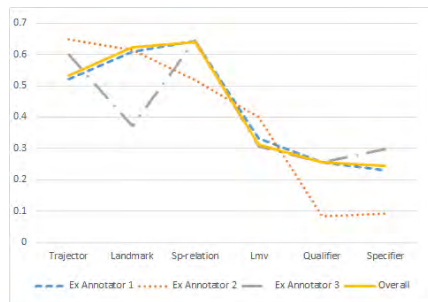


Figure 3 Inter-annotator agreement excluding each annotator in turn.

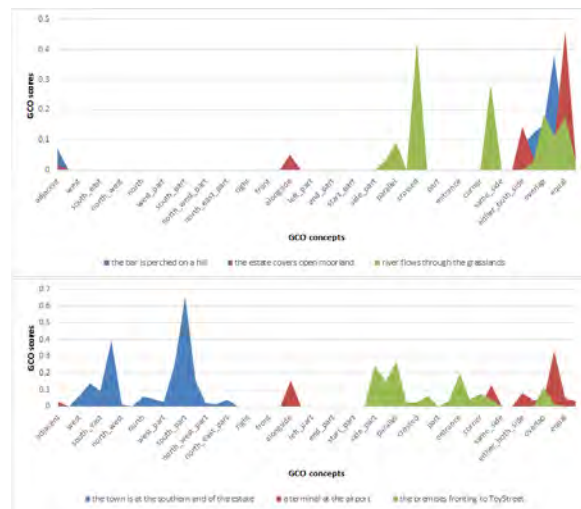


Figure 4 a,b Study 2. GCO score for second three expression.

Turning to Experiment 2, the results highlight the lack of agreement among individual respondents regarding the spatial relation diagram that best reflects a given expression. The respondents in Experiment 2 were also non experts in geographic information science. Figures 4 a and b each show the spread of responses for three example expressions. In contrast to Experiment 1, Experiment 2 used short, simple spatial expressions, and the graph shows the frequency (after weights have been applied as described in Section 2) of selection of each spatial relation for a given expression. Two expressions in 4b show a number of small peaks, with no clearly dominant relation selected by the respondents. Across the entire data set, a similar pattern was observed, with lack of consensus among respondents in selecting spatial relations to match many expressions.

4 Discussion and Conclusion

The results clearly show that it is not straightforward to create a manually annotated data set of natural language descriptions with a broad set of language elements that is based on semantics rather than syntax. Obviously, for an annotated data set for use in machine learning and validation, we would like the agreement to be very strong. Considerations of the level of experience of the annotators and the examination of the influence of specific annotators on particular tags did not result in noticeable improvement. The challenges that were encountered can be summarised as follows: Firstly, it is not unusual for the same place name, geographic feature or moving object to be both a trajector and a landmark, and secondly, the landmark/trajector status of a word may be ambiguous. The following example illustrates both of these cases. In the expression *the church stands beside the post office near the bridge*, the structure of the expression could be:

“trajector+(lmv)+spatial-relation+landmark+spatial relation+landmark”

“trajector+(lmv)+spatial-relation+(trajector and landmark)+spatial relation+landmark”

In the first case, church is a trajector for both the church landmark and the bridge landmark, and in the second case post office is the trajector for the bridge landmark, as well as the landmark for the church trajector. The annotation scheme used in this paper allowed each word to be tagged only as a trajector or a landmark, but not both. The creation of a

tag that indicates a dual role may be a possible methods for addressing this. Resolution of ambiguity is a more difficult problem to solve, and even the most expert and experienced annotators may disagree. A final observation from the results is that, spatial qualifiers and spatial specifiers had only fair inter-annotator agreement (lower than other tags), and while this may be in part due to confusion about when to use each, when questioned, Annotator 2 was able to accurately explain when the spatial specifier tag was used and claimed to find it easy to understand. Confusion in the tagging process was sometimes caused by considerations of grammar, rather than meaning.

In this paper, we have described a semantic annotation scheme that is designed to be both useful and practical, and the methodology used to create an annotated data set. We analysed and presented some of the challenges encountered in the process, and the fundamental difficulties resulting from ambiguity and individual discrepancies in the use of spatial language that make it difficult to define a single, reliable annotated data set at a semantic level. In future work, we intend to do more analysis and test different annotation strategies like single tag per annotator, to see if there is any improvement in the results achieved.

References

- 1 Kenny R Coventry and Simon C Garrod. *Saying, seeing and acting: The psychological semantics of spatial prepositions*. Psychology Press, 2004.
- 2 Torsten Hahmann and Michael Gruninger. A naive theory of dimension for qualitative spatial relations. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, 2011.
- 3 R Ross J Bateman J Hois, T Tenbrink, R Ross, and J Bateman. Gum-space. Technical report, Technical report, Universität Bremen SFB/TR8 Spatial Cognition, 2009.
- 4 John D Kelleher and Fintan J Costello. Applying computational models of spatial prepositions to visually situated dialog. *Computational Linguistics*, 35(2):271–306, 2009.
- 5 Parisa Kordjamshidi, Martijn Van Otterlo, and Marie-Francine Moens. Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Transactions on Speech and Language Processing (TSLP)*, 8(3):4, 2011.
- 6 Marcus Kracht. The fine structure of spatial expressions. *Syntax and semantics of spatial P*, pages 35–62, 2008.
- 7 Inderjeet Mani, Christy Doran, Dave Harris, Janet Hitzeman, Rob Quimby, Justin Richer, Ben Wellner, Scott Mardis, and Seamus Clancy. Spatialml: annotation scheme, resources, and evaluation. *Language Resources and Evaluation*, 44(3):263–280, 2010.
- 8 Kristin Stock, Didier Leibovici, Luciene Delazari, and Roberto Santos. Discovering order in chaos: using a heuristic ontology to derive spatio-temporal sequences for cadastral data. *Spatial Cognition & Computation*, 15(2):115–141, 2015.
- 9 Kristin Stock, Robert C Pasley, Zoe Gardner, Paul Brindley, Jeremy Morley, and Claudia Cialone. Creating a corpus of geospatial natural language. In *International Conference on Spatial Information Theory*, pages 279–298. Springer, 2013.
- 10 Kristin Stock and Javid Yousaf. Context-aware automated interpretation of elaborate natural language descriptions of location through learning from empirical data. *International Journal of Geographical Information Science*, pages 1–30, 2018.
- 11 Leonard Talmy. *Toward a cognitive semantics*, volume 2. MIT press, 2000.
- 12 Joost Zwarts. Prepositional aspect and the algebra of paths. *Linguistics and Philosophy*, 28(6):739–779, 2005.

Improved and More Complete Conceptual Model for the Revision of IndoorGML

Abdullah Alattas

Faculty of Environmental Design, Geomatics department, King Abdulaziz University, Jeddah, Saudi Arabia

Arch.alattas@gmail.com

Faculty of Architecture and the Built Environment, Delft University of Technology, Julianalaan 134, 2628 BL Delft, The Netherlands

a.f.m.alattas@tudelft.nl

Sisi Zlatanova

Faculty of Built Environment, University of New South Wales, NSW 2052, Sydney , Australia
s.zlatanova@unsw.edu.au

Peter van Oosterom

Faculty of Architecture and the Built Environment, Delft University of Technology, Julianalaan 134, 2628 BL Delft, The Netherlands

P.J.M.vanOosterom@tudelft.nl

Ki-Joune Li

Pusan National University, Kumjeong-Gu, Pusan 46241, Korea
lik@pnu.edu

Abstract

With the increasing number of indoor navigation applications, it is essential to have clear and complete conceptual model (in the form of UML class diagram) for IndoorGML. The current version of IndoorGML standard has an incomplete class diagram (incomplete w.r.t. attributes, of which some are appearing in the XML/GML schema), and that provides confusion for the users of the standard. Furthermore, there are some issues related to unclear association names, unclear class names, classes that related to the Primal space and the Dual space, code lists not specific per type (which should have their own code list values), untyped relationships to external object classes, and semantically overlapping classes. In this paper, we propose an enhancement for IndoorGML conceptual model (UML class diagram) to avoid the misunderstanding. We propose a conceptual model that maps the classes of the standard in a better way. This conceptual model is the basis for 1) a database schema when storing IndoorGML data, 2) the XML schema when exchanging IndoorGML data, and 3) when developing IndoorGML applications with an intuitive and clear GUI. Furthermore, the proposed conceptual model provides constraints for more meaningful model and to define more sharply what is considered valid data. This paper briefly reports these preliminary results on the UML conceptual model.

2012 ACM Subject Classification Information systems → Geographic information systems, Software and its engineering → Unified Modeling Language (UML), Information systems → Location based services

Keywords and phrases Navigation, Space, Boundary, CellSpace

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.21

Category Short Paper



© Abdullah Alattas, Sisi Zlatanova, Peter van Oosterom, and Ki-Joune Li; licensed under Creative Commons License CC-BY

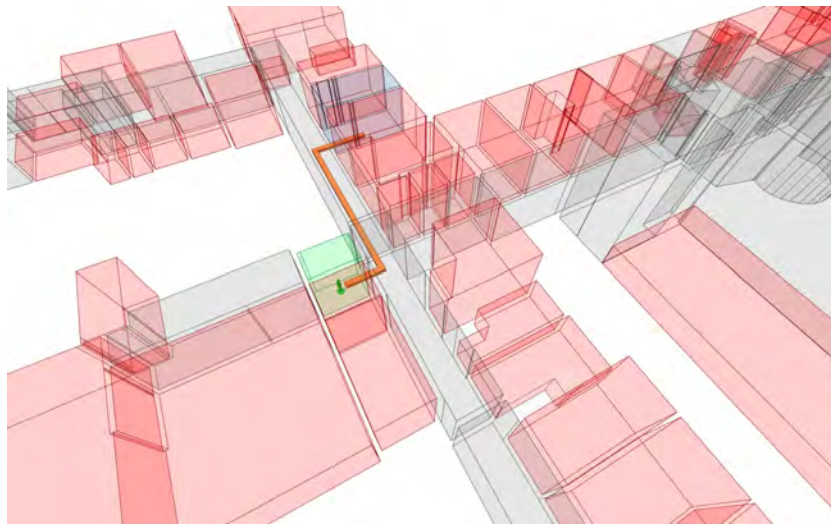
10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 21; pp. 21:1–21:12

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



■ **Figure 1** Indoor navigation path.

1 Introduction

Over recent years, the research area of navigation has become very active with an extensive variety of applications. Navigation is essential but also complex human activity. While initially navigation systems have been established for outdoor environments (such as for cars on the road), presently they have subsequently developed to be an essential field of interest for indoors (Makri et al. [11]). According to (Klepeis et al. [8]) around 87% of the people in the USA spend most of their lives inside buildings and the movement of the user of the indoor environment has been affected by the massive size of the indoor environment. The public buildings in our cities such as airports, train stations, hospitals, offices and university buildings, confront users with difficulties to find their destinations, and thus various research has been carried out that has resulted in many navigation models as shown in Figure 1. In this paper we concentrate on IndoorGML, adopted as a standard by Open Geospatial Consortium (OGC).

IndoorGML delivers a framework for indoor navigation systems to offer a description of the indoor space and provide Geography Markup Language (GML) syntax for encoding geoinformation (Zlatanova et al. [15], Kang and Li [6]). IndoorGML consists of two parts, first the core data model which describes geometry and topology connectivity, and second, a data navigation model that provides semantics for the navigation process (Lee et al [10]). The main purpose is to establish a methodology to classify spaces (rooms, corridors, etc.) and their indoor characteristics rather than represent architectural elements (Li, [9]). However, the current version on IndoorGML has incomplete UML model and that affects the quality of applications that depend on it. In this paper we propose an enhancement for the new version of the standard. We have also discussed alternatives in several cases and provided arguments pro and con each option and based on this selected best option. We classify some critical aspects that we have considered in this process:

- Complete attributes and code list for all classes.
- Better representation for the Primal space and Dual space.
- Clear terminology (vocabulary).
- Introducing geometry as attribute of classes (making the model more clear).

The methodology of this research is based on the following research phases: 1. Analyzing current version of IndoorGML and finding missing and weak parts, 2. Proposing options for solutions, 3. Discussion the pros and cons of the various options, 4. Selection the best option and make this part of improved IndoorGML proposal, 5. (Future work) develop technical model and populate with real data (to assess the conceptual model of IndoorGML) and further fine tune model when needed, 6. (Future work) bring our proposal into the standardization process within OGC (and collect opinion of the members of the IndoorGML team). The output of this investigation will be provided as input to OGC for an enhancement of the future version of the standard.

In Section 2, we discuss the research and developments related to IndoorGML in general and the UML model of IndoorGML, while in Section 3 we propose the enhanced UML model for IndoorGML. Finally, Section 4 concludes this paper.

2 Background

IndoorGML is an OGC standard that presents an elaboration of the indoor space and GML syntax for encoding geoinformation for the purpose of navigation (Zlatanova et al., [14]). IndoorGML defines a model to represent the geometry, topology and semantics of the indoor spaces which are used for the components of navigation network. The indoor and outdoor spaces differ from each other in many characteristics. Based on the indoor requirements for the spatial applications, the standard have to be reviewed with respect to the type of indoor applications. There are two categories indoor spatial applications: 1) managing the building components and facilities, and 2) using the indoor space. The first category mainly focuses on the architecture elements of the building such as walls and roofs (this discipline is called FM, facility management). The second category deals with the usage and localization properties of the indoor space, which refers to representing spaces such as rooms, corridors, and constraints elements such as doors. IndoorGML defines a framework to locate static or mobile objects (agents), and provide spatial information services (navigation) by using their positions in indoor space. IndoorGML represents the spatial character of the indoor spaces and provides information about their connectivity (Lee et al., [10]).

The indoor navigation research community broadly re-uses concepts such as Dual and Primal Space and automatic derivation of Dual Space that are part of IndoorGML (Diakite et al., [3]). Thus, research and developments depend on the standard to build applications based on the spatial framework of the standard. In that regard, software tools, e.g. an editor and a viewer have been developed by (Hwang et al., [4]) to support related studies. Concerns have been expressed about representation in 2D and 3D and the link between indoor and outdoor. (Kim and Lee, [7]) have proposed a semi-automatic approach to create IndoorGML data from images. In the same direction, (Mirvahabi and Abbaspour, [12]) have proposed an automated method to extract IndoorGML data from OpenStreetMap. (Diakite et al., [3]) have proposed a concept study for space subdivision to distinguish two significant aspects: the occupancy of the indoor space that influences the notation of indoor cells, and the description of criteria to support the automation of the space subdivision process. (Diakite and Zlatanova, [1]) have introduced an approach that creates the geometrical and topological valid IfcSpace classes in an IFC model, which can then be utilized for deriving a navigation network. Also, (Ryu et al., [13], Iida et al., [5]) have tried to enhance some characteristics of the current standards such as introduce attributes to support visually impaired people.

However, none of these researches have addressed the issues that relates to the UML model of IndoorGML. For navigation, it is important to include the access rights and/or restrictions of a user (group). When, developed a combined IndoorGML-LADM, we were confronted with the incompleteness of IndoorGML conceptual model (Alattas et al., [2]).

The current IndoorGML UML model contains the classes and their relationships as shown Figure 2. It has four different type of classes (GML, IndoorCore, IndoorNavi, and Not implemented). Most of the classes do not have attributes: no attribute names, no attribute data types. Further, the associations that link the classes has names that bring some confusion to the user. The GeneralSpace and the TransferSpace classes have attributes that contain the same code list values. But, if code lists are values equal, it is unclear what has to separate the code lists. The SpaceLayer class has a relationship with the CellSpace, State, and Transition classes and that create misunderstanding for the user of the standard (as it is not directly clear from the model that CellSpace/ State represents primal space and that Transition represents dual space). Furthermore, including AbstractFeature class is not the best way for illustration, because it has many relationships with other classes. Furthermore the type of the link is a generalization with lines in the illustration to nearly all other classes: spaghetti drawing. The standard represents the geometry data as separate classes and that allows mixing of the geometries to different objects (which could have been sharper typed). In addition, having geometry as separated classes in the illustration of the model increases the number of boxes and lines (i.e. creates spaghetti feeling). Therefore, in this paper we carry out a deeper study on several issues that relate to the UML class diagram and provide an enhancement for the new version of the standard.

3 Proposed UML Model for IndoorGML

In this section we present the proposed improvements, refinements and changes to the IndoorGML conceptual model. The current UML classes of IndoorGML are represented in pink color and the proposed UML classes are presented in light blue color.

3.1 From Classes to Attributes

Solid, Surface, Point, and Curve are geometry classes (as defined in ISO 19107) in the current version of IndoorGML with associations to classes that have geometric representation. Although this approach might be beneficial for keeping consistency, it is rather unclear for implementation. Therefore, we propose to convert the classes into attributes. The CellSpace class will have two additional attributes to represent the geometry data types as shown in Figure 3. The 3DGeometry attributes will have the GM_Solid value, and the 2DGeometry attributes will have the GM_Surface value. The CellSpace class will have a constraint that only one of the attributes (3DGeometry or 2DGeometry) has to be filled to ensure that the user correctly using the standard. The CellSpaceBoundary will have two additional attributes, first 3DGeometry attribute that has the value GM_Surface, and, second 2DGeometry attribute that has the value GM_Curve. The CellSpaceBoundary class will have a constraint that only one of the attributes (3DGeometry or 2DGeometry) has to be filled based on the type of geometry that has been used in the CellSpace class. Because the geometry of the CellSpace can (conceptually) be derived from the geometry of the associated boundaries, this is indicated with a forward slash before attribute name; e.g. /2DGeometry.

The Point Geometry type will be added as an attribute to the NodeInDualSpace class and the RouteNode class (for intermediate point) as an attribute that call Location and has the value GM_Point as shown in Figure 4. The curve geometry type will be added to EdgeInDualSpace class and RouteSegment class (for route parts) as an attribute that call Geometry and has the value GM_Curve as shown in Figure 5.

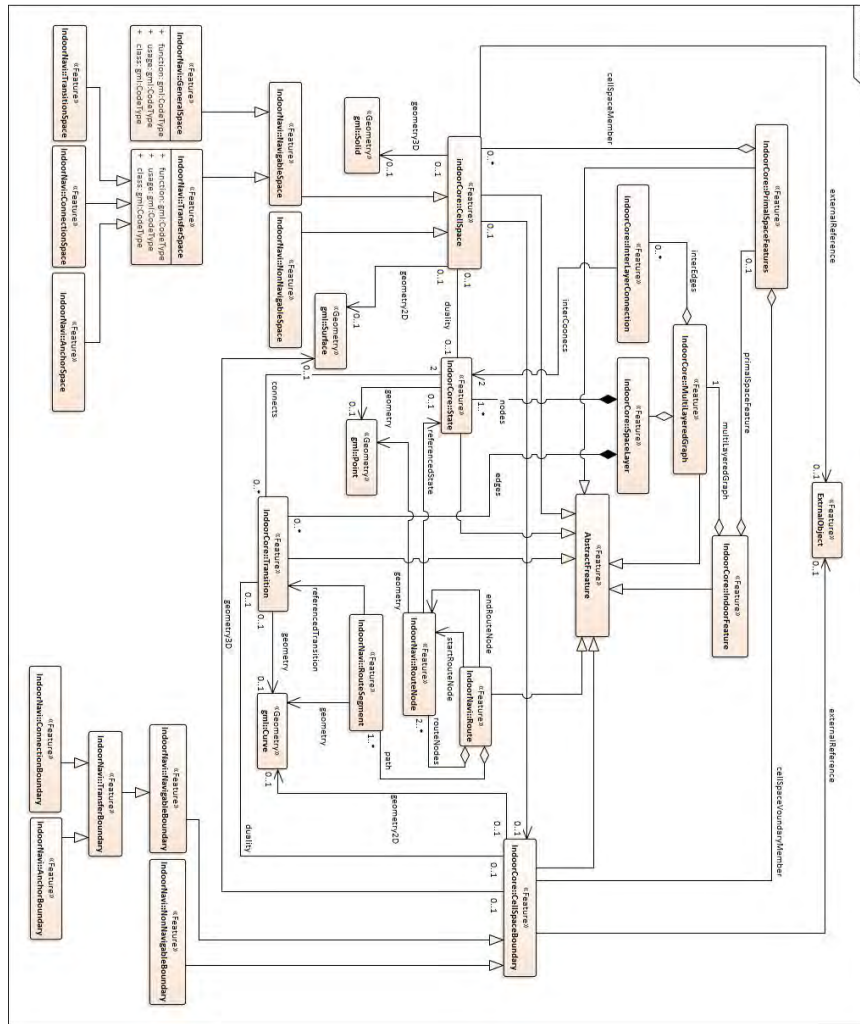


Figure 2 The current UML model of IndoorGML.

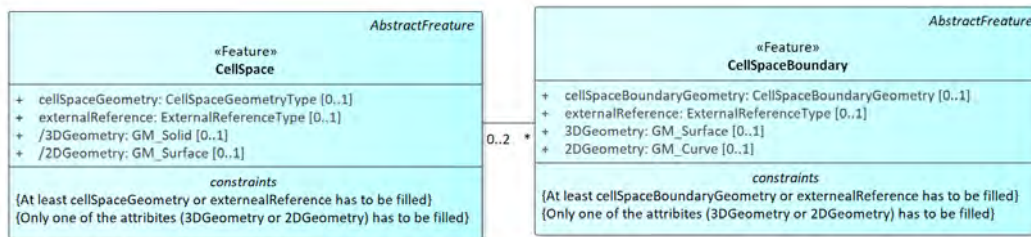
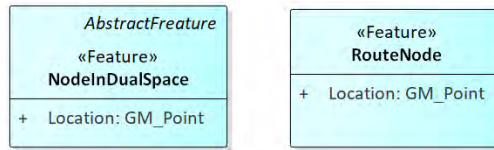


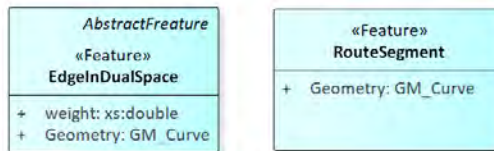
Figure 3 Additional geometry attributes for CellSpace class and CellSpaceBoundary and their constraints.

3.2 ExternalObject Class

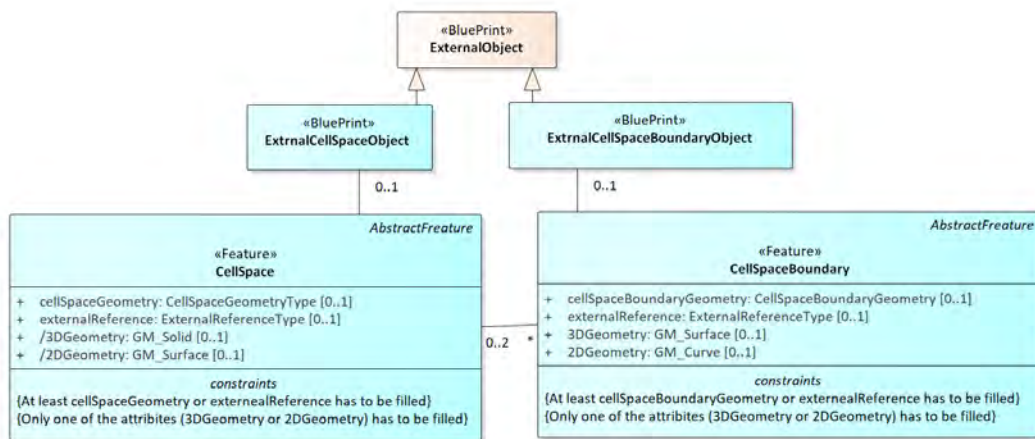
The current UML model contains ExternalObject class that has an association with CellSpace class and CellSpaceBoundary as shown in Figure 1. We propose that the current ExternalObject class to have two external object classes. The new two classes will have



■ **Figure 4** New attributes for NodeInDualSpace class and RouteNode class.



■ **Figure 5** New attribute for EdgeInDualSpace class and RouteSegment class.

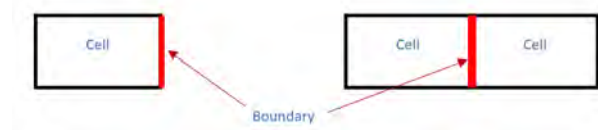


■ **Figure 6** The proposed ExternalObject classes.

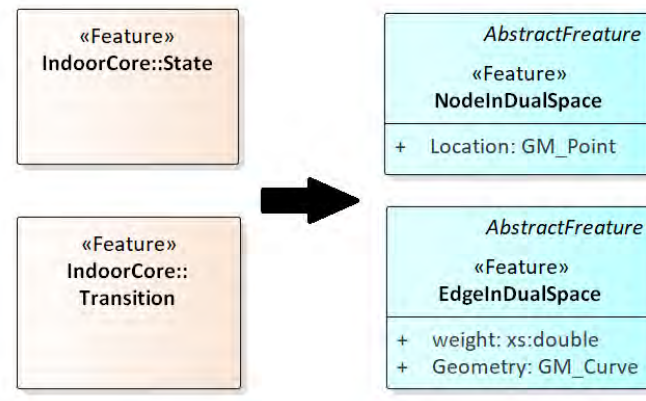
associations with the current ExternalObject class (as superclass), the new subclasses are also more precise typed. The CellSpace class will have an association with a new class that call “ExternalCellSpaceObjec” and it is responsible for providing the object reference of the Space from the ExternalObject class. Also, the CellSpaceBoundary class will have an association with a new class that is called “ExternalCellSpaceBoundaryObject” and it is responsible for providing the object reference of the boundary from the ExternalObject class. This method will bring more flexibility to the representation space and boundary as shown in Figure 6. Also, the type of the class of the ExternalObject has been changed from Feature type to the Stereotype «Blueprint», because this class represents a reference that not include in the model.

3.3 Association Multiplicity of CellSpace and CellSpaceBoundary

The association multiplicity between CellSpace and CellSpaceBoundary in the current version of the standard shows that each CellSpace has many Boundaries, and each CellSpaceBoundary has zero or one CellSpace as shown in Figure 1. However, in reality each CellSpaceBoundary could have one or two (or zero if boundary not used) CellSpace as shown in Figure 6.



■ **Figure 7** CellSpaceBoundary could have one or two CellSpace.



■ **Figure 8** Proposed terms for State and Transition classes.

Furthermore, in case of so called functional areas or virtual spaces, the neighbor cells do share a one boundary. The multiplicity has been modified as shown in Figure 7.

3.4 The terms State and Transition

We propose to change the terms State and Transition into the more intuitive terms Node and Edge. In addition, we suggest adding the Dual terms to each class and that will make them understandable for the user that they are belong to the Dual space. The term State has been changed to NodeInDualSpace and the term Transition has been changed to EdgeInDualSpace as shown in Figure 8.

3.5 Code Lists

The current version of the standard has the same code list values for GeneralSpace class and TransferSpace class (gml:CodeType). We have changed that by adding different names for the code list as shown in Figure 9 (in total 7 different code lists).

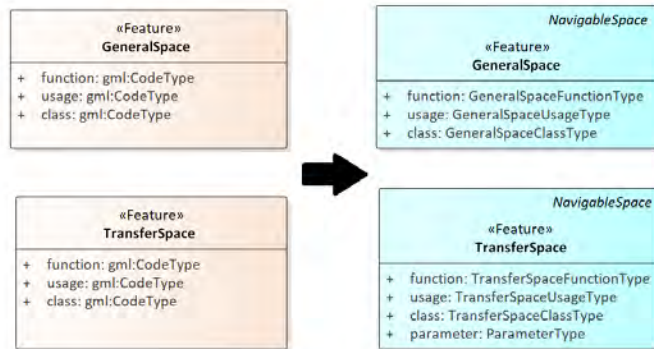
The GeneralSpace class has three attributes (function, usage, and class) and each attribute has a code list value, with example code list values as shown in Figure 10. The Usage attributes has a code list values that represent the user groups of the space such as student group, employee group, and visitor group.

The ConnectionSpace class is a subclass of the TransferSpace and it has three attributes (function, usage, and class) and each attribute has a code list value as shown in Figure 11.

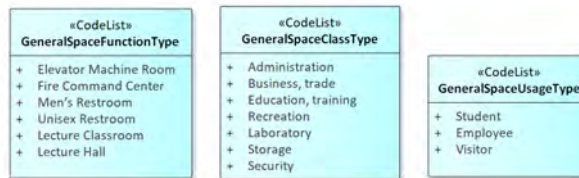
The AnchorSpace class is a subclass of the TransferSpace and it has three attributes (function, usage, and class) and each attribute has a code list value as shown in Figure 12.

The SpaceLayer class has six attributes (usage, terminationDate, function, creationDate, and class). The class attribute has a code list type value which is the SpaceLayerClassType as shown in Figure 13. Note that the values of an enumeration type are fixed (and can not be extended as for code lists).

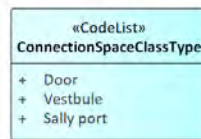
21:8 Improved and More Complete Conceptual Model for the Revision of IndoorGML



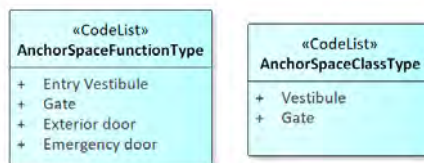
■ **Figure 9** New code list names for GeneralSpace and TransferSpace classes.



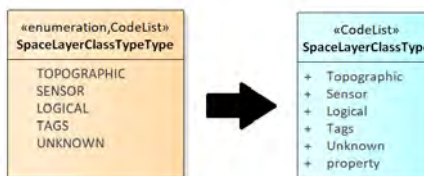
■ **Figure 10** Code list for the attributes of the GeneralSpace class (with example values).



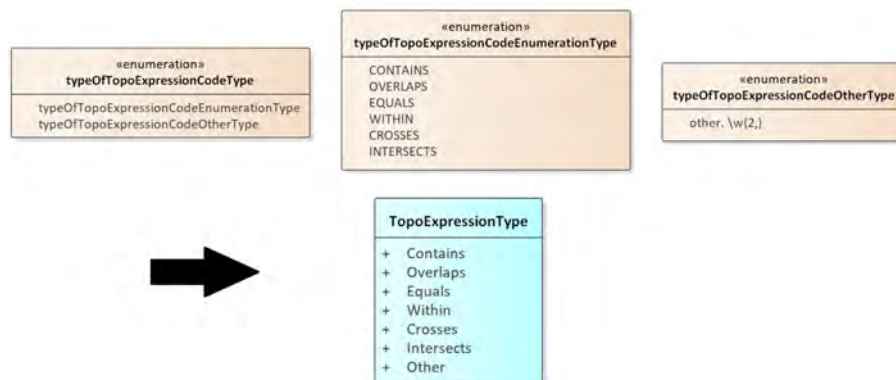
■ **Figure 11** Code list for the attributes of the ConnectionSpace class.



■ **Figure 12** Code list for the attributes of the AnchorSpace class.



■ **Figure 13** Code list values for the attributes of the SpaceLayer class.



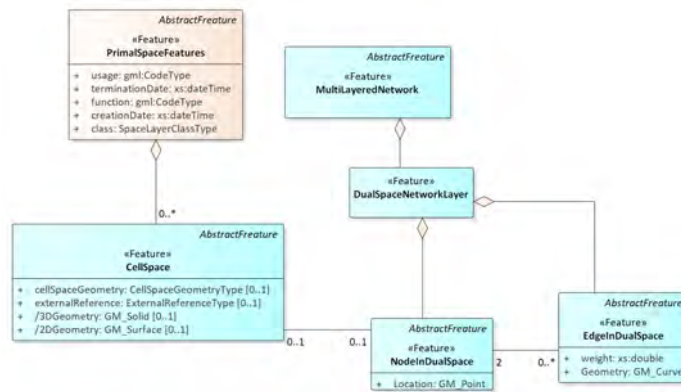
■ **Figure 14** Enumeration values for the attributes of the InterLayerConnection class.

The InterLayerConnection class has two attributes (typeOfTopoExpression and comment). The typeOfTopoExpression attributes has an enumeration value which is the typeOfTopoExpressionCodeType and it consists of two enumeration values (typeOfTopoExpressionCodeEnumerationType and typeOfTopoExpressionCodeOtherType), however, we propose to replace these 3 «enumeration» types with a single «codeList» that has the name TopoExpressionType as shown in Figure 14.

3.6 Classes and Associations

The current UML model contains an association between SpaceLayer class and NodeInDualSpace (State) class and EdgeInDualSpace (Transition) class have been defined as Composition association as shown in Figure 1. However, instead of connecting these two classes to the SpaceLayer, we have proposed a new feature class call DualSpaceNetworkLayer that will be as a collecting class for the Node and the Edge of the Dual space. We want to emphasize that the layers can be for both: the Primal and Dual Spaces. The SpaceLayer class will have an association with the CellSpace class and the SpaceLayer will be collecting class for the spaces of the primal space. The name of MultiLayerGraph class has changed to MultiLayerNetwork because a graph does not need to have geometry and in the case of IndoorGML there is a need for geometries at least for the Nodes. The MultiLayerNetwork will has association with the NodeInDualSpace and EdgeInDualSpace instead of the association with the SpaceLayer because it deals with the Dual space as shown in Figure 15.

Also, the current UML model of IndoorGML standard has defined names for the associations between the classes such as duality, edges, nodes, geometry, and partialBoundaryBy which bring a lot of confusing during the generating of the XML schema as shown in Figure 1 as these association (role) names are very close to the names of the involved classes (and add little/no value). The propose UML model does not include all the defined names of the associations to avoid confusing as shown in Figure 16. Additional, the TransferSpace class and CellSpaceBoundary has parameter attributes that have the type (virtual, real) to allow aggregation and subdivision of CellSpaces. Furthermore, the TransitionSpace class has been removed from the UML class diagram because it is difficult to semantically distinguish this from the ConnectionSpace class.



■ **Figure 15** New DualSpaceNetworkLayer class and their associations with the NodeInDualSpace class and EdgeInDualSpace class.

4 Conclusion

In this paper, we proposed an enhancement for the UML class diagram of IndoorGML standard. We suggested the following improvements for the conceptual model (as input for the revision of the standard within OGC, See Section four for more details):

- The ExternalObject class has two subclasses (ExternalCellSpaceObject and ExternalCellSpaceBoundaryObject). The CellSpace has an association with the ExternalCellSpaceObject and the CellSpaceBoundary have an association with ExternalCellSpaceBoundaryObject to improve the concept behind the ExternalObject class.
- Association multiplicity of CellSpace and CellSpaceBoundary is corrected.
- The terms State and Transition are changed into NodeInDualSpace and EdgeInDualSpace because they better represent the nature of these classes and improve the perception.
- The geometry classes are converted into attributes of the classes that need them to ensure better understanding during the implementation from the user.
- GeneralSpace class and TransferSpace class have different names for the code list and we have created code list classes to define the values for each attribute.
- DualSpaceNetworkLayer is introduced as a collecting class for the node and the edge of the dual space. The SpaceLayer has an association with the CellSpace class only and is a collecting class for the spaces of the primal space.
- TransferSpace class and CellSpaceBoundary have additional attributes that have the value (virtual, real) to allow aggregation and subdivision of CellSpaces.

This paper comes as a proposal for IndoorGML to include the above-mentioned suggestions. Additional investigation is required to define attributes for all classes. This paper reflects the initial developments of a more complete and enhanced conceptual model for IndoorGML. The future work includes additional investigations to define more attributes for the classes as well as development of prototype implementations such as SQL implementation, XML encoding, and Application with GUI. All of them will be based on same conceptual IndoorGML model. Furthermore, we will bring our proposal into the standardization process within OGC. This is expected to validate the proposed model extension further and accelerate the development of indoor navigation applications.


- 4 Jung-Rae Hwang, Hye-Young Kang, and Jin-won Choi. Development of an editor and a viewer for IndoorGML. *Proceedings of the Fourth ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness - ISA '12*, page 37, 2012. doi:10.1145/2442616.2442625.
- 5 Hirokazu Iida, K E I Hiroi, Katsuhiko Kaji, and Nobuo Kawaguchi. A Proposal of IndoorGML Extended Data Model for Pedestrian-Oriented Voice Navigation System. *ACM SIGSpatial Workshop on ISA*, 3(Figure 1), 2015. doi:10.1145/2834812.2834814.
- 6 Hae-Kyong Kang and Ki-Joune Li. A Standard Indoor Spatial Data Model—OGC IndoorGML and Implementation Approaches. *ISPRS International Journal of Geo-Information*, 6(4):116, 2017. doi:10.3390/ijgi6040116.
- 7 M. Kim and J. Lee. Developing a method to generate IndoorGML data from the omnidirectional image. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 40(2W4):17–19, 2015. doi:10.5194/isprsarchives-XL-2-W4-17-2015.
- 8 N E Klepeis, W C Nelson, W R Ott, J P Robinson, A M Tsang, P Switzer, J V Behar, S C Hern, and W H Engelmann. The National Human Activity Pattern Survey (NHAPS): a resource for assessing exposure to environmental pollutants. *Journal of exposure analysis and environmental epidemiology*, 11(3):231–252, 2001. doi:10.1038/sj.jea.7500165.
- 9 Ki Joune Li. Indoorgml - A standard for indoor spatial modeling. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 41(July):701–704, 2016. doi:10.5194/isprsarchives-XLI-B4-701-2016.
- 10 Ki-Joune Li, Jiyeong Lee, Sisi Zlatanova, Thomas H. Kolbe, Claus Nagel, and Thomas Becker. OGC® IndoorGML. *Open Geospatial Consortium*, pages 1–17, 2015. doi:http://www.opengeospatial.org/.
- 11 A. Makri, S. Zlatanova, and E. Verbree. an Approach for Indoor Wayfinding Replicating Main Principles of an Outdoor Navigation System for Cyclists. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 40(4W5):29–35, 2015. doi:10.5194/isprsarchives-XL-4-W5-29-2015.
- 12 S. S. Mirvahabi and R. A. Abbaspour. Automatic extraction of IndoorGML core model from OpenStreetMap. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 40(1W5):459–462, 2015. doi:10.5194/isprsarchives-XL-1-W5-459-2015.
- 13 Hyeong-Gyu Ryu, Taehoon Kim, and Ki-Joune Li. Indoor navigation map for visually impaired people. *Proceedings of the Sixth ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness - ISA '14*, pages 32–35, 2014. doi:10.1145/2676528.2676533.
- 14 S Zlatanova, K J Li, Christiaan Lemmen, and Peter J M van Oosterom. Indoor Abstract Spaces: Linking IndoorGML and LADM. *5th International FIG 3D Cadastre Workshop*, pages 317–328, 2016.
- 15 S. Zlatanova, P. J. M. Van Oosterom, J. Lee, K.-J. Li, and C. H. J. Lemmen. Ladm and Indoorgml for Support of Indoor Space Identification. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-2/W1(October):257–263, 2016. doi:10.5194/isprs-annals-IV-2-W1-257-2016.

Design for Geospatially Enabled Climate Modeling and Alert System (CLIMSYS): A Position Paper

Devanjan Bhattacharya¹

Nova Information Management School, Universidade Nova de Lisboa, Campus de Campolide, Lisbon, Portugal


dbhattacharya@novaims.unl.pt

 <https://orcid.org/0000-0002-3382-9523>

Marco Painho

Nova Information Management School, Universidade Nova de Lisboa, Campus de Campolide, Lisbon, Portugal

painho@novaims.unl.pt

 <https://orcid.org/0000-0003-1136-3387>

Abstract

The paper brings the focus on to multi-disciplinary approach of presenting climate analysis studies, taking help of interdisciplinary fields to structure the information. The system CLIMSYS provides the crucial element of spatially enabling climate data processing. Even though climate change is a matter of great scientific relevance and of broad general interest, there are some problems related to its communication. Its a fact that finding practical, workable and cost-efficient solutions to the problems posed by climate change is now a world priority and one which links government and non-government organizations in a way not seen before. An approach that should suffice is to create an accessible intelligent system that houses prior knowledge and curates the incoming data to deliver meaningful results. The objective of the proposed research is to develop a generalized system for climate data analysis that facilitates open sharing, central implementation, integrated components, knowledge creation, data format understanding, inferencing and ultimately optimal solution delivery, by the way of geospatial enablement.

2012 ACM Subject Classification Information systems → Geographic information systems, Information systems → Expert systems, Information systems → Sensor networks

Keywords and phrases Spatial enablement, climate modeling, natural hazards, spatial data infrastructure, sensor web

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.22

Category Short Paper

Funding D. Bhattacharya has been funded by the European Commission through the GEO-C project H2020-MSCA-ITN-2014, Grant Agreement number 642332, <http://www.geo-c.eu/>

1 Introduction

The focus is growing sharper than ever on climate research activities. Now is the time to respond with a global system about generalized climate modeling at any scale and expert decision support. With the advent of sensors for monitoring, data collections for any event are at unprecedented levels. For example, in climate data processing the major hurdles are

¹ Supported by GEO-C-H2020-MSCA-ITN-2014-642332



© Devanjan Bhattacharya and Marco Painho; licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 22; pp. 22:1–22:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

that the different research groups globally are processing their data in silos, most of the time repeating same processes at each location, creating similar metadata each time, duplicating data, thereby falling behind the rushing stream of more incoming data. The solution could be addressed through integrating data source, spatial data platform, data understanding, knowledge base, inferencing and visualization into a single, well-connected online real-time system. Such a spatial decision support system (DSS) with expert knowledge bases will not only serve the critical research of climate modeling but do so to any research relying on real-time data capture and analysis with spatial domain of data being the unique enabler[8].

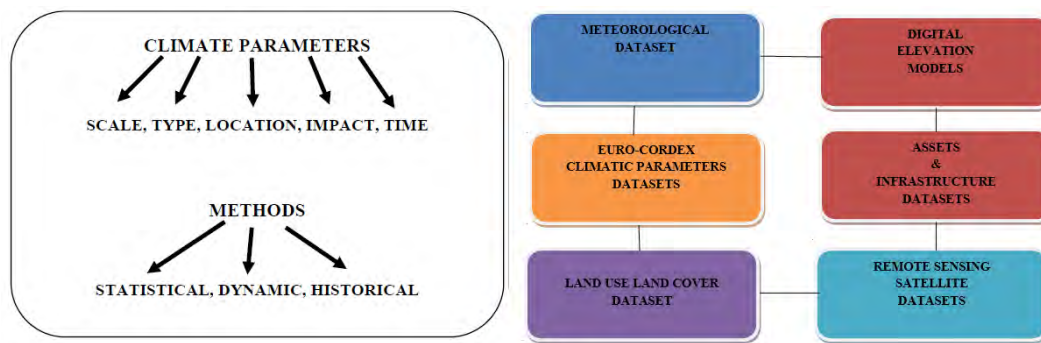
The objective of the proposed research is to develop a generalized system for climate data analysis that facilitates open sharing, central implementation, integrated components, knowledge creation, data format understanding, inferencing and ultimately optimal solution delivery, all through open-source development. It should enable a climate scientist located anywhere to utilize data sources, create algorithms, models and output layers of climate information. The core of the system development will be to design optimal knowledge base (KB) and expert system (ES) for climatic scenarios. The research questions to be answered through this research are: i) how to build open source spatial ontologies for climate phenomenon using causative factors ii) how to connect ontologies for climate to intelligent inferencing logics iii) how to build specialized knowledge bases for a generalized climate modeling DSS iv) how to apply the system to automate procedures such as climate extreme indices and downscaling urban climate extremes v) how to integrate sensor web(SW), other data sources and spatial data infrastructure(SDI) with open source technologies.

2 Background Literature Review

Several studies [5, 2] over the years and recently [9, 1] have heavily stressed the need for developing a system capable of encapsulating the entire essence of climate studies in one platform which can be open, shareable, knowledgeable, and contributable globally. CLIMSYS aims to address these challenges through developing a framework that houses data, metadata, understanding of the data, knowledge to be applied on the data, and output from the data. CLIMSYS would enable a distributed spatial framework that targets to deliver climate based decisions to start with but would be capable of administering spatial functionalities to a variety of social needs.

Climate data are dramatically increasing in volume and complexity, just as the users of these data in the scientific community and the public are rapidly increasing in number. A new paradigm of more open, user-friendly data access is needed to ensure that society can reduce vulnerability to climate variability and change, while at the same time exploiting opportunities that will occur. The burgeoning types and volume of climate data alone constitute a major challenge to the climate research community and its funding bodies. Institutional capacity must exist to produce, format, document, and share all these data, while, at the same time, a much larger community of diverse users clamors to access, understand, and use climate data [8]. Fig 1 shows the interoperability issues, due to multi-input types, in the engineering processes due to the application of many sets of domain data that stresses the multidisciplinary nature of the problem. The engineering process is based on reusing the existing knowledge representation models.

Research to action has been the clarion call from several climate critiques, wherein the papers have concluded that scientists need to relay the valuable work they are doing through impactful interfaces[3]. Through the present paper we want to convey that such an interface is imminently possible through the integration of sensor web and SDIs on top of

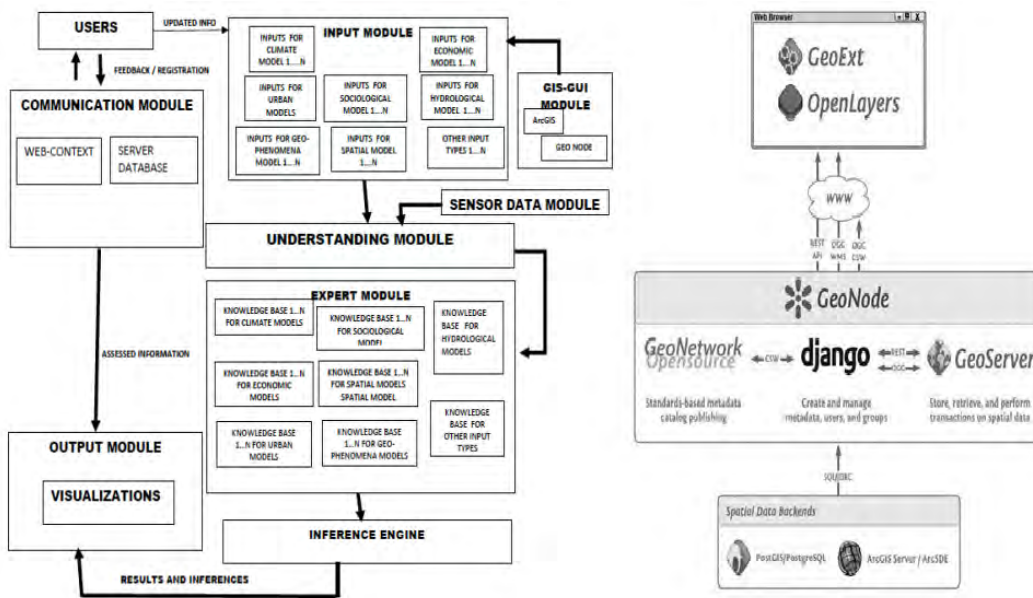


■ **Figure 1** Input Parameters of Climate Data and Methods Adopted.

pertinent expert knowledge bases. We can now delve into the background of SDIs and sensor web. Considerable research and development has been carried out in SDI in recent years. Some are trying to develop high-level middleware services and domain-specific services for problem-solving and scientific discovery in infrastructures [4]. For example, the Group on Earth Observation (GEO) Model Web initiative proposes to provide environmental models as services and integrating distributed models in infrastructures. With these systems it is seen that they tend to be case-specific and restricted. Also the design is not broad enough to accommodate increasing number of formats. Hence CLIMSYS is to be designed to be more generalized, integrable with multiple domains and formats and the biggest addition will be the availability of pluggable KBs that infuse better understanding of the data. With the integration with sensor web CLIMSYS will provide long term benefits.

3 Methodology

CLIMSYS utilizes a distributed SDI including data models, applications and services based on OGC standards and their benchmarking and evaluation are the objectives of this proposed research. The initial architecture as shown in Fig 2a for the shared data concept has been implemented to categorize and modularize input domains, sensor-web module, data understanding module, expert KB, inferencing, and output. In Fig 2b the GIS graphical user interface(GUI) has been expanded to show the implementation of GEONODE structure. It handles the spatial database and spatial analysis. GeoNode provides the distributed SDI environment (Fig 2b). The concept of plug-ins to interact amongst themselves from one framework to another makes the integration of SDI and sensor web possible. The web-enablement in Fig 3a is where the architecture to capture geospatial elements and transmitting over the web is taken care of by webGIS standards and the open interfaces are utilized for latching on to the sensor network through a set of GML Clients. Therefore, a consistent set of encoding and interface standards are mandatory for adapting and integrating sensor networks into an SDI application. In CLIMSYS, we present how the reused models were interconnected, starting from the analysis of the interoperability needs of the existing and planned data sources, the use of a core ontology as integration strategy, and the modeling of concepts that carry out the interconnection among the reused models. The work then must solve the key interoperability issues using visualization tools and representative scenarios. Experiments on an information recovery study stress the potential of the proposed ontology, its limitations, and future challenges in the modeling process. We expect to contribute with ideas about an ontology engineering process for semantic interoperability of multidisciplinary



■ **Figure 2** a) CLIMSYS Proposed Architecture b) GeoNode and database functionalities.

domains (Fig 3a, b), as well as to present experiences from applying this process. As can be ascertained from Fig 1, the disparity in data formats in sensor web (Fig 3a) needs proper ontology to understand the data contents and semantic context (Fig 3b). The joining of sub-systems happens at corresponding levels like input module with data layer of sensor web and database backend of GeoNode, understanding module and expert module interface with web service layer of sensor web and middleware of GeoNode. The Output module interfaces with the application layer and front end of GeoNode.

4 Results

The joining of geospatial datasets and knowledge bases has been done to utilize the complete set of information available in each of them. There are many open source geospatial datasets available such as GeoNames, Open Street Map, Natural Earth and to get a comprehensive dataset with the union of all available information it is important that such datasets are linked optimally without redundancy or loss of information. The multi-interfacing that is captured by Fig 2a allows for spatial interface, input, storage, incremental upgrades, and output communication. The interfaces use Java apps with Python codes. The Geonode architecture has PostGIS and PostGresql backend, and HTML frontend (Fig 2b). The HTML frontend of GeoNode displays the global basemap (Fig 4) and the sensed data[6, 7] input and processed by CLIMSYS is layered on top of the basemap. By clicking on the place-name (Czech Republic) the temperature data and snowing data for the region (Jan-Mar 2018) are displayed.

5 Discussions and Conclusions

A gamut of information about the environment - land, air, water, weather, climate and natural and man-made risks can be harnessed by seamless and rapid access to sensors. In addition, sensors are critical components of building, transportation, utility and industry

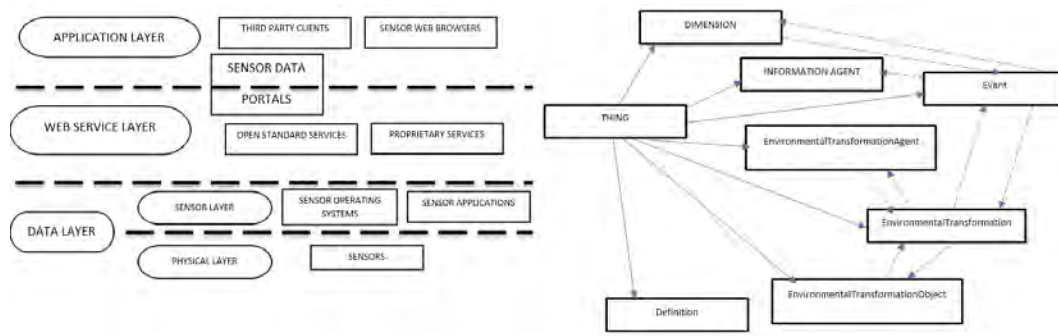


Figure 3 a) The layers for integration of sensor web with SDI b) Ontological snapshot of an environmental process for system development.

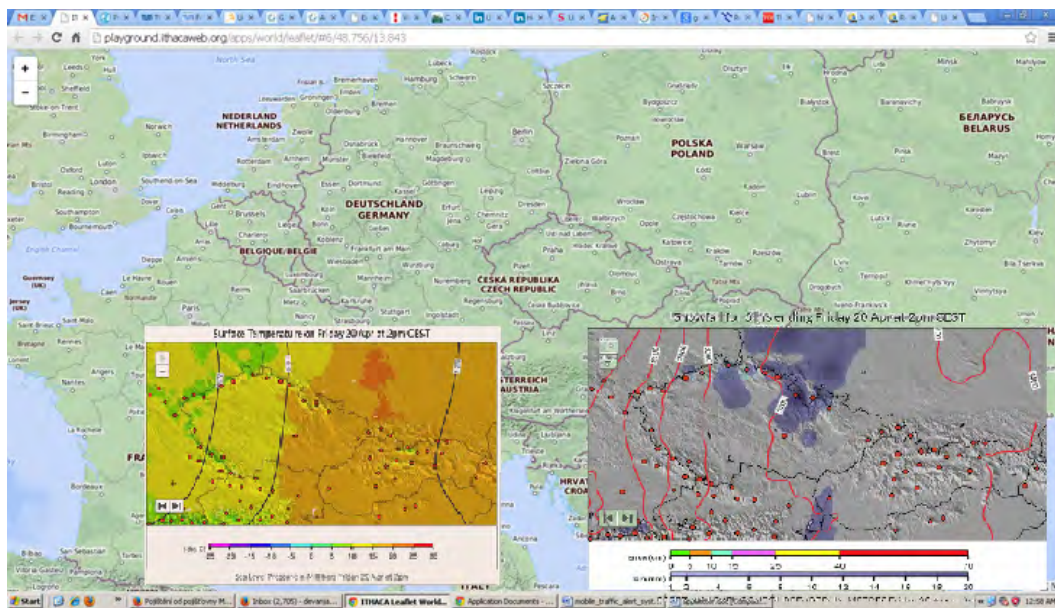


Figure 4 Integrated sensor data for temperature(left box) and snowing(right box) in Czech Republic over GeoNode basemap.

infrastructure. The ability to harness and render this information in a location context is a major challenge. Until recently though, there were no facilitating standards to make it easier to discover, access and integrate this information. Therefore, a consistent set of encoding and interface standards are mandatory for adapting and integrating sensor networks into an SDI application. Both, SDI (web mapping) standards and sensor web enablement standards from OGC, have to meet at a common ground and connect together. The integration of sensor web and SDI in open source domain could be achieved possibly by setting up one to one correspondence between their services through functions calling and methods calling.

CLIMSYS can deliver an integrated sensor web and SDI which can solve a lot of challenges that stand-alone, disconnected, case-specific, and customized systems lack. The next level of capability for both SDI and sensor web would be to evolve into a new realm of a location enabled and semantically enriched Geospatial Web or Geosemantic Web but additionally with spatial analytics capabilities. The SDI has the capacity to integrate with distributed computing and database platforms and enable the Geospatial Web with capabilities of data

democratization. Hence in conclusion, it is of pressing importance to geospatial studies to integrate SDI with Sensor Web. The integration can be done through merging the common OGC interfaces of SDI and Sensor Web. Through CLIMSYS, Sensor Web and SDI are going to keep expanding in the next decade. Sensors are going to be so ubiquitous that similar to the world wide web the addition of vast number of sensors will keep happening like new data sources of present internet. The concept of CLIMSYS has to keep evolving to help overall development.

References

- 1 Gregory Giuliani, Stefano Nativi, Andre Obregon, Martin Beniston, and Anthony Lehmann. Spatially enabling the global framework for climate services: Reviewing geospatial solutions to efficiently share and integrate climate data and information. *Climate Services*, 8:44–58, 2017. doi:10.1016/j.cliser.2017.08.003.
- 2 C.S.B. Grimmond, M. Roth, T.R. Oke, Y.C. Au, M. Best, R. Betts, G. Carmichael, H. Cleugh, W. Dabberdt, R. Emmanuel, E. Freitas, K. Fortuniak, S. Hanna, P. Klein, L.S. Kalkstein, C.H. Liu, A. Nickson, D. Pearlmutter, D. Sailor, and J. Voegt. Climate and more sustainable cities: Climate information for improved planning and management of cities (producers/capabilities perspective). *Procedia Environmental Sciences*, 1:247–274, 2010. doi:10.1016/j.proenv.2010.09.016.
- 3 J.N. Lavis, J. Lomas, M. Hamid, and N.K. Sewankambo. Assessing country-level efforts to link research to action. *Bulletin of the World Health Organization*, 84(8):620–626, 2006.
- 4 Steve H.L. Liang and Chih-Yuan Huang. Geocens: A geospatial cyberinfrastructure for the world-wide sensor web. *Sensors*, 13(10):13402–13424, 2013. doi:10.3390/s131013402.
- 5 S. Nativi, M. Craglia, and J. Pearlman. Earth science infrastructures interoperability: The brokering approach. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(3):1118–1129, June 2013. doi:10.1109/JSTARS.2013.2243113.
- 6 NOAA. NOAA national centers for environmental information. <https://www.ncdc.noaa.gov/data-access/quick-links#dsi-3505> (accessed on 20 Apr 2018).
- 7 NOAA. NOAA national ice center. http://www.natice.noaa.gov/Main_Products.htm (accessed on 20 Apr 2018).
- 8 Jonathan T. Overpeck, Gerald A. Meehl, Sandrine Bony, and David R. Easterling. Climate data challenges in the 21st century. *Science*, 331(6018):700–702, 2011. doi:10.1126/science.1197869.
- 9 O. Rössler, A. M. Fischer, H. Huebener, D. Maraun, R. E. Benestad, P. Christodoulides, P. M. M. Soares, R. M. Cardoso, C. Pagé, H. Kanamaru, F. Kreienkamp, and D. Vlachogiannis. Challenges to link climate change data provision and user needs—perspective from the cost-action value. *International Journal of Climatology*, 0(0), 2017. doi:10.1002/joc.5060.

Geographical Exploration and Analysis Extended to Textual Content

Raphaël Ceré

Department of Geography and Sustainability, University of Lausanne, Switzerland

Raphael.Cere@unil.ch

Mattia Egloff

Department of Language and Information Sciences, University of Lausanne, Switzerland

Mattia.Egloff@unil.ch

François Bavaud

Department of Language and Information Sciences & Department of Geography and Sustainability, University of Lausanne, Switzerland

Francois.Bavaud@unil.ch

Abstract

Textual and socio-economical regional features can be integrated and merged by linearly combining the between-regions corresponding dissimilarities. The scheme accommodates for various squared Euclidean socio-economical and textual dissimilarities (such as chi² or cosine dissimilarities derived from document-term matrix or topic modelling). Also, spatial configuration of the regions can be represented by a weighted unoriented network whose vertex weights match the relative importance of regions. Association between the network and the dissimilarities expresses in the multivariate spatial autocorrelation index δ , generalizing Moran's I , whose local version can be cartographed. Our case study bears on the Wikipedia notices and socio-economic profiles for the 2251 Swiss municipalities, whose weights (socio-economical or textual) can be freely chosen.

2012 ACM Subject Classification Mathematics of computing → Probability and statistics, Information systems → Clustering

Keywords and phrases Spatial autocorrelation, Weighted spatial network, Document-term matrix, Multivariate features, Soft clustering

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.23

Category Short Paper

1 Introduction

Spatial analysis deals with notions of “*where*” (the spatial configuration of regions), “*what*” (the regional features) and “*how much*” (the relative importance of regions, as given by their surface, the population size or terms size). *The aim of this contribution is to propose a formalism and a case study showing how to **directly incorporate textual information**, in the frequent situation where each region is described by a text.* In a nutshell, both socio-economic and textual features can be encoded in a dissimilarity matrix between regions, and linearly combined in a flexible way, producing new dissimilarities mixing both kind of features. The latter can be further used for multidimensional scaling, or distance-based clustering.

Socio-economic features can be spatially auto-correlated, and so are the textual features. Section 2 presents a general formalism for assessing and testing spatial autocorrelation and its local indicators, able to deal with multivariate features. Its application requires the



© Raphaël Ceré, Mattia Egloff, and François Bavaud;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 23; pp. 23:1–23:7

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

dissimilarities to be squared Euclidean, which leaves open many possibilities and variants, in particular regarding the information retrieval processing of the document-term matrix.

The formalism also represents the spatial configuration of the regions as an unoriented weighted network, where the node weights represent the importance of regions, and the edge weights is a measure of accessibility, larger between spatially close regions. Requiring the sum of the edge weights associated to a region to equal the regional weight is natural and mathematically convenient. Among various possible choices, we adopt here the *diffusive weighted specification*, yielding a family of weight-compatible networks index by a single parameter $t > 0$, the diffusion time. The regional weights themselves can be chosen as proportional to the residential population, or proportional to the document sizes, and this choice has a deep impact on the behaviour of the quantities under consideration, as illustrated in the case study presented in section 3.

2 Formalism and definition

We consider a set of n regions, characterized by textual descriptions, as well as by socio-economic features. The former are typically specified by a $n \times v$ document-term matrix X^{text} , giving, after the usual textual pre-processing, the number of occurrences of term $w = 1, \dots, v$ in the document describing region $i = 1, \dots, n$. The latter are specified by a $n \times p$ matrix X^{se} contains the p socio-economic features of interest, such as the proportions of inhabitants belonging to specific ages, nationalities, professional types, the proportions of buildings of a given type, etc.

Regions differ by their importance, as specified by relative weights $f_i > 0$ with $\sum_{i=1}^n f_i = 1$. Regional weights can be chosen as reflecting the document sizes f_i^{text} , or the population share f_i^{se} as in standard socio-economic geographic analysis. Finally, the spatial configuration of the n connected regions is specified by a binary $n \times n$ adjacency matrix $A = (a_{ij})$.

2.1 A general framework for spatial autocorrelation

Dissimilarities between regional features may, on average, be smaller between spatially close regions, and this precisely constitutes the issue of spatial autocorrelation. A general framework, permitting to attribute differing weights to regions, whose spatial proximity is modelled by a weighted unoriented network, and whose features can be multivariate, relies on two ingredients :

1. a $n \times n$ symmetric joint probability matrix $E = (e_{ij})$, referred to as the *exchange* matrix, giving the probability to select a pair ij of regions, the idea being that e_{ij} is proportional to the relative weights f_i and f_j of the regions, and decreasing with their spatial distance.
 2. a $n \times n$ symmetric dissimilarity matrix $D = (D_{ij})$, where $D_{ij} = \|\vec{x}_i - \vec{x}_j\|^2$ is a squared Euclidean distance between suitably normalized multivariate regional features \vec{x}_i and \vec{x}_j .
- In addition, and crucially, the exchange matrix is required to be *weight compatible*, that is its margins yield the regional weights, that is $e_{i\bullet} = \sum_{j=1}^n e_{ij} = f_i$, where f_i can be interpreted as the probability to select region i .

The global inertia, respectively local inertia, measures the average dissimilarity between randomly selected regions, respectively between neighbours. Their comparison provides an autocorrelation index δ which constitutes a multivariate generalization of *Moran's I*. They read, in order,

$$\Delta = \frac{1}{2} \sum_{i,j=1}^n f_i f_j D_{ij} \quad \Delta_{\text{loc}} = \frac{1}{2} \sum_{i,j=1}^n e_{ij} D_{ij} \quad \delta = \frac{\Delta - \Delta_{\text{loc}}}{\Delta} \quad (1)$$

The values of δ range in $[-1, 1]$, and its standardized value $z = (\delta - E_0(\delta))/\sqrt{\text{Var}_0(\delta)}$ can be tested in the normal approximation [2, 3].

Regional dissimilarities D_{ij} can, as in spatial econometrics and quantitative geography, reflect their socio-economic profiles, but also, and more originally, the textual content of their description, or a mixture of both. All the involved similarities should be squared Euclidean, and this constitutes a necessary and sufficient condition for the application of the formalism. For comparison sake, they should also be preliminary standardized as $D_{ij} \leftarrow D_{ij}/\Delta$.

Local multivariate indicators of spatial autocorrelation. Local multivariate indicators of spatial autocorrelation [1], measuring the average scalar product of the deviations at a region and at its neighbours, can be constructed as

$$\delta_i = \frac{(WB)_{ii}}{\Delta} \quad \text{with} \quad W = \text{diag}(1/f)E \quad \text{and} \quad B = -\frac{1}{2}HDDH', \quad \text{where} \quad H = I - \mathbf{1}f' \quad (2)$$

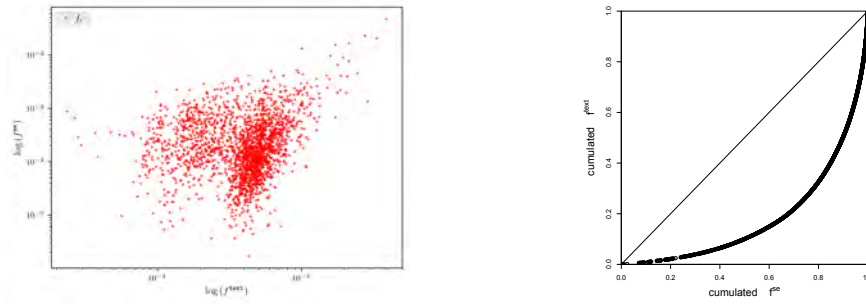
and satisfy $\sum_i f_i \delta_i = \delta$. Here W is the row-standardized $n \times n$ matrix of spatial weights, and constitutes the transition matrix of a reversible Markov chain with stationary distribution f . Also, $B = (B_{ij})$ is the $n \times n$ matrix of scalar products $B_{ij} = (\vec{x}_i - \bar{x})'(\vec{x}_j - \bar{x})$ corresponding to the dissimilarities $D_{ij} = \|\vec{x}_i - \vec{x}_j\|^2$, where $\bar{x} = \sum_i f_i \vec{x}_i$.

Spatial configuration: weighted spatial network. In practice, the weight compatible exchange E matrix, specifying the spatial configuration of regions under the form of weighted spatial network, must be constructed from the given regional weights f (which may be taken as f^{text} or f^{se}) and the adjacency matrix A . That is, $E \equiv E(f, A)$, and among differing possibilities, we adopt here the *diffusive kernel construction*, which essentially consists in considering a *time-continuous Markov process* whose infinitesimal generator is given by the Laplacian of the adjacency matrix (e.g. [11, 9]). Imposing weight-compatibility $E\mathbf{1} = f$, as detailed in [2, 3, 4] yields a time-dependent exchange matrix $E(t) = E(f, A, t)$ with limits $\lim_{t \rightarrow 0} e_{ij}^{(t)} = f_i \delta_{ij}$ (reducible network made of n disconnected regions) and $\lim_{t \rightarrow \infty} e_{ij}^{(t)} = f_i f_j$ (complete weighted network, free of distance-deterrence effects).

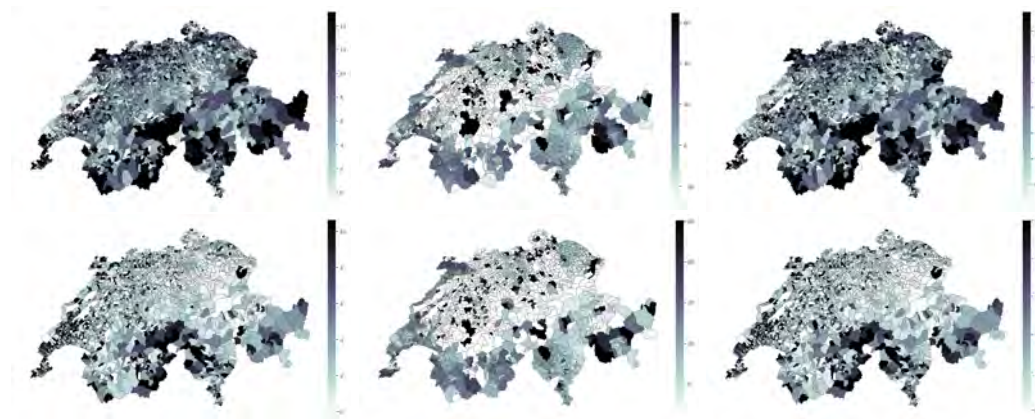
Socio-economic dissimilarities between regions. Socio-economic dissimilarities between regions can be obtained as $D_{ij} = (x_i - x_j)^2$, for numerical univariate features x , or as generalized chi-squared dissimilarities $D_{ij} = \sum_{l=1}^m \rho_l (q_{il}^\theta - q_{jl}^\theta)^2$ for categorical features with m modalities, where ρ_l is the proportion of modality l , q_{il} the ratio of observed cross-counts to their expected value under independence, and $\theta > 0$ a distortion factor overweighting for $\theta > 1$ (respectively $\theta < 1$) the contribution of high (respectively low) region-modality associations. In any case, all those dissimilarities are squared Euclidean, and so are their p -variate mixtures $D_{ij}^{\text{se}} = \sum_{k=1}^p \alpha_k D_{ij}^{(k)}$, where $D^{(k)}$ is the standardized dissimilarity for the k -variable, and $\alpha_k \geq 0$ the freely adjustable corresponding contribution, thus allowing the generation of flexible socio-economic dissimilarities adapted for particular contexts.

Textual dissimilarities between regions. Each region is described by a document, such as historical or geographical notices; or political or administrative documents; or, in our case study, Wikipedia English articles on Swiss municipalities. After usual textual preprocessing (see e.g. [10]), the resulting document-term matrix X^{text} , serves in turn to the generation of textual dissimilarities between regions :

- as straightforward chi-square dissimilarities on N , possibly generalized (see above)



■ **Figure 1** Left: logarithmic scatter plot of the weights f^{se} versus f^{text} illustrates the disparity between population and textual weights. Right: Lorenz curve associated to the Gini coefficient $G = 0.63$ between f^{se} and f^{text} .



■ **Figure 2** Local indicators of spatial autocorrelation $\delta_i(t)$ of equation (2) for the Swiss municipalities at diffusive time $t = 1$. *Top left to right*: dissimilarities are respectively D^{se} , $D^{\chi^2_{\theta=1}}$, and $(D^{\text{se}} + D^{\chi^2_{\theta=1}})/2$, with f^{se} as the reference weight. *Bottom left to right*: the resulting local indicators with the same dissimilarities and reference weight f^{text} . A large δ_i indicates strong and parallel feature deviations between municipality i and its neighbours. The notable pattern differences between top and bottom maps reveals the influence of the weight choice.

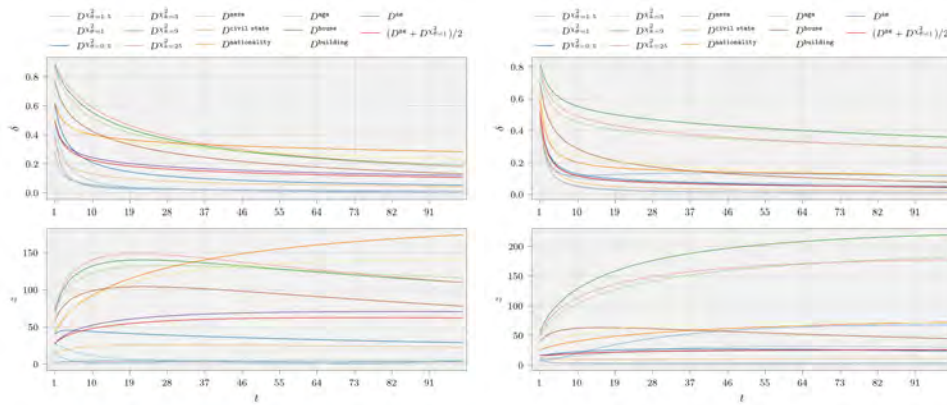
- from *topic modelling* (see e.g. [5]) on N , yielding in turn membership probabilities (of documents relatively to the topics), on which generalized "topic" chi-square dissimilarities can again be computed.

Socio-economic and textual dissimilarities can be combined as mixtures $\lambda D^{\text{se}} + (1 - \lambda) D^{\text{text}}$, where $\lambda \in (0, 1)$, which are still squared Euclidean. They can serve at implementing soft k-means clusterings detailed in [6, 7], and extended to textual content in [8].

3 Case study

We illustrate our general approach for spatial autocorrelation upon the $n = 2251$ Swiss municipalities in 2016, exploring the balance between socio-economical and textual features.

Socio-economic dissimilarities of Swiss municipalities: the $p = 6$ socio-economical features X^{se} bearing on sex, age, nationality and civil status of the permanent population (defining the socio-economic weights f^{se}), as well as the count of houses and buildings, constitute census values provided by the FSO. After standardization, their corresponding chi-squared dissimilarities contribute in equal parts to the overall socio-economic dissimilarity $D^{\text{se}} =$



■ **Figure 3** Spatial autocorrelation $\delta(t)$ of equation (1) measured for all Swiss municipalities, at diffusive times $t = 1, \dots, 99$, for various dissimilarities, with weights f^{se} proportional to the population (left) and f^{text} proportional to the number of terms (right).

$$(D^{sex} + D^{age} + D^{nationality} + D^{civil\ status} + D^{house} + D^{building})/6 .$$

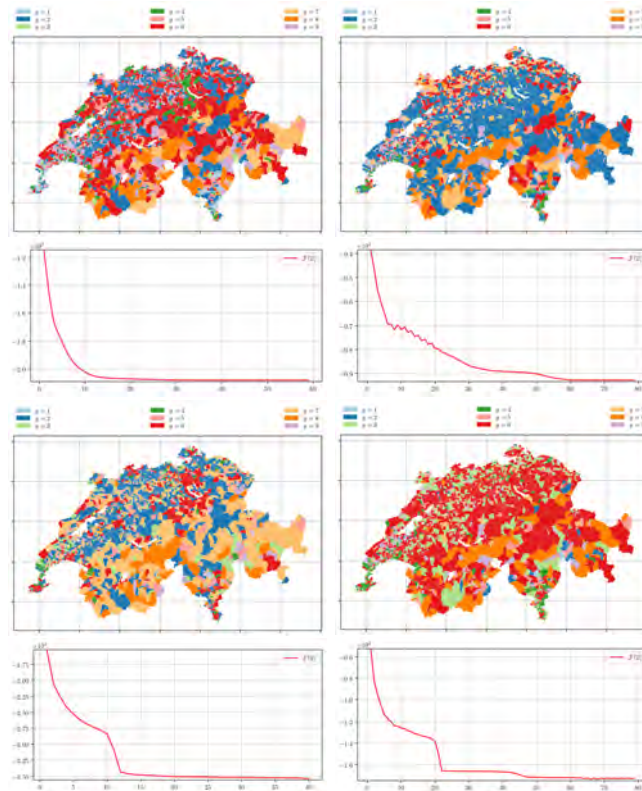
Textual dissimilarities of Swiss municipalities: for each municipality, we use the Wikipedia pages obtained through the Federal Statistical Office (FSO) number. They are further geo-referenced and textually pre-processed (see [8] for more details). Two dissimilarities will be investigated (see section 2): $D^{\chi^2_k}$, resulting from topic modelling with $k = 3, 9, 25$ topics, and $D^{\chi^2_\theta}$, the generalized chi-squared dissimilarity on the original document-term matrix.

Combination of Socio-economic dissimilarities and Textual dissimilarities: the autocorrelation index $\delta(t)$ and its standardized value $z(t)$ are depicted in figure 2 for differing diffusion times $t > 0$, after preliminary choice of the weights f , well contrasted (figure 1), and whose large influence on the analysis is apparent.

Figure 2 depicts the disparate values of the local indicators $\delta_i(t = 1)$, whose range is much larger for the chi2 textual document-term dissimilarities under socio-economic weights, and whose values can be negative, indicating a strong spatial contrast yet to be fully understood. Finally, figure 3 depicts the contrasted behavior of $\delta(t)$ and $z(t)$ for various diffusion times, various dissimilarity choices, and for the two set of weights. Although differing by order of magnitudes, the associated spatial autocorrelations are always significant at level 5% (that is $|z(t)| > u_{.95} = 1.96$), with the exception of D^{sex} and D^{age} for $f = f^{text}$, and $D^{\chi^2_{\theta=1}}$ for $f = f^{se}$, which loose their significance for t large.

The “spatial+feature” clustering. The “spatial+feature” clustering method introduced in [6, 7] and extended to textual content in [8] attempts to create clusters containing nodes both strongly connected (as in network clustering) and similar regarding their features (as in distance-based clustering), and does so by running an iterative procedure, decreasing at each step the *free energy* $F[Z]$ (a generalized negative log-likelihood) of the *soft membership matrix* $Z = (z_{ig})$, given the probability that region i belongs to group $g = 1, \dots, m$. Starting from an initial membership Z^0 , the iteration converges to a final membership Z^∞ , which constitutes a local minimum of the free energy, and constitutes a generalized soft k-means procedure (spherical Gaussian mixtures) taking into account the spatial configuration of the objects to be clustered.

Figure 4 depicts the final clustering, made hard by assigning each region i to group $G[i] = \arg \max_{g \in \{1, \dots, m\}} z_{ig}^\infty$, with $m = 9$ groups. In all four cases, the initial membership Z^0



■ **Figure 4** Hard assignment of the final soft attribution Z^∞ for all Swiss municipalities at diffusive time $t = 1$, for $m = 9$ groups, and decrease of the free energy. Left: socio-economic dissimilarities D^{se} with parameters $\beta = 8, \alpha = 0.1$ (see [7, 8]), and weights f^{se} (top) and f^{text} (bottom). Right: mixed dissimilarities $(D^{\text{se}} + D^{\chi_{\theta=1}^2})/2$ with parameters $\beta = 1.4, \alpha = 0.1$, and weights f^{se} (top) and f^{text} (bottom).

consists of an official attribution of the $n = 2251$ Swiss municipalities in $m = 9$ urban-rural categories, provided by FSO, and updated in 2017 [12].

In guise of conclusion. As illustrated by the case study, the proposed formalism sets up a general methodology able to incorporate directly textual content in the characterization of regions, on equal footing with more usual geographical information such as socio-economic features. A crucial step is the systematic use of squared Euclidean dissimilarities, which can be freely linearly combined. The regional weights can also be chosen as reflecting the population or area regional importance; or, more originally, the regional textual importance – a choice better adapted for e.g. destination image and impressions in tourism studies.

References

- 1 Luc Anselin. Local indicators of spatial association - LISA. *Geographical analysis*, 27(2):93–115, 1995.
- 2 François Bavaud. Testing spatial autocorrelation in weighted networks: the modes permutation test. *Journal of Geographical Systems*, 3(15):233–247, 2013.

- 3 François Bavaud. Spatial weights: Constructing weight-compatible exchange matrices from proximity matrices. In M. et al. Duckham, editor, *Geographic Information Science*, pages 81–96, Cham, 2014. Springer.
- 4 François Bavaud, Maryam Kordi, and Christian Kaiser. Flow autocorrelation: a dyadic approach. *Springer Nature 2018*, 2018.
- 5 David M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.
- 6 Raphaël Ceré and François Bavaud. Multi-labelled Image Segmentation in Irregular, Weighted Networks: A Spatial Autocorrelation Approach. In *GISTAM 2017 - Proceedings of the 3rd International Conference on Geographical Information Systems Theory, Applications and Management*, volume 1, pages 62–69, 2017.
- 7 Raphaël Ceré and François Bavaud. Soft image segmentation: on the clustering of irregular, weighted, multivariate marked networks. Accepted for Springer Book of GISTAM 2017: Communications in Computer and Information Science CCIS series, 2018.
- 8 Mattia Egloff and Raphael Ceré. Soft Textual Cartography Based on Topic Modeling and Clustering of Irregular, Multivariate Marked Networks. In C et al. Cherifi, editor, *Complex Networks & Their Applications VI*, pages 731–743. Springer, 2018.
- 9 François Fouss, Marco Saerens, and Masashi Shimbo. *Algorithms and models for network data and link analysis*. Cambridge University Press, 2016.
- 10 Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.
- 11 Alexander J. Smola and Risi Kondor. Kernels and regularization on graphs. In *COLT*, volume 2777, pages 144–158. Springer, 2003.
- 12 Laurent Zecha, Florian Kohler, and Viktor Goebel. Niveaux géographiques de la Suisse. Typologie des communes et typologie urbain-rural 2012. Technical report, Office fédéral de la statistique (OFS), 2017.

Evaluating Efficiency of Spatial Analysis in Cloud Computing Platforms

Changlock Choi

Department of Geography, Kyung Hee University, Seoul, South Korea
hihi7100@khu.ac.kr

Yelin Kim


Department of Geography, Kyung Hee University, Seoul, South Korea
yelin910@khu.ac.kr

Youngho Lee

Department of Geography, Kyung Hee University, Seoul, South Korea
emfo0124@khu.ac.kr

Seong-Yun Hong

Department of Geography, Kyung Hee University, Seoul, South Korea
syhong@khu.ac.kr

 <https://orcid.org/0000-0001-5049-8810>

Abstract

The increase of high-resolution spatial data and methodological developments in recent years has enabled a detailed analysis of individuals' experience in space and over time. However, despite the increasing availability of data and technological advances, such individual-level analysis is not always possible in practice because of its computing requirements. To overcome this limitation, there has been a considerable amount of research on the use of high-performance, public cloud computing platforms for spatial analysis and simulation. In this paper, we aim to evaluate the efficiency of spatial analysis in cloud computing platforms. We compared the computing speed for calculating the Moran's I index between a local machine and spot instances on clouds, and our results demonstrated that there could be significant improvements in terms of computing time when the analysis was performed parallel on clouds.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases spatial analysis, parallel computing, cloud services

Digital Object Identifier 10.4230/LIPICs.GIScience.2018.24

Category Short Paper

Funding This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP; Ministry of Science, ICT & Future Planning) (No. 2017R1C1B5015090).

1 Introduction

The widespread use of social media and location-based services has produced a large amount of geospatial data [4]. Much of these data are made up of point data, such as OpenStreetMap's PoI and geotagged Twitter posts. Therefore, spatial analysis on point-based data is also widely used for practical and scientific purposes. Point data that involve millions of points are becoming common, and they cause a problem of storage space and memory shortage



© Changlock Choi, Yelin Kim, Youngho Lee, and Seong-Yun Hong;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 24; pp. 24:1–24:5

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

■ **Table 1** Test environments for the Moran’s I index.

Environment	Processor	Number of cores	RAM
Local machine	3.4GHz	3	6GB
Spot instance (t2xlarge)	2.3GHz–2.4GHz	4	16GB
Spot instance (m44xlarge)	2.3GHz–2.4GHz	16	64GB

of the computer due to the data scale. There is certainly a need for a new approach for analysing big geospatial data that are difficult to be handled by existing techniques [5].

Cloud computing is one of the alternatives for the analysis of big geospatial data. Many attempts have been made to solve the problem using cloud computing platforms, as it can provide a better analysis environment in terms of cost effectiveness, stability, and computing efficiency. The use of cloud services for spatial analysis is cost effective, because it allows users to lease hardware resources only when they are required. It can be more stable than running own high-performance servers because the cloud computing service providers maintain and manage the facilities. In this short paper, our purpose is to confirm the efficiency of spatial data analysis in cloud computing platforms. To achieve this goal, we compare the time taken for calculating the Moran’s I index on a local machine with those on virtual machines (or spot instances). We use the statistical software R for the experiments, but due to the fact that R utilises only one of the machines’ cores for its computation, the existing functions are adjusted to make the calculation parallel.

2 Background

2.1 Cloud computing with R

Cloud computing is a term that encompasses the hardware and system software in the data center that provide applications and services that are delivered as services over the Internet. These services have long been called Software as a Service (SaaS), and data centers, hardware and software are what we call the cloud. With the advent of cloud computing, developers are free to increase capital expenditures and operational costs, and use unprecedented, low-cost, resilient resources to deliver services [1].

Amazon Elastic Compute Cloud (EC2) is part of Amazon Web Services (AWS) and provides virtual computing environments called *instances*. There are many different types of instances available in EC2, each of which has a different combination of CPU, memory, storage, and networking capacity. Users can choose an instance based on their purpose—general purpose, computing optimisation, memory optimisation, accelerated computing, and storage optimisation. The use of such cloud computing platforms can be more cost effective than constructing a physical computing environment.

There are, however, limitations in using R on cloud services. Most cloud service providers increase the computing performance of spot instances by increasing the number of cores. However, since R can use only one core by default, the increasing number of cores on instances does not affect the computing performance of spatial analysis. To illustrate this point, we selected two spot instances from EC2 and compared the computing time between the instances and between a local machine and them (Table 1).

■ **Table 2** Computing time for single-core and multi-core environments.

Environment	Single-core (in seconds)	Multi-core (in seconds)
Local machine	26342.72	12655.43
Spot instance (t2xlarge)	37503.21	11889.89
Spot instance (m44xlarge)	36976.54	6458.94

2.2 Parallel computing with R

There have been many attempts to solve the problems caused by the size of spatial data in geography using the cloud platforms and parallel computing. Parallel computing is a technique for decomposing and concurrently manipulating data, or concurrently executing process components to complete a task [7]. A common method of parallel computing is to decompose a data set into smaller units, distribute it to multiple operators, and then collect and reconstruct the results after analysis [2].

In this work, we calculate the Moran's I index using Monte Carlo simulations, and each trial runs independently. This can be considered an application of *embarrassingly-parallel*, which means no interactions or communications exist between the operations during the parallel computing process [3]. We have modified the existing Moran's I function in R using the `parallel` package to enable this sort of parallel computing, and use it in each of the described computing environments to compare the efficiency.

3 Methods and results

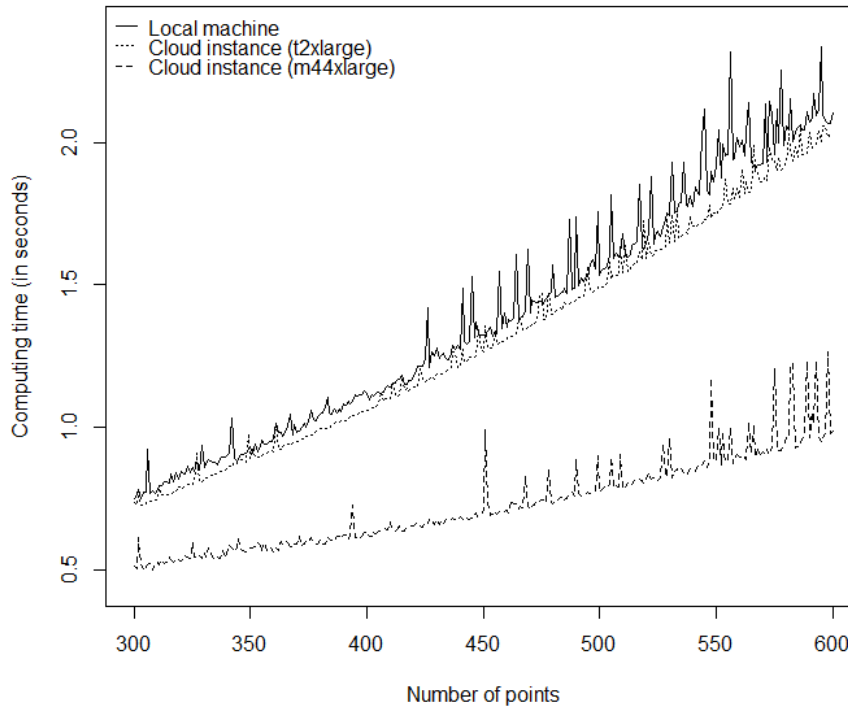
This paper uses Moran's I to compare the computational efficiency of spatial analysis on cloud services. Moran's I is an index for describing spatial autocorrelation of point patterns [6]. Theoretically, the range of Moran's I is from -1 to 1, with positive autocorrelation closer to 1, and negative autocorrelation closer to -1.

The Moran's I index is calculated for hypothetical data that contain 300–500 sets of coordinates and values. The coordinates and the values were generated from a uniform distribution, and the number of repetitions in the Monte Carlo simulation was set to 1,000. Each simulation was repeated 30 times. Table 2 presents the total computing time in each environment, and Figure 1 shows how the average computing time changes with the number of points (i.e., data size).

As shown in Table 2, the local machine took slightly over 26,000 seconds, while both spot instances, **t2xlarge** and **m44xlarge** took about 37,000 seconds. In addition, **m44xlarge** shows about four times more computational efficiency than **t2xlarge** in terms of catalog performance. These results seem to be derived from the performance enhancements of single-core and cloud computing services—a feature of R mentioned above.

On the other hand, in the case of parallel computing, it was confirmed that the computing time using parallel computing is less than that of the local machine. Also, as the size of data increases, the gap tends to increase more and more. However, when comparing **t2xlarge** and **m44xlarge**, there is less difference compared to actual performance difference. This is probably a problem of the parallel computing process. Parallel computing, when compared to a single-core computing, requires at least two additional processes, distribution of data and aggregation of results, and this might cause the difference in time.

Table 3 shows the minimum, mean, and maximum values for each environment, and it indicates a similar conclusion to that from Figure 1. When comparing the mean values, the



■ **Figure 1** Computing time by the number of points.

`t2xlarge` instance does not show a significant difference in the computing time with the local machine, but the `m44xlarge` instance has shortened the time from 1.5 to 2 times for the same number of points. However, when comparing the maximum values, the time taken for analysis fluctuates, possibly due to the instability of the system. When the calculation is repeatedly performed, the differences between the mean and the maximum values become clearly apparent.

4 Conclusions

As we have demonstrated, the use of single-core programs for big data analysis is limited, because it takes a considerable amount of time to operate or does not properly reflect the evolving computing environment. In particular, spatial analysis using spatial data requires a new approach, because the number of data increases and the computing resources required for analysis increase exponentially. Therefore, the need for high-performance computing technology that is capable of rapidly computing and processing large-scale data has begun to be emphasised.

This paper attempts to verify cloud computing as an alternative method to solve the above problems from the empirical point of view. Cloud computing platforms provide a better analysis environment in three ways. First, it is more economical to lease the hardware of the desired performance at the user's desired time through cloud computing than to build the high-performance resource at the initial cost. In general, users are tempted to perform high-performance analysis because their computing resources are time-sensitive and their replacement cycle is short. Second, cloud services meet the stability of analytics in the sense that the service providers take the responsibility for maintaining and servicing data. Finally,

■ **Table 3** Computing time by the number of points.

Environment		Number of points			
		300	400	500	600
Local machine	Max	0.94819	0.13022	0.16920	2.27118
	Mean	0.74900	1.11907	1.53280	2.10250
	Min	0.65706	0.92256	1.37751	1.86864
Spot instance (t2xlarge)	Max	0.73618	1.08571	1.51585	2.04415
	Mean	0.72592	1.05769	1.49446	2.02816
	Min	0.70898	1.04644	1.47618	2.01196
Spot instance (m44xlarge)	Max	0.61105	0.74626	0.92479	1.36301
	Mean	0.50443	0.63304	0.77270	0.98782
	Min	0.48263	0.60369	0.73902	0.94593

multi-core analysis on cloud computing platforms ensures the efficiency of analysis. In this paper, we demonstrated that the time for calculating Moran's I can be significantly improved (i.e., reduced) when parallel computing is used on cloud services.


In this study, a parallel processing structure of SIMD (Single Instruction Stream) method is used for the calculation. This means that the same operation is simultaneously performed on the data set assigned to each operator. When using multiple instruction streams (MIMD), different operations can be performed simultaneously on an allocated data set, resulting in more efficiency in parallel computing. In the future, it will be possible to verify the most effective approach to spatial analysis when various parallel processing structures are used.

References

- 1 Michael Armbrust, Armando Fox, Rean Griffith, Anthony D Joseph, Randy H Katz, Andrew Konwinski, Gunho Lee, David A Patterson, Ariel Rabkin, Ion Stoica, and Matei Zaharia. Above the clouds: A Berkeley view of cloud computing. Technical report, Electrical Engineering and Computer Sciences, University of California, Berkeley, 2009.
- 2 Yuemin Ding and Paul J Densham. Spatial strategies for parallel spatial modelling. *International Journal of Geographical Information Systems*, 10(6):669–698, 1996.
- 3 Ian Foster. *Designing and building parallel programs*, volume 78. Addison Wesley Publishing Company Boston, 1995.
- 4 Michael F Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221, 2007.
- 5 Rob Kitchin. Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography*, 3(3):262–267, 2013.
- 6 Patrick A. P. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950.
- 7 Michael J Quinn. *Designing efficient algorithms for parallel computers*. McGraw-Hill, 1987.

Towards the Usefulness of User-Generated Content to Understand Traffic Events

Rahul Deb Das¹

Department of Geography, University of Zurich
Winterthurerstrasse 190, 8057 Zurich, Switzerland
rahul.das@geo.uzh.ch
 <https://orcid.org/0000-0002-3379-3516>

Ross S. Purves

Department of Geography, University of Zurich
Winterthurerstrasse 190, 8057 Zurich, Switzerland
ross.purves@geo.uzh.ch

Abstract

This paper explores the usefulness of Twitter data to detect traffic events and their geographical locations in India through machine learning and NLP. We develop a classification module that can identify tweets relevant for traffic authorities with 0.80 recall accuracy using a Naive Bayes classifier. The proposed model also handles vernacular geographical aspects while retrieving place information from unstructured texts using a multi-layered georeferencing module. This work shows Mumbai has a wide spread use of Twitter for traffic information dissemination with substantial geographical information contributed by the users.

2012 ACM Subject Classification Information systems → Geographic information systems, Information systems → Information retrieval, Computing methodologies → Natural language processing, Computing methodologies → Artificial intelligence, Human-centered computing → Ubiquitous and mobile computing

Keywords and phrases Urban mobility, traffic, UGC, tweet, event, GIR, geoparsing

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.25

Category Short Paper

Funding We are grateful to the Swiss National Science Foundation (SNSF) grant number 166788.

Acknowledgements We would like to thank Alan MacEachren, Natalia Andrienko, Gennady Andrienko, Liao Din, Martin Schorcht, Florian Lautenschlager, and Stefan Kasberger for their valuable comments during the initial stage of this project in VGIScience Summer School'17 in Dresden.

1 Introduction

Retrieving geographical information pertaining to events is important for planning and decision making processes, for instance in identifying locations that demand special attention. With the emergence of user-generated content (UGC), it is now possible to detect various urban events and their geographical locations more ubiquitously. Events may be related to, for example, urban mobility [6], natural disasters [13, 3] or environmental conditions [17].

¹ Corresponding author



UGC derived from social-media platforms are often unstructured and pose challenges if we are to relate vague and ambiguous references in natural language to specific locations [7]. This paper introduces a framework to deal with such challenges while detecting traffic events for managing urban resources and transportation infrastructure.

Currently traffic information is collected through static, and physical sensors e.g., loop detectors or CCTV cameras installed at different locations in a city [8]. Since these sensors are static, they provide limited spatial coverage and come with high installation and maintenance costs. In order to address these issues, this paper leverages the concept of *citizens as sensors* [5] where the citizens contribute information (in)voluntarily, which can be used to characterize traffic events.

We use Twitter to both analyze traffic in real time and gain insights into patterns over time. Our contributions are as follows.

- Unlike previous works [6, 9] we leverage untagged tweets to extract the locations of traffic events through text analysis.
- We assess the usefulness of UGC (e.g., Twitter) to detect traffic events in India where many of the metro cities are highly congested [15] with limited physical traffic infrastructure.
- We develop a hybrid multi-layered geoparser that can retrieve traffic event locations from unstructured texts tweeted in India where place names are often mentioned in local languages.

In Section 2 we briefly review the state of the art. Section 3 and 4 explain the framework and its evaluation, before Sections 5 and 6 discuss some limitation of our approach and propose directions for future work.

2 Related work

Twitter is a ubiquitous UGC source where people post information, reactions and opinions about a vast array of topics [2]. In the past Twitter has been used to detect traffic events, however, mostly identifying traffic related information from geotagged tweets. For example, D'Andrea and colleagues developed a model that could detect traffic related tweets in real time in Italy using Support Vector Machine (SVM) with an accuracy of 95.75% [1]. They used a balanced data set with 665 instances each in training and testing. Kurniawan and colleagues developed a real-time tweet classification model using geotagged tweets in order to provide traffic related information in Indonesia [9]. Similarly Salas and others developed a SVM based supervised model to detect incidents in London using Twitter data [14]. They used a balanced data set for training and testing. In these papers traffic events and their locations are assumed to be the location in the tweet metadata.

Wanichayapong and colleagues developed a model to classify tweets as traffic or non-traffic related through syntactic analysis in Thailand. They also classified traffic related tweets in point and link category depending on the location of the traffic events [16]. They achieved 76.85% accuracy for point category and 93.23% accuracy for link category. Gu and others presented a real-time traffic incident detection model which was evaluated in Philadelphia and Pittsburgh in the USA. They developed the model based on a semi-Naive Bayes classifier and achieved 90.5% accuracy [6]. Since most of tweets are not explicitly georeferenced, various models have been proposed to extract locations from tweet content and metadata [4]. For example, Gelernter and Balaji proposed a hybrid model to georeference tweets in New Zealand [4]. In a slightly different work Pereira and others used text analysis to predict incident durations from the authoritative structured text [12].

In this work we propose a model that goes beyond existing traffic detection models that leverage geotagged tweets. Instead, we use untagged tweets to understand traffic conditions through a hybrid multi-layered geoparser (c.f [4]) by applying a mixture of spatial rules and localized spatial references evaluated in Indian context.

3 Methodology

We propose a hierarchical model that can detect tweets relevant to traffic and then extract spatial information from the tweet to provide more information about the traffic event. The methodology is divided into three stages.

3.1 Data collection

To evaluate the model a data set has been collected in Mumbai ² using a keyword *traffic* from 1st January to 28th February, 2017. Manual annotation was performed to label whether a tweet was related to traffic. Since all the tweets contain the keyword *traffic*, a number of criteria were set during the annotation process. A tweet is labelled as a relevant tweet if it contains information about a traffic event along with either a spatial reference (*where*) or a temporal reference (*when*) or a cause (*why*). Through this process, 2614 tweets were annotated over two months in Mumbai where the count of traffic related tweets was 755. Another manual annotation was performed to extract the place names mentioned in the traffic related tweets to use them as ground truth to evaluate the performance of the georeferencing module (c.f Section 3.3). A tweet may have more than one place name. In that case all the unique place names are annotated.

3.2 Tweet classification

After preprocessing (to eliminate emoticons and non-ASCII characters and trim white space from tweet content) three different classifiers were tested. Those were rule-based (PART), tree-based (Decision Tree (DT)), and a probabilistic classifier (Naive Bayes (NB)).

To create the features to train the classifiers the tweet text is converted to a numerical form where each word is assigned a weight based on its term frequency-inverse document frequency (tf-idf) as follows.

$$tf = T_t \quad (1)$$

$$idf = \log[N/(1 + D_t)] \quad (2)$$

$$tf - idf = tf * idf \quad (3)$$

Where T_t is the total count of term 't' in tweet 'D'. 'N' is the total number of tweets in the corpus and D_t is the total number of tweets containing the term 't'.

² <https://www.numbeo.com/traffic/rankings.jsp>, last accessed April, 2018

3.3 Georeferencing module

In the third stage a 3-tier tweet georeferencing module (GM) was developed that can retrieve geographical information from the traffic relevant tweets. Initially a pre-trained supervised geoparser e.g., StanfordNLP [11] was used (1st tier). However, due to lack of training on the local data set (as in Mumbai) two more rule-based layers have been implemented. In the first rule-based layer (2nd tier) if a token is a proper noun (NNP) or a common noun (NN) and if it is preceded by a spatial preposition then the token is deemed to be a place name. We used 17 spatial prepositions e.g., *towards, from, to, at, on, near*. The second rule-based layer (3rd tier) considers vernacular place names in Mumbai (e.g., *naka: toll plaza, marg: road, bhavan: building, chowpatty: fishermen's colony*) and various spatial object types in English (e.g., *building, park, flyover*). Any NNP or NN token that is followed by one of these vernacular names or an object type is deemed to be a place name. We extracted 84 vernacular names and object types.

Once the spatial references are retrieved place names are resolved by assigning coordinates using OpenStreetmap (OSM). To disambiguate place names *Maharashtra* (the local state name) was used as a spatial context (c.f [10]).

4 Evaluation and results

4.1 Detecting traffic related tweets

The models are evaluated using 3-fold cross validation. While detecting traffic related tweets, a NB classifier performs best with 0.80 recall and 0.52 precision, while a rule-based model (PART) yields 0.67 precision and 0.57 recall and the DT model gives precision 0.65 and recall 0.57. For non-traffic tweets a NB classifier yields 0.89 precision whereas a PART and DT yield 0.88 and 0.87 recall respectively.

4.2 Performance of tweet georeferencing module

As the texts in tweets are often unstructured – involve abbreviation and typos, while measuring the accuracy of georeferencing module, first a complete match was performed. If the retrieved place name in tweet_k does not completely match with any of the annotated place names in the same tweet, then a fuzzy matching was performed. If the cosine similarity (CoSim) between the retrieved place name and the annotated place name is greater than a threshold (0.4) then the retrieved place name is considered as a true positive.

When using the StanfordNLP alone without the second and third tiers (rule base) over two months of traffic related tweets, the georeferencing module yields precision of 0.60 and recall 0.34. However, using all the tiers the georeferencing module yields 0.71 precision and 0.61 recall. Using all the three tiers total 451 places are resolved from the retrieved place names out of 767 resolved places from the annotated ones (Fig 1). As can be seen the proposed model effectively retrieves locations that are subject to traffic events with 58.88% places being successfully resolved (Fig 1).

5 Discussion

Currently the model uses tweets that contain only the keywords *traffic*, but in future more keywords will be incorporated. The georeferencing module presented in this work sometimes fails to detect place names consisting of two tokens followed by a vernacular name or object type. For example, *teen Haath Naka* has been recognized as *Haath Naka*, which is detected

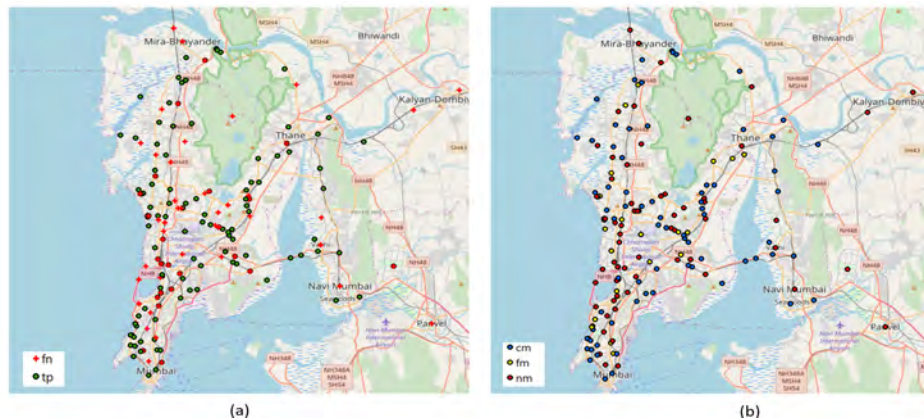


Figure 1 (a) Locations of the traffic events resolved from correctly retrieved tweets i.e true positives (tp) and location of annotated events that could not be retrieved i.e false negatives (fn); (b) Retrieved locations of the traffic events that completely match with the annotated place names (cm), fuzzy match (fm) and annotated places that could not be retrieved (nm) by the model.

Table 1 Spatial granularity in the text.

Tweet	Place Type	Geometry Type
@MumbaiPolice heavy traffic at bkc near income tax office...	Building (income tax office)	Point
Traffic movement on S V Road at Andheri and Jogeshwari is lot better today.	Road name (S V Road), Region (Andheri, Jogeshwari)	Polyline, Polygon

through fuzzy matching. We also observed that people use geographical information at different granularities while tweeting about traffic events (Table 1).

Similar to the earlier works while classifying the tweets, a *k*-fold cross validation has been used to evaluate the tweet classification model over two months data. It has been observed while tweeting people in Mumbai react in two ways, either they report or share traffic events or they request respective authority (e.g., @MumbaiPolice) to resolve a traffic issue. An extension of this work will investigate if the model performs equally well on a data set collected separately in a different time period.

Although in this research a small number of tweets have been analyzed based on only a single keyword, the approach is scalable and adaptive to more traffic related keywords and more tweets. In terms of the size of the data set past studies have also showed promising results with small data sets [1, 14]. The main focus of this paper was on detecting traffic relevant tweets and their respective locations. However, it is also important to identify the reasons behind the traffic events, which requires more complex syntactic and semantic analysis of the text.

6 Conclusions

In this paper a traffic event detection model has been introduced. The model can be useful both in real-time as well as in historical manner and can detect tweets relevant to traffic authorities, urban planners and daily commuters to understand the traffic events and their geographical locations both for short-term and long-term planning. In this research we showed Twitter has potential for detecting traffic events in Indian cities if we build a georeferencing model capable of dealing with unstructured, vague and vernacular text in natural language.

An important limitation of our work is that India is a multi-lingual country and our analysis focused on English. Nonetheless, vernacular terms are often used while communicating about an event with a spatial reference in English. Here the implemented multi-layered geoparser shows its effectiveness in resolving 58.88% of local places that are subject to have traffic events, which was not possible using a pre-trained NER due to lack of local traffic related corpora.

Future work will consider tweets with more traffic related keywords and explore temporal patterns of tweeting behavior reacting to traffic events. Although the study is performed in India, but the same approach can be useful to other places.

References

- 1 E. D. Andrea, P. Ducange, B. Lazzerini, and F. Marcelloni. Real-time detection of traffic from twitter stream analysis. *IEEE Transactions on Intelligent Transportation Systems*, 16(4):2269–2283, 2015.
- 2 Farzindar Atefeh and Wael Khreich. A survey of techniques for event detection in twitter. *Computational Intelligence*, 31(1):132–164, 2015.
- 3 Andrew Crooks, Arie Croitoru, Anthony Stefanidis, and Jacek Radzikowski. #Earthquake: Twitter as a distributed sensor system. *Transactions in GIS*, 17(1):124–147, 2013.
- 4 Judith Gelernter and Shilpa Balaji. An algorithm for local geoparsing of microtext. *GeoInformatica*, 17(4):635–667, 2013.
- 5 Michael F. Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221, 2007.
- 6 Yiming Gu, Zhen Qian, and Feng Chen. From twitter to detector: Real-time traffic incident detection using social media data. *Transportation Research Part C: Emerging Technologies*, 67:321–342, 2016.
- 7 Livia Hollenstein and Ross S Purves. Exploring place through user-generated content: Using flickr to describe city cores. *Journal of Spatial Information Science*, 1(1):21–48, 2010.
- 8 Akira Kinoshita, Atsuhiko Takasu, and Jun Adachi. Real-time traffic incident detection using a probabilistic topic model. *Information Systems*, 54:169–188, 2015.
- 9 D. A. Kurniawan, S. Wibirama, and N. A. Setiawan. Real-time traffic classification with twitter data mining. In *8th International Conference on Information Technology and Electrical Engineering (ICITEE)*, pages 1–5, Yogyakarta, Indonesia, 2016.
- 10 Jochen L. Leidner, Gail Sinclair, and Bonnie Webber. Grounding spatial named entities for information extraction and question answering. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references - Volume 1*, Stroudsburg, USA, 2003.
- 11 Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Baltimore, Maryland, USA, 2014.
- 12 Francisco C. Pereira, Filipe Rodrigues, and Moshe Ben-Akiva. Text analysis in incident duration prediction. *Transportation Research Part C: Emerging Technologies*, 37:177–192, 2013.
- 13 Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World Wide Web*, pages 851–860, Raleigh, North Carolina, USA, 2010. ACM.
- 14 A. Salas, P. Georgakis, and Y. Petalas. Incident detection using data from social media. In *IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 751–755, Yokohama, Japan, 2017.

- 15 Azeem Uddin. Traffic congestion in indian cities: Challenges of a rising power, draft. Report, General Motors India, 2009.
- 16 Napong Wanichayapong, Wasawat Pruthipunyaskul, Wasan Pattara-Atikom, and Pimwadee Chaovalit. Social-based traffic information extraction and classification. In *IEEE 11th International Conference on ITS Telecommunications*, St. Petersburg, Russia, 2011.
- 17 Yuchao Zhou, Suparna De, and Klaus Moessner. Real world city event extraction from twitter data streams. *Procedia Computer Science*, 98:443–448, 2016.

Unfolding Urban Structures: Towards Route Prediction and Automated City Modeling

Paolo Fogliaroni

Vienna University of Technology, Austria
paolo.fogliaroni@geo.tuwien.ac.at

Marvin Mc Cutchan

Vienna University of Technology, Austria
marvin.mccutchan@geo.tuwien.ac.at

Gerhard Navratil

Vienna University of Technology, Austria
gerhard.navratil@geo.tuwien.ac.at

Ioannis Giannopoulos

Vienna University of Technology, Austria
igiannopoulos@geo.tuwien.ac.at

Abstract

This paper extends previous work concerning intersection classification by including a new set of statistics that enable to describe the structure of a city at a higher level of detail. Namely, we suggest to analyze sequences of intersections of different types. We start with sequences of length two and present a probabilistic model to derive statistics for longer sequences. We validate the results by comparing them with real frequencies. Finally, we discuss how this work can contribute to the generation of virtual cities as well as to spatial configuration search.

2012 ACM Subject Classification Information systems → Geographic information systems, Information systems → Data analytics, Information systems → Probabilistic retrieval models

Keywords and phrases intersection types, spatial structure, spatial modeling, graph theory

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.26

Category Short Paper

1 Introduction

Modeling the structure of a city is a long-term goal in the GIScience community, as well as in other communities such as Urban Planning, Transportation Planning, Civil Engineering, and Spatial Cognition. Indeed, developing a formal model for describing the structure of a city can be beneficial for a variety of scenarios. For example, to look for structurally similar areas in different cities or different areas of the same city, to generate virtual look-alike cities (i.e., virtual environments exposing a similar structure to a reference city), or to (re)design a street network to minimize the probability of traffic congestion.

The structure of a city can be regarded as consisting of topological and metrical information. In this work we introduce an approach for capturing a topological aspect of the structure that will be complemented in future work with distance and directional information to obtain a complete structural representation.

In [8] a novel approach was introduced that approaches the problem by analyzing the intersections making up the street network of a city. The paper presents a classification for



© Paolo Fogliaroni, Marvin Mc Cutchan, Gerhard Navratil, and Ioannis Giannopoulos;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 26; pp. 26:1–26:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

intersections and introduces and formally defines so-called regular intersections that are used as a baseline for comparing real intersections. The introduced model is utilized to derive statistics about and compare four cities as well different districts of the same city.

We extend the previous work by introducing a novel metric for the representation of urban structures. Building on top on the intersection data provided in [8] we draft a model to predict sequences of consecutive intersections. We start by counting the occurrences of sequences of two intersections and we present a probabilistic model to infer the frequencies of longer sequences. We validate the model by comparing the inferred frequencies with real data for sequences of 3 and 4 intersections.

Finally, we envision how this model can be used in future work to automatically generate virtual look-alike environments that expose a similar structure with respect to a reference city and to find structurally similar areas in different cities or in different parts of the same city.

2 Related Work

While our work takes on different disciplines such as spatial cognition, network analysis, graph theory, and space syntax, at the best of our knowledge this is a novel approach to the problem of understanding the structure of a city.

Probably the most famous work about the analysis of urban spaces is the work from Lynch [12]. In this work, Lynch analyzes properties of cities that affect the perceptual and cognitive aspects. He argues that the environmental image of a city consists of three main components: identity, structure, and meaning. The structure of a city is described as the spatial relations occurring among the city objects as well as between those and the observer.

Spatial networks (see [1] for a detailed survey) are spatially embedded graphs representing spatial features and connections among them. A typical example of spatial networks are street networks where intersections are reported as nodes of the graph and street segments as its edges. In [14] an open source toolbox for ArcGIS is introduced that allows for computing five types of network centrality measures on spatial networks: reach, gravity index, betweenness, closeness, and straightness (see [14] for details about these metrics).

Graph theory provides the mathematical foundation to topological analysis (see [4] for an extensive discussion on the topic). Network analysis and other spatial studies typically model the domain of investigation by means of a graph or a hypergraph and employ typical graph properties (e.g., node degree, and reachability) and operations (e.g. shortest path, connected components) to perform the necessary analyses.

Finally, space syntax [11, 10] is a set of theories aiming at identifying how urban structure affects social structure. Theories of space syntax typically represent the urban space as a graph, using different abstractions for nodes and edges. In simple terms, space syntax approaches model spatial environments with a dual graph where nodes represent empty space (e.g., streets in a street network) and edges represents some sort of connection among them (e.g., intersections). One of the earliest approaches [11] to space syntax is based on the concepts of axial line and convex space. However, it has been argued [2] that the lack of formality in the original definition of these concepts does not allow for an automatic generation of a so-called axial map. Another popular approach to space syntax resorts to the concept of isovist: the set of all points visible from a given vantage point in space and with respect to an environment [3].

2.1 Types of Intersections

In [8] an intersection is classified according to two main metrics. First, the number n of street segments stemming out of an intersections (i.e., its branches). An intersection with n branches is called an n -way intersection.

Second, the angular arrangement of the branches of an n -way intersection I^n . This is described as the angular distance $\Delta(I^n, R^n)$ between I^n and the corresponding *regular* n -way intersection R^n : an intersection whose branches split a revolution (2π) into n equal angles, each of width $\frac{2\pi}{n}$. $\Delta(I^n, R^n)$ is the minimum sum of angles that we have to rotate the branches of I^n to perfectly match R^n , while preserving the circular order of I^n 's branches.

Finally, intersections are classified according to the type of transportation mode they allow. *Path*-intersections allow only for pedestrian transit; *road*-intersections are passable by both pedestrians and cars; *car*-intersections only allow cars.

3 Predicting Route Sequences

3.1 Modeling

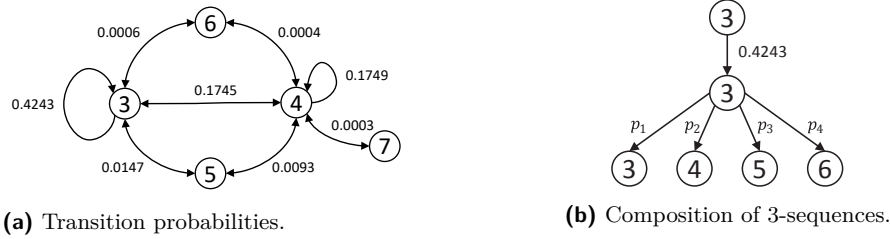
In [8] statistics about the intersections of a city have been derived: the type of intersections (3-way, 4-way, ...) and their angular distance to the corresponding regular intersection. Assume to represent the street network of a city as a graph $G = (V, E)$ whose nodes V represent intersections and whose edges E represent street segments among them. Then, the type of an intersection denotes the degree of a node and the corresponding statistics provide a first approximate description of the graph and, thus, of the city structure.

In order to fully characterize the city structure we shall compute more information. In this paper we focus only on topological information. More specifically, we focus on the prediction of intersection sequences as we route through the street network.

Say $\pi = \langle t_1, \dots, t_n \rangle$ is the shortest path between two nodes in the graph, where t_i denotes the type of node that is traversed – i.e., its degree or branching factor – and (t_i, t_{i+1}) is an edge of the graph – i.e., a street segment connecting two consecutive intersections. So, for example $\pi = \langle 3, 4, 5 \rangle$ denotes a path starting at a 3-way node, passing through another 4-way node and terminating in a 5-way node. We say that π is a sequence of three consecutive intersections. In short we denote this as a 3-sequence of type [3,4,5]. Paths can overlap but cannot be identical – i.e., two paths π_i and π_j can share proper sub-paths. So, the number of paths starting at a node is equal to its degree. In this work we consider undirected graphs – i.e., we do not account for traffic direction – and want to efficiently compute the number of occurrences of n -sequences of any possible type $[x_1, x_2, \dots, x_n]$.

Assume that we know the number of occurrences of 2-sequences of all possible types. See Figure 1a for the transitional probabilities of the city of Vienna obtained by counting. Then we can derive the statistics for all n -sequences with $n > 2$ by probabilistic reasoning. Assume that $P([x_1, x_2])$ is the probability that a 2-sequence of type $[x_1, x_2]$ occurs. Such a 2-sequence can only be followed by another 2-sequence starting at a node with degree x_2 – this is illustrated in Figure 1b. Then the probability $P([x_1, x_2, x_3])$ that a 3-sequence of type $[x_1, x_2, x_3]$ occurs is equal to the probability that a 2-sequence of type $[x_1, x_2]$ occurs times the probability that a 2-sequence of type $[x_2, x_3]$ occurs rescaled over the possible 2-sequences that start on a node of type x_2 . This can be generalized as follows:

$$P([x_1, x_2, \dots, x_n]) = P([x_1, x_2]) \cdot \prod_{i \in C} \left[P(S_i) + P(S_i) \cdot \left(1 - \sum_{j \in A_i} P(S_j) \right) \right] \quad (1)$$



■ **Figure 1** The left graph illustrates the transition probabilities (computed for the city of Vienna, Austria) from one type of intersection to another. We omitted the transitions with zero probability. The right figure illustrates an example where the probabilities have to be reassigned after the first transition is known since the options to continue are constrained.

■ **Table 1** Distribution of 2-sequences (a) and 3-sequences (b, c) of intersections.

(a) distribution of 2-sequences.			(b) distribution of 3-sequences.				(c) distribution of 3-sequences.			
Type	Count	Percentage	Type	Count	Percentage	Predicted	Type	Count	Percentage	Predicted
[3,3]	3778	0.4243	[3,3,3]	3952	0.2525	0.2494	[4,3,4]	693	0.0442	0.0422
[3,4]	1554	0.1745	[3,3,4]	1404	0.0897	0.1026	[4,3,5]	58	0.0037	0.0035
[3,5]	131	0.0147	[3,3,5]	121	0.0077	0.0086	[4,3,6]	2	0.0001	0.0001
[3,6]	6	0.0006	[3,3,6]	4	0.0002	0.0003	[4,4,4]	1709	0.1092	0.0502
[4,4]	1558	0.1749	[3,4,3]	1998	0.1276	0.0499	[4,4,5]	57	0.0036	0.0026
[4,5]	83	0.0093	[3,4,4]	1550	0.0990	0.0501	[4,4,6]	5	0.0003	0.0001
[4,6]	4	0.0004	[3,4,5]	98	0.0062	0.0026	[4,4,7]	2	0.0001	0.0001
[4,7]	3	0.0003	[3,4,6]	2	0.0001	0.0001	[4,5,4]	75	0.0047	0.0001
[5,5]	6	0.0006	[3,4,7]	2	0.0001	0.0001	[4,5,5]	2	0.0001	0.00001
Total count: 8904			[3,5,3]	234	0.0149	0.0004	[4,6,4]	3	0.0001	0.0000004
			[3,5,4]	146	0.0093	0.0002	[4,7,4]	5	0.0003	0.0000002
			[3,5,5]	13	0.0008	0.0002	[5,3,5]	6	0.0003	0.0002
			[3,6,3]	12	0.0007	0.000001	[5,4,5]	6	0.0003	0.0001
			[3,6,4]	11	0.0007	0.000001	[6,4,6]	2	0.0001	0.0000003
			Total count: 15650				Total count: 15650			

where S_i denotes a generic 2-sequence, C is the set of 2-sequences that have to be concatenated to $[x_1, x_2]$ to obtain the n -sequence $[x_1, x_2, \dots, x_n]$ and S^B is the set of admissible 2-sequences that can follow the generic i -sequence $[x_1, x_2, \dots, x_i]$. We show in Section 3.2 that probabilistic inference through the formula given in Equation 1 is reliable.

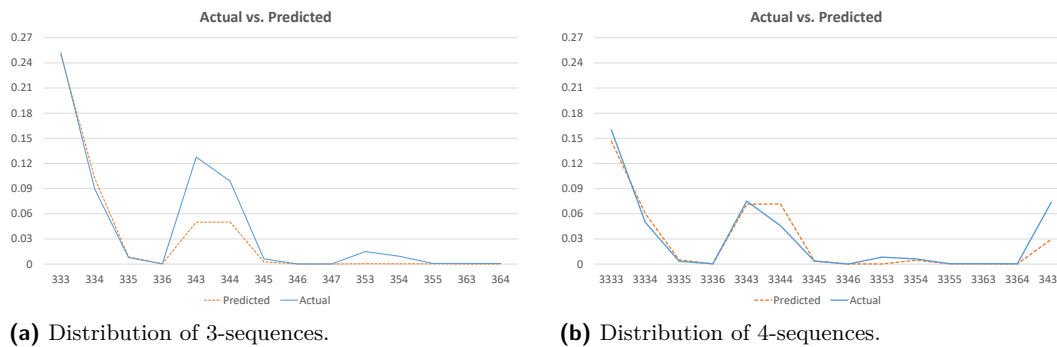
Since we can infer the probability of any n -sequence with $n > 2$, we only have to compute the probabilities for 2-sequences. This can be done straightforwardly by checking all the edges E in the graph. Since, in the worst case the number of edges is quadratic with the number of vertices, this can be done in $O(|V|^2)$. In practice, since we are dealing with graphs representing street networks we expect the number of edges to be much lower than that.

3.2 Validation

To validate the model presented in Section 3.1 we analyzed¹ a subset of the OpenStreetMap² dataset of the city of Vienna: districts 1, 3, 4, 5, 6, 7, 8, and 9 that, together, form a connected region. More specifically, we used the intersection dataset computed in [8]. We only considered intersections of type *Road* and we focused on pedestrian navigation (i.e., we

¹ The analysis has been performed on a PostGIS database.

² <http://www.openstreetmap.org/>



■ **Figure 2** Validation of predicted distributions computed with Equation 1 for intersection sequences of length 3 (a) and 4 (b).

assumed that each street segment to be traversable in both directions).

We counted the occurrences of n -sequences of intersections, with $n \in \{2, 3, 4\}$. The result of this operation for $n = 2$ (reported in Tables 1a) has been used in equation 1 to generate predictions about the distribution of n -sequences of length $n = 3$ and $n = 4$. Note that, since we assumed that each street segment is traversable in both directions, the results for pairs $[x, y]$ and $[y, x]$ are the same and we only report them once. Tables 1b and 1c show the prediction and the actual count for each type of 3-sequence. A graphical representation is reported in Figure 2a. The results for the case of 4-sequences is only reported in graphical form in Figure 2b.

3.3 Discussion and Outlook

The results produced by the introduced model look very promising (see Figure 2) and can be already utilized for a variety of applications.

The data we derived can be put together in a graph representation. This would allow for looking for structurally similar areas in different regions by applying graph matching algorithm – e.g., the algorithms presented in [15] or in [6]. The first [15] is one of the first algorithms conceived for subgraph isomorphism and is still today one of the most used techniques. It enumerates all (sub)graph matchings employing a tree search with backtracking and forward checking. It basically creates the matching incrementally; at each step it tries to match a new node. If the matching fails it backtracks to the last matched subgraph. The forward checking is used to prune the search space by looking at node adjacency. The more recent algorithm presented in [6] is based on a depth-first search strategy, also employing a set of forward-checking rules to prune the search space. For a survey on graph-matching techniques, please refer [5].

We plan to extend the work presented in this paper by also including spatial relations and semantics. The former include other quantitative measurements such as the angles formed by consecutive intersections (as already computed in [8]) or the distance between two intersections. Similarly, one can also include qualitative spatial relations such as relative direction, orientation, and visibility as done, for example, in [9].

Semantics can be included in different ways. For example, one may extend the model by considering not only intersections but also point of interests of a given types (e.g., recreational and sightseeing features). Extending the representation in such a way would allow for semantic similarity analysis and search among different regions.

We argue that the statistics derived in this paper, extended with more information as described in the paragraphs above, provide the base for the generation of virtual look-alike environments. The idea is that of incrementally generating a (mostly³) planar graph that fits to the different statistics that we generated: intersection type (3-way, 4-way,...) and shape (angular distance to regular intersections), length of intersection sequences (2-sequences, 3-sequences,...) and type of intersection sequences (3-3, 3-4, ...). A simple solution would be to resort to a brute-force procedure that deploys in the plane a number of intersections of type 3-way, 4-way, and so on according to the given statistics and then tries all possible combination of connecting those. Clearly, this is computationally very expensive and may become unfeasible already for small graphs. More feasible approaches would resort to the adaptation of random graph generation techniques [7, 13]. These techniques are capable of generating a graph uniformly at random, so they have to be adapted to fit the statistical distributions derived with our model.

References

- 1 Marc Barthélemy. Spatial networks. *Physics Reports*, 499(1-3):1–101, 2011.
- 2 Michael Batty and Sanjay Rana. The automatic definition and generation of axial lines and axial maps. *Environment and Planning B: Planning and Design*, 31(4):615–640, 2004.
- 3 Michael L Benedikt. To take hold of space: isovists and isovist fields. *Environment and Planning B: Planning and Design*, 6(1):47–65, 1979.
- 4 C. Berge. *Graphs and Hypergraphs*. Elsevier Science Ltd., Oxford, UK, UK, 1985.
- 5 Donatello Conte, Pasquale Foggia, Carlo Sansone, and Mario Vento. Thirty years of graph matching in pattern recognition. *International journal of pattern recognition and artificial intelligence*, 18(03):265–298, 2004.
- 6 Luigi P Cordella, Pasquale Foggia, Carlo Sansone, and Mario Vento. Performance evaluation of the VF graph matching algorithm. In *Image Analysis and Processing, 1999. Proceedings. International Conference on*, pages 1172–1177. IEEE, 1999.
- 7 Paul Erdos. On random graphs. *Publicationes mathematicae*, 6:290–297, 1959.
- 8 Paolo Fogliaroni, Dominik Bucher, Nikola Jankovic, and Ioannis Giannopoulos. Intersections of Our World. In *Proceedings of the 10th International Conference on Geographic Information Science (GIScience)*, Leibniz International Proceedings in Informatics (LIPIcs), Dagstuhl, Germany, 2018 (in print). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.
- 9 Paolo Fogliaroni, Paul Weiser, and Heidelinde Hobel. Qualitative spatial configuration search. *Spatial Cognition & Computation*, 16(4):272–300, 2016. doi:10.1080/13875868.2016.1203327.
- 10 Bill Hillier. *Space is the machine: a configurational theory of architecture*. Space Syntax, 2007.
- 11 Bill Hillier and Julienne Hanson. *The social logic of space*. Cambridge univ. press, 1989.
- 12 Kevin Lynch. *The Image of the City*. MIT Press, 1960.
- 13 Sadegh Nobari, Xuesong Lu, Panagiotis Karras, and Stéphane Bressan. Fast random graph generation. In *Proceedings of the 14th international conference on extending database technology*, pages 331–342. ACM, 2011.
- 14 Andres Sevtsuk and Michael Mekonnen. Urban network analysis. *Revue internationale de géomatique-n*, 287:305, 2012.
- 15 Julian R Ullmann. An algorithm for subgraph isomorphism. *Journal of the ACM (JACM)*, 23(1):31–42, 1976.


³ In first approximation we assume a street network to be a planar graph: so we exclude special situations such as tunnels that may break planarity.

Deconstructed and Inverted Multi-Criteria Evaluation for On-The-Fly Scenario Development and Decision-Making

Martin Geilhausen

Zurich University of Applied Sciences, Institute of Natural Resource Sciences, Campus Grüental, Wädenswil, Switzerland


martin.geilhausen@zhaw.ch

 <https://orcid.org/0000-0003-1797-7208>

Patrick Laube

Zurich University of Applied Sciences, Institute of Natural Resource Sciences, Campus Grüental, Wädenswil, Switzerland

patrick.laube@zhaw.ch

 <https://orcid.org/0000-0002-5926-3177>

Abstract

We propose a variation of the conventional spatial multi-criteria evaluation workflow for suitability analysis that allows efficient on-the fly scenario development for decision-making. Our approach proposes to reconstruct the conventional MCE workflow in order to exclude computationally expensive geoprocessing from the iterative scenario development. We then introduce a procedure that replaces costly iterations of spatial operations with one off-line preprocessing step followed by iterations of much less computationally expensive database queries. We illustrate our approach for deconstructed and inverted multi-criteria analysis with a case study aiming at selecting suitable sites for wind turbines in the Swiss alps.

2012 ACM Subject Classification Information systems → Geographic information systems, Information systems → Data analytics, Information systems → Expert systems

Keywords and phrases Multi-criteria evaluation, efficiency, decision-making, data structures

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.27

Category Short Paper

Acknowledgements We thank Andreas Fürholz und Valentin Stahel for many fruitful discussions about deconstructing and inverting MCE for practical decision-making and the Institute of Natural Resource Sciences, IUNR, ZHAW, for supporting our research.

1 Introduction

Spatial multi-criteria evaluation (MCE) is a formalized procedure for spatial decision problems [7], and represents one of the key applications of GIS. MCE applications include land suitability evaluation [4] or selecting suitable sites for wind farms [6]. Many of these applications have contributed to the GIScience theory by introducing computational techniques for improving the MCE workflow, proposing optimization approaches, performing sensitivity studies, handling uncertainties, as well as visualizing multi-faceted MCE results [3, 8, 2].

MCE is typically data-rich and computationally expensive, which can make it impractical for decision-making processes requiring iterative scenario development. Therefore we propose a variation of the conventional MCE that specifically aims at optimizing the workflow in such



© M. Geilhausen and P. Laube;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 27; pp. 27:1–27:7

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

a way that decision makers can exploit the full depth of MCE results for efficient on-the-fly scenario development. We achieve this by (a) proposing a deconstructed and rearranged MCE workflow where computationally expensive steps can be precomputed and hence excluded from the interactive and iterative scenario development, and (b) proposing a procedure for inverse criteria evaluation that reduces the computational costs for adjusting MCE criteria in scenario development to a feasible minimum. The work emerges from an applied research project on selecting suitable sites for wind turbines in the Swiss Alps.

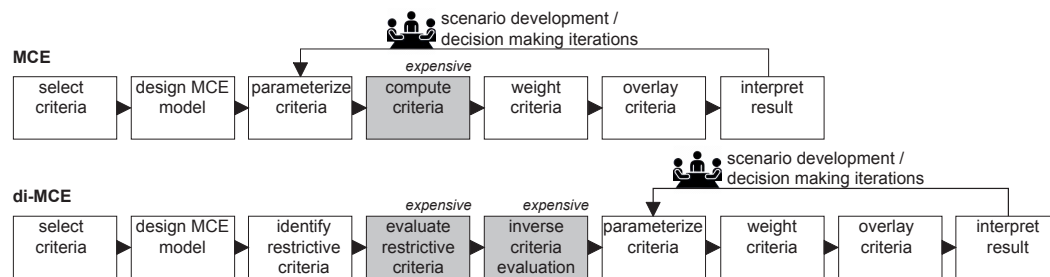
Conventional MCE typically follows a standardized workflow (Figure 1): selecting the criteria (e.g. “not within a given distance to power lines”), defining a model for translating them into spatial relations (line buffer with radius b_l), parameterizing the criteria ($b_l < 110m$), computing the respective spatial operations, standardizing and weighting the value scores (0 or 1 for not-suitable/suitable), aggregating the value scores (overlay operation), and finally interpret and validate the results, e.g. using sensitivity analysis [8]. There are several types of spatial criteria in MCE. Many criteria value locations by spatial properties, e.g. slope or soil type. This paper, however, focuses on criteria that value locations by properties of their neighborhoods. This is typically done with some form of a distance relation expressed by a buffer, e.g. “within 200m of a main road”. The latter type of criteria can then further be separated into selection (“suitable locations must be within 200m of a main road”) and exclusion criteria (“suitable location must not be within 150m of a power line”).

MCE is primarily a planning and decision-making tool, so not surprisingly, participatory concepts are increasingly used [5, 9]. The input of decision makers is also required when potentially conflicting interests have to be balanced in multi-objective evaluation [8]. At the same time, MCE is typically data-rich, which means it requires time-consuming computing and produces a wealth of data. These two aspects both hinder interactive decision-making [4]. The adjustment of a single parameter of a neighborhood criterion (e.g. increasing the exclusion distance to power lines from 110m to 150m) may trigger costly recomputing of spatial buffers and overlay operations. Under such conditions, efficient on-the-fly scenario development for decision-making is challenging.

The overarching objective of our work is developing a MCE workflow for neighborhood criteria that allows for fast and simple on-the-fly scenario development for interactive planning sessions with decision makers or for on-line decision-making tools. This leads to the following research question: How can the conventional MCE workflow be modified such that adjusting criteria parameters does not require computationally expensive spatial operations?

2 Deconstructed and inverted MCE (di-MCE)

We propose a variation of the classic MCE workflow based on two key ideas. First, we deconstruct the MCE procedure into its constituent operations and re-assemble them in such a way that fast and efficient scenario development becomes feasible. For those operations that have to be repeated frequently in scenario development, we propose secondly a procedure that inverts the perspective of the spatial criteria evaluation. Costly spatial operations are precomputed and for the scenario development phase replaced by more efficient SQL queries. We subsequently refer to *deconstructed and inverted MCE*, in short *di-MCE*.



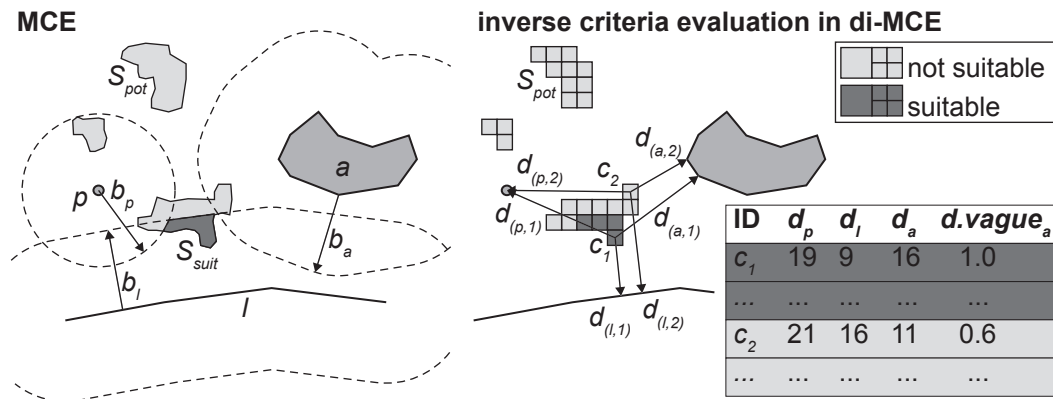
■ **Figure 1** Comparison of conventional MCE with di-MCE. In the conventional case, the iteration loop for scenario development includes adjusting criteria parameters and recomputing the criteria, which can be costly and impractical. di-MCE, instead, suggests re-assembling the workflow such that computationally expensive steps are excluded and outside the decision-making iteration loop.

Deconstructing MCE. The idea of deconstructing¹ the MCE workflow lies in disaggregating the data analysis process into its constituent steps and excluding the computationally expensive steps from the iterative scenario development phase. In our experience, most MCE studies feature some criteria that are more spatially restrictive than others and most often also non-negotiable. These could, for example, be a maximal slope or a minimal wind speed for positioning a wind turbine. We hence propose analyzing the complete set of criteria and isolating those that most reduce the resulting suitable space. Instead of including the entire set of criteria into the scenario development iterations, we propose precomputing such *restrictive criteria* and thereby excluding them from scenario development (Figure 1). Note that in *di-MCE* the computationally expensive evaluation of restrictive criteria is excluded from the iterative scenario development loop. This results in two MCE phases, where the first (off-line) phase results in the intermediate result of the *potentially suitable space* (S_{pot}). Ideally, S_{pot} only covers a small fraction of the entire study area (Figure 2). For S_{pot} we then propose an inverse criteria evaluation approach, where the computationally expensive steps can again be precomputed, and separated from the interactive and iterative decision-making.

Inverse criteria evaluation. Neighborhood criteria in conventional MCE typically focus on the spatial features that support or limit the suitability of the solution space. That means, suitability criteria are implemented using buffers that *expand from* supporting or limiting features (note the direction of the arrows in the left Figure 2). We propose inverting the perspective and focusing instead on S_{pot} identified in the previous step, and then evaluating spatial relations *directed towards* the supporting or limiting features (now note the opposite direction of the arrows in the right Figure 2).

Allowing for this inverse perspective, we tessellate the S_{pot} and for each tessellated unit compute a nearest neighbor distance d to the nearest feature of every remaining criterion. Note that for simplicity we chose a regular raster data structure for tessellating S_{pot} , resulting in candidate cells c_i . However, our approach also works for irregularly tessellated spaces, e.g., based on land-use parcels. This step translates the topological relation (“within buffer of width b ”) into a numeric attribute of a candidate cell. Again, for simplicity, we focus so far on simple distance relations to supporting or limiting features, acknowledging that more complex distance functions could be used.

¹ The term deconstructed is inspired from cookery, proposing the deconstruction of classics, e.g. as in “Deconstructed Pavlova”, the antipodean pastry classic.



■ **Figure 2** Conventional MCE adds distance buffers to points (p), lines (l), and areal features (a) with buffer parameters b_p , b_l , and b_a , resulting in the overall suitable area S_{suit} (dark grey). di-MCE precomputes distances ($d_{(p,n)}$, $d_{(l,n)}$, and $d_{(a,n)}$) from S_{pot} – tessellated into candidate cells c_i – to the nearest point, line, or area. In the example, c_1 is suitable, but c_2 is not. Crisp buffers can be dissolved into vague criteria.

Combining all computed distances for all spatial units of S_{pot} results in the cell attribute table (CAT). In unfavorable criteria constellations, this transformation can be computationally expensive. However, the distances have to be computed only once, which can be done in advance. The inversion transforms the structure of intermediate MCE results. The spatial criteria do no longer come in the form of buffer vector data or raster cost-surfaces, but as numeric data in a table. This in turn means, that adjusting parameters does not require costly recomputing of geoprocessing operations (such as buffer, overlay, or map algebra operations) but only adjusting SQL queries on attribute tables. In short, we precompute the computationally expensive spatial operations for all candidate cells and then make use of SQL queries for the final site selection in the iterative scenario development – the dark grey cells in Figure 2. Going back to Figure 1 the deconstruction idea becomes evident again. The computationally expensive calculation of the distances is precomputed and hence excluded from the scenario development. Hence, the parametrization of all negotiable criteria can happen *after* the costly spatial processing.

The inversion furthermore allows for an efficient inclusion of multi-criteria trade-offs and vagueness. First, the balancing of objectives, even conflicting objectives, can be implemented into CAT queries, using sophisticated SQL functions combining multiple attributes. The nearest neighbor distance values in the CAT secondly allow also for a straightforward inclusion of vagueness into criteria evaluation. Membership functions can be applied to nearest neighbor distances, dissolving unrealistically crisp buffer boundaries into gradual memberships.

3 Case study: Positioning wind turbines

We illustrate our approach with the very research project that highlighted to us the shortcomings of conventional MCE. The study aimed at finding suitable areas for positioning wind turbines in a region of the Swiss alps. The criteria covered technical requirements (maximal slope, accessibility for construction), economic requirements (e.g. minimal wind speed of $4.5 \frac{m}{s}$), and a set of regulatory requirements given through a federal guideline [1]. A subset of the approximately 50 criteria and their parameterization for two types of wind turbines T_1 and T_2 is given in Table 1. Note that most criteria are of the neighborhood type.

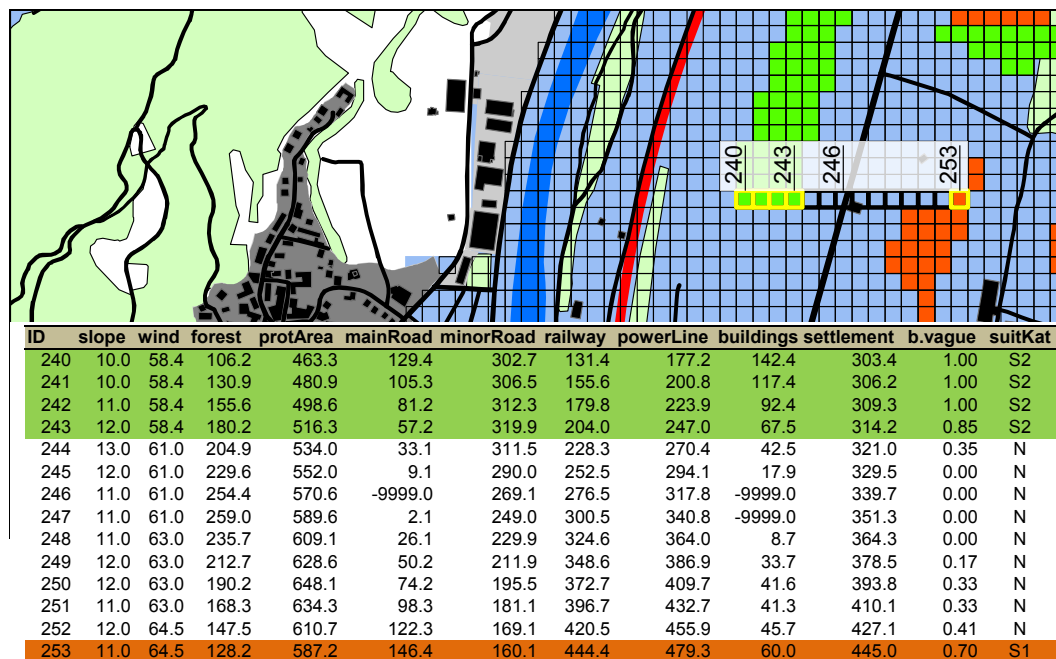


Figure 3 Excerpt from the results of the wind turbine project. The map features a number of spatial layers required for evaluating the suitability criteria, e.g. streets (black), railways (red), settlement areas (grey), and buildings (black). For a small horizontal transect of cells c_{240} to c_{253} the computed nearest neighbor distances d_{nn} are displayed in the cell attribute table below. Note, -9999.0 as in c_{246} codes the case when the candidate cell touches the feature (e.g. mainRoad).

Table 1 Eight out of approx. 50 criteria for positioning two types of wind turbines T_1 and T_2 .

Criterion	T_1	T_2	Criterion	T_1	T_2
forest	not within	not within	mainRoad	$d > 17$	$d > 33$
protArea	not within	not within	minorRoad	$d > 17$	$d > 33$
buildings	$d > 50$	$d > 50$	railway	$d > 17$	$d > 33$
settlement	$d > 100$	$d > 100$	powerLine	$d > 110$	$d > 190$

The available wind field, a slope threshold and an accessibility criterion were identified as restricting criteria and consequently used for computing the *potentially suitable space* (S_{pot}). The map in Figure 3 illustrates a small space depicting a subset of all criteria (streets, railways, settlement areas, building, forest) as well as the precomputed *potentially suitable space* (S_{pot}). In correspondence with Figure 2 (S_{pot}) was tessellated into $25m * 25m$ cells, the blue layer in the background indicates the areas with enough wind. Finally, the overall suitable area S_{suit} is depicted with orange and green cells (suitability categories S1 and S2).

The map also features a short transect of candidate cells c_{240} to c_{253} for which the table below the map shows a subset of the CAT. Whereas for slope and wind the actual values of the respective field variable are given (which were used for computing (S_{pot}), all other attributes are nearest neighbor distances d_{nn} to supporting or limiting spatial features. The last two columns illustrate the use of vagueness and the final suitability category. The distances in the table can now be directly compared with the criteria parameters in Table 1.

The case study can illustrate the advantages of di-MCE. Consider the criterion “sites for wind turbines must not be within a defined distance to power lines”. This parameter is clearly

a function of the size of the turbine, Table 1 indicates $d > 110m$ for the smaller type T_1 , and $d > 190$ for the larger T_2 . Assuming the scenario for a new turbine type of intermediate size, the suitability for each candidate cell can easily be recalculated from the precomputed distance value in the CAT, without costly repetition of spatial operations. The column **b.vague** in Figure 3 finally illustrates the inclusion of vagueness. To this end, the criterion “distance to buildings” has been dissolved into a vagueness value using a membership function (0 for $d_b < 25$, 1 for $d_b > 75$, and a linear function in between).

4 Discussion and conclusions

The goal of our work is re-structuring the MCE workflow in such a way that on-the-fly scenario development becomes feasible. This explicitly does not mean reducing the overall computing load of MCE. Depending on the criteria constellation, the proposed inverse criteria evaluation may even add to the overall computation cost. However, with our deconstructed and re-assembled workflow, the crucial step of adjusting criteria parameters appears in the sequence of operations *after* the costly spatial operations, making iterative scenario development perfectly feasible. Replacing spatial operations on vector or raster data with queries on attribute tables offers the additional benefit of the straightforward integration of vague criteria and the balancing of conflicting objectives. Once the nearest neighbor distances are computed, SQL queries allow for very flexible transformation and combination of multiple criteria.

Our approach is most suited for MCE projects with (i) frequent stakeholder interaction, (ii) a set of criteria with one or two criteria being rather restrictive and non-negotiable, and (iii) a predominant use of distance-based neighborhood criteria. In our wind turbine site selection case study all three preliminaries were given. We argue, however, that the majority of MCE studies comply with at least some of these preliminaries, hence offering at least partially to benefit from the advantages of di-MCE. In the wind turbine case study we only considered simple distance-based nearest neighbor criteria. More complex neighborhood functions could be conceptualized and implemented within di-MCE. We are currently working on more complex neighborhood functions, comparable to focal and zonal map algebra operations (e.g. %-forest cover within distance d around a candidate cell).

References

- 1 Mattia Cattaneo and Leonhard Zwiauer. Concept windenergy – Federal guidelines for planning windenergy installations. Technical report, Swiss Federal Office for Spatial Development, 2017.
- 2 Bakhtiar Feizizadeh, Piotr Jankowski, and Thomas Blaschke. A GIS based spatially-explicit sensitivity and uncertainty analysis approach for multi-criteria decision analysis. *Computers & geosciences*, 64:81–95, 2014.
- 3 Montserrat Gómez-Delgado and Stefano Tarantola. Global sensitivity analysis, GIS and multi-criteria evaluation for a sustainable planning of a hazardous waste disposal site in Spain. *Int.J. of Geographical Information Science*, 20(4):449–466, 2006.
- 4 Isabel Jaisli, Patrick Laube, Sonja Trachsel, Pascal Ochsner, and Sarah Schuhmacher. Suitability evaluation system for the production and sourcing of agricultural commodities. *Computers and Electronics in Agriculture*, 2018.
- 5 Piotr Jankowski. Towards participatory geographic information systems for community-based environmental decision making. *J. of env. management*, 90(6):1966–1971, 2009.

- 6 Dionysis Latinopoulos and K Kechagia. A GIS-based multi-criteria evaluation for wind farm site selection. A regional scale application in greece. *Renewable Energy*, 78:550–560, 2015.
- 7 Jacek Malczewski. GIS-based multicriteria decision analysis: A survey of the literature. *Int. J. of Geographical Information Science*, 20(7):703–726, 2006.
- 8 Jacek Malczewski and Claus Rinner. *Multicriteria decision analysis in geographic information science*. Springer, 2016.
- 9 Beni Rohrbach, Patrick Laube, and Robert Weibel. Comparing multi-criteria evaluation and participatory mapping to projecting land use. *Landscape and Urban Planning*, 176:38–50, 2018.

Space-Time Representation of Accessible Areas for Wheelchair Users in Urban Areas

Amin Gharebaghi

Center for Research in Geomatics, Université Laval, Quebec, Canada
amin.gharebaghi.1@ulaval.ca

Mir Abolfazl Mostafavi

Center for Research in Geomatics, Center for Interdisciplinary Research in Rehabilitation and Social Integration, Université Laval, Quebec, Canada
mir-abolfazl.mostafavi@scg.ulaval.ca

Abstract

Providing personalized information on the accessibility of urban places for people with disabilities can significantly increase their social participation. This information should be adapted with respect to their needs at the specific time and space. Location-based technologies are considered as proper services to provide such information and encourage mobility of these people in urban areas. However, generally these services focus on the spatial conditions of the accessibility and ignore users' capabilities and time dependent constraints. This is much more challenging for people with disabilities given the diversity of their physical capabilities and preferences. To address this issue, we propose an approach to measure the space-time accessibility of urban areas considering environmental characteristics, users' capabilities, and time constraints. The proposed approach is unique and it highlights time constraint that is rooted in time geography theory. Unlike the classical time geography, which suggests a uniform travel velocity, we consider a variable travel velocity in the proposed approach, which is more relevant to the mobility of people with disabilities. To implement the proposed method, a Fuzzy approach is applied to evaluate the wheelchair speeds for the segments of a pedestrian network. The proposed approach is implemented in Saint-Roch, Quebec City for a case study and the results are presented and discussed.

2012 ACM Subject Classification Human-centered computing → Accessibility technologies

Keywords and phrases Mobility, Wheelchair users, Accessibility, Time geography, Potential travel areas

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.28

Category Short Paper

1 Introduction

The last two decades have seen a growing trend towards the space-time accessibility measures that allow geo-visualization of human activity patterns and evaluation of the accessibility for people through space and time [10]. Spatial accessibility is the result of interaction between the individual and the environment [2]. For people with disabilities, this is significant as it is in accordance with the definition of the handicap process; a path can be accessible for some while it can be inaccessible for others even if the environment is the same. People with disabilities schedule their activities considering not only spatial conditions but also temporal constraints as well as their capabilities. Hence, in order to assess the accessibility of urban areas, three main elements including environmental factors, personal factors, and the individual travel time budget should be taken into account (Figure 1).



© Amin Gharebaghi and Mir Abolfazl Mostafavi;
licensed under Creative Commons License CC-BY

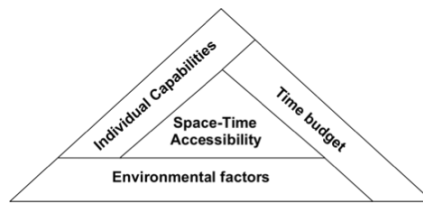
10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 28; pp. 28:1–28:6

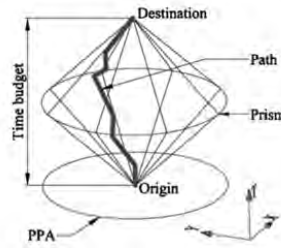
Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



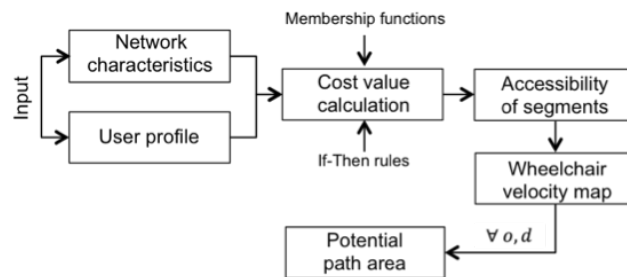
■ **Figure 1** The main elements of space-time accessibility measure.



■ **Figure 2** Time geographical concepts [10].

Time-geography concepts introduced by [3] are efficient tools to model the participation of people with different capabilities through space and time. Although this theory has a conceptual attraction and strength, very few studies have been reported on its applicability for the real world situations. This is mainly because of the difficulties of the abstraction, modeling and implementation of the real world complexities into the GIS [6]. Time geography theory relies on the concepts such as space-time prism, space-time paths, and potential path areas. The space-time prism is the package of all possible space-time paths between specified locations and times, which emphasizes on the individual ability to participate in the activities. The spatial footprint of the space-time prism is the potential path area, which is the geometric region in the space that is accessible for a moving object (for more details please refer to [8]). These concepts are visualized in Figure 2. As shown in this figure, the classical time-geography concepts suggest a uniform travel velocity, which does not represent all the complexities of the real-world situation. For example, the travel velocity of people with disabilities and specifically wheelchair users mostly confined to the characteristics of the environment (e.g. surface quality) and their capabilities. Indeed, the accessibility level of segment (ALS), the wheelchair speed (WS), and ultimately the needed travel time (TT) of segments change from an individual to others. Therefore, these principles should be adapted for visualizing the potential travel areas (PTA) of wheelchair users. In this paper, the notion of travel area is used for an area representing a set of points reachable for a wheelchair user within a specified time, which corresponds to the potential path of traditional time geography. Although in recent years few authors have slightly adapted the time geography concepts to make it more suitable for the reflection of real-world situations [4, 5, 11, 10, 9, 7], no researches took into account the capability of people with disabilities to generate their PTA at the specified time budget. In order to address this issue, we aim to generate such areas considering time intervals for traveling of manual wheelchair users.

Following the introduction section, the paper begins with elaborating the proposed methodology in section 2. Section 3 explains the assessment of spatial accessibility of pedestrian network, which is the central part of the space-time accessibility measure process. In section 4, we employ the proposed methodology in the study area for two wheelchair users who have different level of capabilities. In this part, the spatial accessibility maps and the potential travel area through the time intervals are generated. The paper is included in section 5.



■ **Figure 3** An overview of the proposed methodology.

2 Methodology

In reality, the WS -specially the speed of manual wheelchairs - depends on both the characteristics of the path (e.g. the surface quality and the slope) and the user capabilities. This principle should be reflected in the space-time accessibility evaluation process. Indeed, to measure the space-time accessibility, three fundamental data are required including (1) the characteristics of the travel environment; (2) the user capabilities regarding different characteristics of the environment; and (3) the WS of different ALS. In this paper, we propose a framework to measure the space-time accessibility of urban areas for manual wheelchair users. To fulfill the proposed methodology, first, we calculate the spatial ALS in the given network (i.e. study area) based on the user capabilities. To evaluate the user capabilities, the perceived ability (i.e. confidence) of a person is measured while performing a given task. Indeed, the user confidence is identified as a stronger predictor of performance than the skill itself [12]. The spatial ALS is evaluated for segments of pedestrian network by aggregating the user confidences with respect to different characteristics of the segments. To realize, If-Then rules approach in a fuzzy environment is employed. The details of this process are given in the following section. Following that, the WS of network's segments are calculated based on the calculated ALSs (i.e. $WS = f(ALS)$). Finally, the TTs and consequently the PTAs are generated within the different time intervals. Figure 3 depicts the overview of the proposed methodology.

3 Evaluation of the spatial accessibility as the fundamental part of the methodology

In order to evaluate the spatial ALS, a cost value for a segment representing the ALS should be calculated. This value is computed by aggregating the user confidences with respect to the different properties of that segment. These properties are mostly determined by crisp values such as 5% as the slope of a segment. However, in many cases the precise quantitative values are often inadequate to describing real-life situations and people use a more qualitative way to characterize environmental factors that affect mobility (e.g. narrow sidewalks). In our study, the fuzzy logic approach [13] is utilized to meet these requirements. To carry out the fuzzy logic approach, first, the transformation from the crisp values into a non-crisp fuzzy environment is conducted. This process is called fuzzification, which is performed by defining membership functions. A membership function is a mathematical function which maps the association of a value to a set between 0 and 1. Thus, the values of the segments' properties are transferred into fuzzy set classes using predefined membership functions [1]. Following the fuzzification process, the user confidence values should be associated to the defined fuzzy subsets. For example, high level of confidence might be associated to the gentle slope. In this



■ **Figure 4** The spatial accessibility map for an individual.

paper, five fuzzy sets are considered to indicate the user's confidence level including Very Low (VL), Low (L), Medium (M), High (H), and Very High (VH). These values are measured regarding three characteristics of the network's segments including Slope (S), Width (W), and Surface Quality (SuQ). The If-Then rules are subsequently defined to aggregate the user confidences and, consequently, calculate the ALS as the output variable. For example:

If (theS.Con is VL) and (theSuQ.Con is L) Then (the segment is NA)

where S.Con refers to the user confidence with respect to slope values and SuQ.Con refers to the user confidence with respect to the surface quality values of a segment. Once the rules are defined and the aggregation step is performed, the ALS can be derived. To realize, a defuzzification technique is applied to produce exact numerical values from the fuzzy values based on the defined membership functions and defined rules. The output values are determined the ALS through four categories of Not Accessible (NA), Low Accessible (LA), Accessible (A), and Very Accessible (VA).

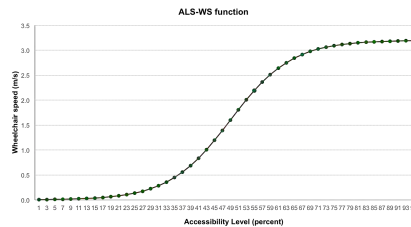
4 Experiments and results

Following the evaluation of the spatial ALS, The proposed methodology calculates the WS, TT, and ultimately PTA. This process is simulated for an individual who wants to travel within Saint-Roch, Quebec City. The required inputs including a graph of pedestrian network-containing nodes, edges, and their attribute tables- is collected from several data sources including collections of Ville de Québec, 2015 and web portal of Ville de Québec (i.e. S, W, SeL, and SuQ). The confidence values regarding different parameters of network (Table 1), and the WS-ALS function as s-functions (Figure 5) are simulated. The evaluation of the spatial ALS is carried out using the fuzzy approach for a part of study area. The results of this step are visualized as accessibility map in a web-based GIS tool, which is called MobiliSIG (Figure 4). Following this, the WSs for each segment based on their spatial ALS are extracted from the defined functions and ultimately the TTs of each segment for each subject are calculated. In other words, the extracted values of WSs from ALS-WS functions are used to calculate the required time for each segment (i.e. $Time = f(SegmentLength, Wheelchair\ speed)$). The travel times are used as the weights of segments to calculate the time of network vertices. Figure 6 shows a simulation of ALS-WS function an individual. To understand the whole process, we illustrate the input and output data for couple of segments shown by Table 2.

Finally, the potential travel areas is generated from a given origin using the time geography concepts, which contains fundamental information about the overall directionality of the

■ **Table 1** The confidence values regarding different parameters of network.

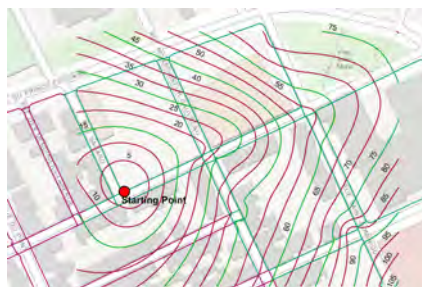
#	Slope			Width			Surface Quality		
	Gentle	Moderate	Steep	Narrow	Moderate	Wide	Good	Fair	Poor
User Confidence	90	65	20	15	70	100	90	60	35



■ **Figure 5** A simulation of ALS-WS function.

■ **Table 2** The input and results for couple of segment examples.

Segment Id			1	2	3	4	5	6	7	8
Input	Segment Attribute	Length (m)	100	250	50	200	150	150	50	100
		S (%)	3	8	-2	4	3	2	-7	-4
		W (m)	1.5	1	2	1.5	15	1.5	1.3	1.7
		SuQ	Good	Bad	Good	Fair	Fair	Good	Bad	Fair
Results	ALS		0.8	0.25	0.85	0.5	0.68	0.68	0.35	0.7
	WS (m/s)		2.4	0.75	2.55	1.5	2.04	2.04	1.05	2.1
	TT (min)		0.7	5.6	0.3	2.2	1.2	1.2	0.8	0.8



■ **Figure 6** The potential travel area for a wheelchair user in 5s time intervals.

network and are useful for the assessment of space – time accessibility (Figure 6). In this figure, a network potential travel area is calculated in 5s time intervals. The contour lines indicate the feasible traveling parts of network in the time intervals. This knowledge on time-space accessibility would provide insights on how wheelchair users can schedule their daily activities using accessible paths and in the given time budget.

5 Conclusion

In this paper, we proposed an approach to measure the space-time accessibility of urban areas for manual wheelchair users. The originality of the method is in its focus on the people with limited mobility while considering time constraints. The approach was carried out in two steps including spatial accessibility evaluation of the pedestrian network segments, and the travel time evaluation of the segments. To perform the first step, we accounted the segments'

properties (i.e. the slope, the width, and the surface quality) and the users' confidences. In this process, we were benefited from the fuzzy logic approach and defined the if-then rules to aggregate the users' confidences regarding the segments' properties. Then, the required travel time was evaluated based on the spatial accessibility levels of segments. The process was carried out for each segment employing the time geography theory. Unlike the classical time geography concepts, we considered the variable travel speeds for the manual wheelchairs. Finally the spatial accessibility map and the potential travel areas in the different time intervals were generated. The process was implemented in our study area – Saint-Roch, Quebec City – for a case study. The achievements of this research would be employed in the location-based services designed for people with disabilities to provide insights on how these people schedule their daily activities by accessible paths and in their time budget.

References

- 1 Amin Gharebaghi, Mir-Abolfazl Mostafavi, Seyed Chavoshi, Geoffrey Edwards, and Patrick Fougeyrollas. The Role of Social Factors in the Accessibility of Urban Areas for People with Motor Disabilities. *ISPRS International Journal of Geo-Information*, 7(4):131, 2018. doi:10.3390/ijgi7040131.
- 2 Amin Gharebaghi, Mir-Abolfazl Mostafavi, Geoffrey Edwards, Patrick Fougeyrollas, Patrick Morales-Coayla, François Routhier, Jean Leblond, and Luc Noreau. A Confidence-Based Approach for the Assessment of Accessibility of Pedestrian Network for Manual Wheelchair Users. In *International Cartographic Conference*, pages 463–477. Springer, 2017.
- 3 Torsten Hägerstrand. What about people in Regional Science? In *Papers of the Regional Science Association*, volume 24, pages 6–21, 1970. doi:10.1007/BF01936872.
- 4 M D Hendricks, M J Egenhofer, and K Hornsby. Structuring a wayfinder's dynamic space-time environment. *Conference on Spatial Information Theory*, pages 75–92, 2003.
- 5 Mei-Po Kwan. Space-Time and Integral Measures of Individual Accessibility: A Comparative Analysis Using a Point-based Framework, 2010. doi:10.1111/j.1538-4632.1998.tb00396.x.
- 6 Mei-Po Kwan and Joe Weber. Individual accessibility revisited: implications for geographical analysis in the twenty-first century. *Geographical Analysis*, 35(4):341–353, 2003.
- 7 Jinyung Lee and Harvey J. Miller. Measuring the impacts of new public transit services on space-time accessibility: An analysis of transit system redesign and new bus rapid transit in Columbus, Ohio, USA. *Applied Geography*, 93(February):47–63, 2018. doi:10.1016/j.apgeog.2018.02.012.
- 8 Harvey J. Miller. Time Geography and Space-Time Prism. *International Encyclopedia of Geography: People, the Earth, Environment and Technology*, pages 1–19, 2017. doi:10.1002/9781118786352.wbieg0431.
- 9 Tijs Neutens, Matthias Delafontaine, Tim Schwanen, and Nico van de Weghe. The relationship between opening hours and accessibility of public service delivery. *Journal of Transport Geography*, 25:128–140, 2012. doi:10.1016/j.jtrangeo.2011.03.004.
- 10 Tijs Neutens, Nico Van de Weghe, Frank Witlox, and Philippe De Maeyer. A three-dimensional network-based space-time prism. *Journal of Geographical Systems*, 10(1):89–107, 2008. doi:10.1007/s10109-007-0057-x.
- 11 Martin Raubal, Harvey J Miller, and Scott Bridwell. User-centred time geography for location-based services. *Geografiska Annaler: Series B, Human Geography*, 86(4):245–265, 2004. doi:10.1111/j.0435-3684.2004.00166.x.
- 12 Paula Wendy Rushton. *Measuring confidence with manual wheelchair use: a four phase, mixed-methods study*. Phd thesis, University of British Columbia, 2010.
- 13 Lofti Zadeh. The concept of a linguistic variable and its application to approximate reasoning—II. *Information Sciences*, 8(4):301–357, 1975. doi:10.1016/0020-0255(75)90046-8.

Spatial Periodicity Analysis of Urban Elements Application to the Ancient City of Amida

Jean-François Girres

Université Paul-Valéry Montpellier 3, IRD, UMR GRED, Montpellier, France
jean-francois.girres@univ-montp3.fr

Martine Assenat

Université Paul-Valéry Montpellier 3, EA CRISES 4424, Montpellier, France

Robin Ralite

Master Géomatique de Montpellier, Montpellier, France

Ester Ribo-Delissey

Master Géomatique de Montpellier, Montpellier, France

Abstract

The characterization of urban structures using morphological indicators is the subject of many applications in the domains of urban planning and transport, but also in less traditional disciplines, such as urban archeology. When reading actual urban plans, it may be possible to identify relics of ancient cities, and to characterize them with the help of appropriate indicators. In this context, we propose a method for the characterization of the spacing between urban elements based on the analysis of their spatial periodicity. The purpose of this method is to detect specific distances in the actual urban structure, potentially characteristic of ancient measurement units. This method is implemented in a GIS software, to facilitate its use by historians and archeologists, and is illustrated by an application on the ancient roman city of Amida (Diyarbakir, Turkey).

2012 ACM Subject Classification Human-centered computing → Accessibility theory, concepts and paradigms

Keywords and phrases Spatial analysis, Periodicity, Urban structures, Archeology

Digital Object Identifier 10.4230/LIPICs.GIScience.2018.29

Category Short Paper

Acknowledgements We want to thank Eloise Noc, who provided the geographical databases on the city of Diyarbakir, which allowed to perform this study.

1 Introduction

The characterization of urban morphology has been the subject of many contributions over the last 50 years. Numerous indicators have been proposed in different fields of application to characterize urban structures, based on the analysis of their constituent elements (e.g. road networks, or buildings). For example, in the domains of urban planning and transports, several indicators have been developed to characterize urban networks [4] [2]. In the field of cartography, indicators have also been proposed to orchestrate operations of cartographic generalization according to specific urban patterns [5]. If lots of indicators are available, some authors [6] consider that many of them are not appropriate to characterize urban structures, especially because they are not expressed in spatial units, which can be problematic in terms of interpretation or comparison between cities. In the domain of urban archeology, the use of



© Jean-François Girres, Martine Assenat, Robin Ralite, and Ester Ribo-Delissey;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 29; pp. 29:1–29:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

indicators easily interpretable is an issue to characterize ancient urban structures. In this article, we propose a method for the characterization of urban structures based on the analysis of the spatial periodicity of urban elements in a particular orientation. The proposed method aims at determining specific spacings between urban elements (or topographic elements considered as a limit), potentially characteristic of ancient measurement units, which provide evidences of the persistence of ancient urban structures in the actual city plan. The proposed method, implemented in a GIS software, is applied on the city plan of Diyarbakir, to reveal and characterize relics of the ancient roman city of Amida.

2 Spatial periodicity analysis methods

The analysis of the distances between urban elements can provide pertinent information to characterize actual or ancient urban structures. For example, indicators of urban morphology can be constructed using the average distance between intersected streets along a road. In this article, we focus on the analysis of the most occurrent spacing between urban elements in a given orientation. These urban elements are not necessarily streets, but they can also be building walls or cadastral boundaries for example. To determine the most occurrent spacing between urban elements, the proposed methodology is established in two steps: (1) definition of a reference orientation of the studied urban elements, (2) analysis of the spatial periodicity of the elements inscribed in the specified orientation.

2.1 Orientation of urban elements

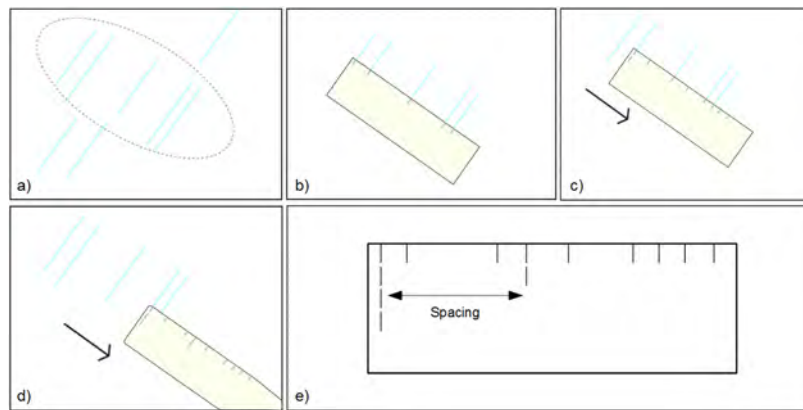
The definition of the orientation of geographic features can be a difficult task according to the geometric primitives and the level of complexity of the analysis. Indeed, for polylines, the orientation is generally defined according to the azimuth inscribed between the first and the last vertex of the geometry. However, in particular configurations, this simple definition of orientation does not necessarily reflect the heterogeneity of internal orientations of the polyline, which may be indicative of underlying urban structures. For polygons, the definition of the orientation is more complex and several measures can be proposed to determine it [3]. For example, the orientation of buildings can be defined using the longest side of the polygon, or the smallest bounding box. As a consequence, these different definitions of orientation can also generate inconsistencies for complex shapes. To overcome these problems for the definition of the orientation of urban elements, we propose to perform a disaggregation of polyline and polygon geometries into a set of segments, whose orientations will be measured separately. Once the geometry of urban elements is disaggregated into a set of segments, a selection of the segments following a particular orientation is performed. The reference orientation is computed with an angular tolerance, generally between 1° and 2° .

2.2 Analysis of spatial periodicity using the “paper band” method

When the orientation is defined, the most occurrent spacing between elements of a urban structure is determined by analyzing the spatial periodicity using the so-called “paper band” method. This method derives its name from the work of archaeologists who aimed at defining the most frequent spacing between urban elements with the help of a band of paper that they slid according to a defined orientation.

Methodology The “paper band” method works as exposed thereafter and in figure 1:

- Retrieving of a set of aligned segments according to a given orientation (a)
- Band initialization perpendicularly to segments, and pointing of intersected segments (b)



■ **Figure 1** Measurement procedure using the paper band method.

- Band translation at the following segment, and pointing of intersected segments (c)
- Band translation at the following segment, and pointing of intersected segments (d)
- ...
- Stop translation at the last segment and measurement of the most frequent distance (e)

Goals and issues. This method can be used to determine specific spacings between urban elements in a particular (or dominant) direction. In the field of urban archeology, the detected spacings can correspond to ancient measurement units fossilized in the actual urban structure. Nevertheless, the manual application of the “paper band” method remains problematic when dealing with large volumes of data, such as entire urban plans. Indeed, this method is redundant and time-consuming, and in addition, it is exposed to inaccuracies related to human intervention, which can be detrimental for a good restitution of specific spacings. Therefore, an implementation of this method in a GIS tool is proposed to facilitate the automation of these tasks.

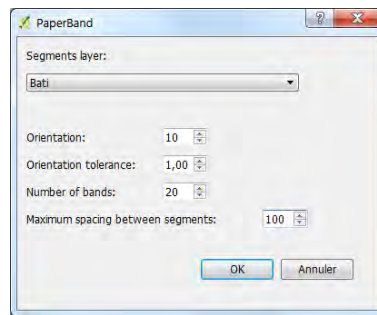
3 GIS Implementation

In order to automate the periodicity analysis method presented in the previous section, an extension of the QGIS GIS software has been developed. The QGIS GIS software has been selected because of its large community of users and its rich documentation, as well as its facility to implement plug-ins using the QGIS Python API and the PyQt library for the development of user interfaces. The developed plug-in, called “paper band”, automates the study of the spatial periodicity of a set of input segments, according to a given orientation.

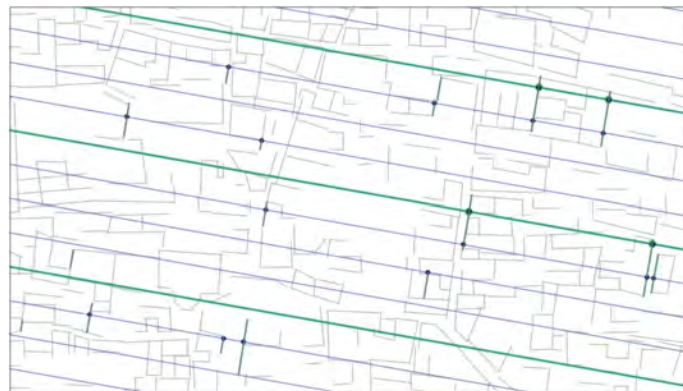
Input parameters. The parameters necessary for the analysis of spatial periodicity are: the input segment layer; the orientation (in degrees); the angular tolerance (in degrees); the number of bands (i.e. the resolution of the analysis); the maximum spacing between elements.

For example, in the figure 2, the analysis relates to the buildings having walls oriented between 9° and 11° (10° with an angular tolerance of $\pm 1^\circ$). The analysis is carried out using 20 bands, and the maximum allowed spacing between walls is 100 meters.

Output results. Once the extension is executed, a layer of points is generated, corresponding to the intersections between search bands and segments of the input layer, as exposed in



■ **Figure 2** User interface of the “paper band” plug-in.



■ **Figure 3** Impact of the resolution of the analysis (with 20 or 80 bands).

figure 3. The distances between intersected points are iteratively computed for each band, and the most occurrent spacing is determined.

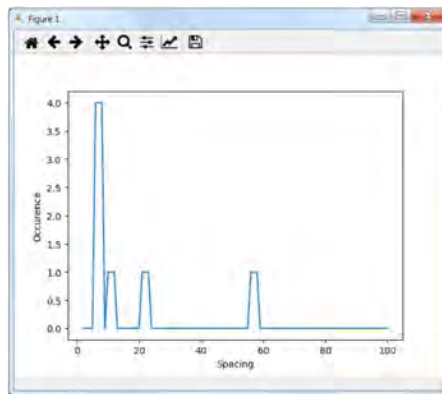
Figure 3 illustrates the impact of the resolution (i.e. the number of bands configured) on the input segments used for the periodicity analysis. If a high resolution may be time consuming, a too low resolution can ignore numerous reliable elements for the periodicity analysis, which could affect final results. One solution is to perform the analysis using various resolutions to study the sensibility of the spacings between elements according to the chosen resolution.

As a result, the plug-in finally generates a diagram representing the distribution of distances between urban elements (figure 4). A spreadsheet containing the computed distances is also generated. For instance, in figure 4, it is found that the most frequent spacing is about 10 meters.

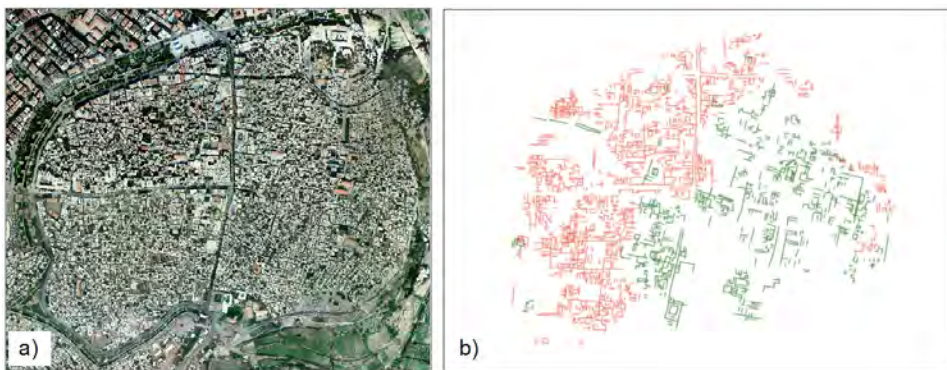
4 Application to the characterization of the ancient city of Amida

The proposed method for the characterization of the spacing between urban elements is applied on the urban structure of the city of Diyarbakir, which is built on the site of the ancient roman city of Amida.

Study area. The city of Diyarbakir is the main kurdish city of southeastern Turkey. Its site is established on the ancient roman city of Amida, which presents the characteristics of an ideal roman city: the city presents a quadripartite plan, and is surrounded by a wall. Indeed, in a roman city, the urban structure is built on two axes: the *cardo* (oriented North-South) and the *daecumanus* (oriented East-West).



■ **Figure 4** Graphic representation of spatial periodicity.



■ **Figure 5** Aerial photography (a) and the two urban structures (b) of the city of Diyarbakir.

Assumptions. The date of the founding of the roman city of Amida remains controversial [1]. Indeed, a first foundation would date from the time of the Sévères (green plan on figure 5b), and a second would date from Constance II (red plan on figure 5b). Each of these two foundations remain in the actual city plan, through two characteristic structures, one oriented North-South, and the other one with an angle of about 10° .

Despite its successive occupations (e.g. Byzantines or Ottoman Empire), the city of Diyarbakir retains the relics of these two plans in its current urban structure. So, it seems possible to identify ancient urban structures inherited from the antiquity through the constituent elements of the current city, such as streets, walls, monuments, or parcel alignments for example.

Objectives. In order to confirm the assumptions concerning the existence of two different structures in the current city plan of Diyarbakir, we seek to characterize the spacings between their constituent urban elements. In this paper, we only focus on the characterization of the inherited urban structure from the roman city of Amida founded during Constance II (oriented in a North-South direction). More particularly, the analysis of the spacing between urban elements will seek to reveal the use of roman measurement units.

So, the analysis of the spatial periodicity was carried out using the cadastral plan of the city of Diyarbakir, using an orientation of 0° , with a tolerance of $\pm 2^\circ$. The analysis was performed with a resolution of 80 bands.

First results. The results show that the distribution of spacings between urban elements has two peaks: one between 3 and 4 meters and the second between 8 and 10 meters. The first measure could correspond to the distance of streets and secondary roads, which were about 3.5 meters in roman period, but it could also correspond to an old unit of measure called the “roman perch”, equivalent to 2.964 meters or 10 roman feet. This unit of measure could also explain the second peak of the distribution, which would be equivalent here to 30 roman feet, since this measurement was used for the sizing of rooms in the roman period. These results are obviously preliminary, and need to be established on other orientations, and to be enriched with the help of other urban elements, such as excavated buildings for example.

5 Conclusion and further works

This article has presented a method of morphological characterization of urban structures, by analyzing the spatial periodicity between its constituent elements. The proposed method, known as the “paper band” method, is used to determine the most occurrent spacing between the elements of a urban structure, according to a specific orientation. This method is particularly relevant in the field of urban archeology, in order to characterize spacings corresponding to ancient measurement units. More generally, the proposed method provides additional metrics to characterize spatial distances in urban structures. The application of this method on the cadastral plan of the city of Diyarbakir offered opportunities to illustrate the persistence of the roman city of Amida in the actual urban structure, by revealing characteristic distances corresponding to ancient roman feet. This work obviously remains to be extended, especially by studying spatial periodicity on complementary elements, such as buildings for example, and in a larger range of orientations. To conclude, this study is all the more justified because the city of Diyarbakir is currently the subject of important destructions of its buildings, which are as many relics of the ancient roman city of Amida.

References

- 1 M. Assénat and A. Pérez. Amida 4. Constance et Amida. *Anatolia Antiqua*, 23:199–217, 07 2014.
- 2 A. Cliff, P. Haggett, J. Ord, K. Bassett, and R. Davies. *Elements of spatial structure : a quantitative approach*. Cambridge [Eng.] ; New York : Cambridge University Press, 1975.
- 3 C. Duchêne, S. Bard, X. Barillot, A. Ruas, J. Trévisan, and F. Holzapfel. Quantitative and qualitative description of building orientation. In *6th ICA Workshop on Generalisation and Multiple Representation, 28-30 April, Paris (France)*, 04 2003.
- 4 K. J. Kinsky. *Structure of Transportation Networks: Relationships Between Network Geometry and Regional Characteristics*, volume 84. University of Chicago. Department of Geography. Research papers., 07 1967.
- 5 A. Ruas. A method for building displacement in automated map generalisation. *International Journal of Geographical Information Science*, 12(8):789–803, 1998. doi:10.1080/136588198241509.
- 6 D. Smith. *Patterns in human geography. An introduction to numerical methods*. David and Charles Newton Abbot, 1975.

Gaze Sequences and Map Task Complexity

Fabian Göbel

Institute of Cartography and Geoinformation, ETH Zurich, Zurich, Switzerland
goebelf@ethz.ch

Peter Kiefer

Institute of Cartography and Geoinformation, ETH Zurich, Zurich, Switzerland
pekiefer@ethz.ch

Ioannis Giannopoulos

Department of Geodesy and Geoinformation, Vienna University of Technology, Vienna, Austria
igiannopoulos@tuwien.ac.at

Martin Raubal

Institute of Cartography and Geoinformation, ETH Zurich, Zurich, Switzerland
mraubal@ethz.ch

Abstract

As maps are visual representations of spatial context to communicate geographic information, analysis of gaze behavior is promising to improve map design. In this research we investigate the impact of map task complexity and different legend types on the visual attention of a user. With an eye tracking experiment we could show that the complexity of two map tasks can be measured and compared based on AOI sequences analysis. This knowledge can help to improve map design for static maps or in the context of interactive systems, create better map interfaces, that adapt to the user's current task.

2012 ACM Subject Classification Human-centered computing → Empirical studies in HCI

Keywords and phrases eye tracking, sequence analysis, map task complexity

Digital Object Identifier 10.4230/LIPICs.GIScience.2018.30

Category Short Paper

Funding Swiss National Science Foundation Grant No.: 200021_162886.

1 Introduction

Maps are visual representations of spatial context that communicate geographic information and allow for spatial problem analysis [13]. The design of “better” maps is a key goal in cartography. However, the definition of “better” is vague and has been a topic of research for a long time. In his book, MacEachren provides an overview of how maps work at different levels and how design choices interact with the processing of information from a map [10].

Visual attention is a valuable source of information for cartographic design both when evaluating a map design or adapting the interface [5]. Tracking and analyzing visual attention on maps through eye tracking experiments has been proposed and used in cartography for quite some time (see [9], for an overview). Compared to other methods for evaluating a map design, such as a “think aloud protocol”, eye tracking does not introduce additional cognitive load or affect the task. Research questions that have been addressed by eye tracking experiments range from testing the differences between expert and novice map users [12], evaluating cartographic design decisions [1], or analyzing task complexity and cognitive processes [11].



© Fabian Göbel, Peter Kiefer, Ioannis Giannopoulos, and Martin Raubal;
licensed under Creative Commons License CC-BY

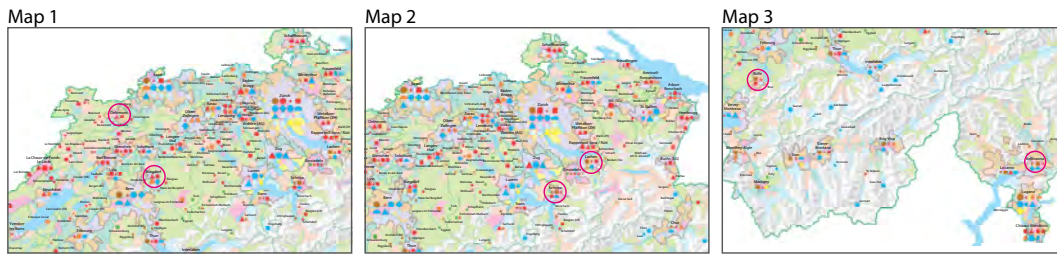
10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 30; pp. 30:1–30:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



■ **Figure 1** The three maps used in the experiment (legend excluded). The magenta circles on the maps indicate the cities that were subject of the tasks. These circles were not visible during the experiment. The map material is based on the economic map from the Swiss World Atlas¹.

Depending on the purpose of the analysis, different measures are commonly used for the analysis of gaze data collected during the interaction with maps. Some measures, such as average fixation duration [2], are not related to the map content and may provide general insights about the cognitive state of the user. Content-related measures, on the other hand, enable an analysis of which elements of the map or interface the user has paid attention to [8], thus allowing for a more detailed evaluation of the map or interface. For instance, Cöltekin et al. [3] used sequence analyses on Areas of Interest (AOI) to study individual and group differences for a geovisual analytics tasks on two different map interfaces.

In this short paper, we suggest to use compressed string analysis of eye tracking data to evaluate the impact of task complexity and different legend types on the visual attention of a user. The two gaze based legend types described in [6] and a traditional legend were tested on three different map extents. This result can help to improve map design or in the context of interactive systems, create better map interfaces, that adapt to the user's current task.

In this research we investigate whether the complexity of two map tasks can be measured and compared based on fixation sequences. In order to address this research question, we choose to analyze the mean fixation duration and perform a sequence analysis based on AOIs. The short paper is structured as follows: We first explain the experiment including an introduction to the task, the map and the legends used. Furthermore, we explain the procedure and the AOIs used. In the results section we report on average fixation duration and gaze sequences. Finally we discuss the results and provide an outlook on future work.

2 Experiment

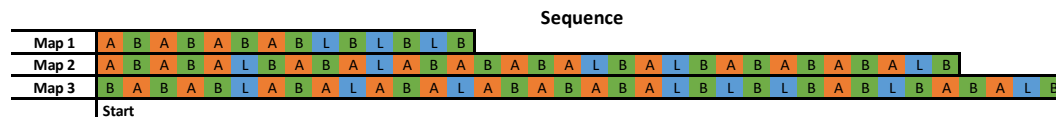
We intended to test the search behavior and interaction with a map legend while performing a common comparison task. For this we chose three maps (Figure 1) with varying symbol density and the three legend types, one traditional and two that adapt to the users' gaze as described in previous work [6]. This results in a 3×3 within-subjects design, with three maps and three legend types. Each participant performed the task on each of the three maps extents once. Map extents, legend type and ordering were counterbalanced based on a Latin square. In the following, we explain the task in more detail.

2.1 Task, Map and Legend

The task of the user was to inspect two cities (A and B) on the map, and determine and name the industries that differ between the two. Visual inspection of the legend was required in order to interpret the meaning of the differing symbols. Before starting the actual task, the location of the two cities was presented to the participant in order to avoid search and only measure task-related gaze behavior. Panning and zooming was not possible.

■ **Table 1** Number of symbols shown for the two cities that needed to be compared on the three maps. These numbers are taken as a measure for task complexity: the comparison task on Map 1 was less complex than that on Map 2, which in turn was less complex than that on Map 3.

	Number of symbols				total	different	symbol density	Visual angle between Cities
	in City A	in City B	in City A but not in City B	in City B but not in City A				
Map 1	2	4	0	2	6	2	high	5.7°
Map 2	5	5	2	2	10	4	high	9.1°
Map 3	4	6	2	4	10	6	low	34.3°



■ **Figure 2** Example sequences of one participant’s dwells on three different AOIs: the two cities whose symbol sets had to be compared (A, B) and the legend (L).

We expected the chosen approach to result in a very structured and predefined way of solving the task: first the participant looks at city A then at city B in search for symbols that differ. After finding at least one, the participant will search within the legend to determine its meaning. This structured approach allows us to break down the analysis to a sequence analysis on only three different AOIs (Figure 2).

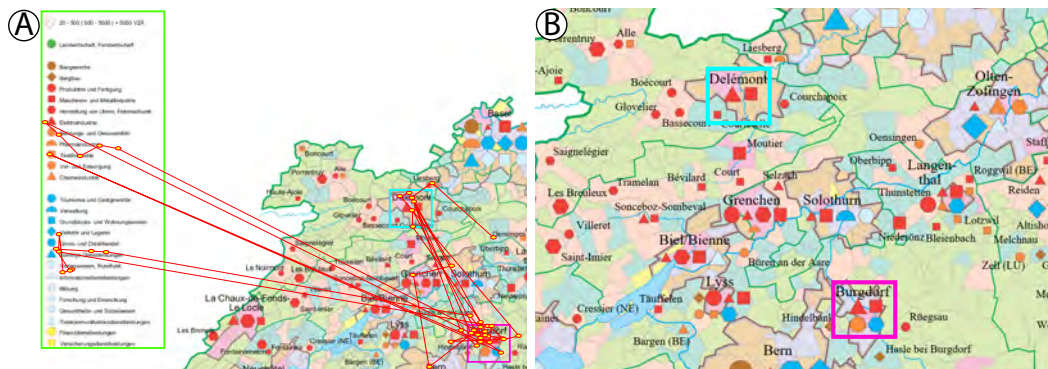
As with this study we focus more on task difficulty and not on the design of the map itself, we employed an economic map from the Swiss World Atlas which had been designed by experienced cartographers to teach geography in schools¹. We can identify four characteristics that among others, increase the search space and thus contribute to a higher task difficulty:

- Total number of symbols in a map extent
- Number of symbols per city
- Number of symbols that differ between two cities
- Distance between the cities

Based on this, we chose three map extents for our experiment (Figure 1). Map 1 and 2 feature a higher density of symbols compared to Map 3. The distance between the relevant cities is the shortest in Map 1, however, still exceeds the area that can be inspected with one fixation, followed by Map 2 and Map 3. Table 1 shows that the cities contained three to six symbols each, and that two to six differed between them. For instance, Map 1 has only two symbols that differ between the two cities. Furthermore, these symbols are all in city B. We assume that this makes the task the easiest on Map 1, followed by Map 2 and Map 3 as with them more map features differ.

The design of the legend was from the original map and a total of 26 symbols were shown (Figure 3). We tested a traditional legend and the two gaze based legend types described in our previous work [6] namely *fixed adaptive*, where the content of the legend is adapted to highlight the symbols that were visible within the last fixation on the map and *dynamic adaptive* which also adapts its placement to appear always at the bottom right position of the current field of view.

¹ <https://schweizerweltatlas.ch/en/>



■ **Figure 3** (A) shows a part of Map 1 with the AOI for the legend in green, the AOIs for City A in cyan and for City B in magenta. The scan path is highlighted in red and fixations in yellow. Figure 2 shows the result as a sequence (see first row). (B) shows the cities and symbols that need to be compared in detail.

In our previous experiment we could show that with the gaze-based legends, participants spent less task time on the legend compared to the traditional legend [6]. Here, however, we are interested in analyzing the impact of task difficulty on the gaze sequence and on usage of the legend.

2.2 Participants and Setup

18 participants (7 female) took part in our experiment with most of them having a professional background in Geomatics or Cartography. Their average age was 31.9 ($SD = 4.4$).

During the study, we collected gaze data using a Tobii TX 300 eye tracker. Additionally, we used a chin rest to keep the distance between participants and display ($23''$, 1920×1080 px) constant (60 cm). Before each run we performed a 9-point calibration.

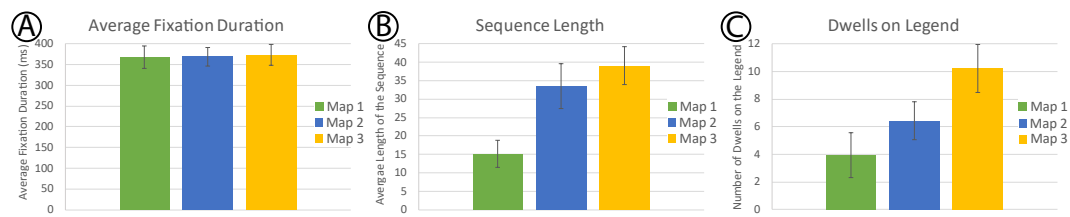
2.3 Procedure

After filling out a demographic questionnaire, participants proceeded with a test run to familiarize themselves with the given legend type. Next, a preview map without the symbols was provided to show the locations of the two cities in question. When the participant indicated that she was ready, the actual task began, however, there was no time constrain to fulfill the task. These steps were repeated three times to test all different maps and legend types. This assured that each possible combination of map \times legend was tested six times.

2.4 Area of Interest

As we are mainly interested in which sequence visual attention was spent on the map and the legend, for each task, we annotated the following three AOIs: Legend, City A and City B. In case of cities, the size of the AOIs comprised the city name and all related symbols (Figure 3). For the legend, the AOI was dynamically adapted to the size (and placement) of the legend which requires to track fixations in real time. Based on the gaze data coming at 300 Hz from the eye tracker, we used our online implementation of the I-DT algorithm first introduced in [4] to calculate fixations (80 px dispersion and 200 ms window size).

For deriving sequences from the gaze data, we denoted the fixations to AOIs in order of appearance. Consecutive fixations on the same AOI are handled as one visit, called dwell.



■ **Figure 4** Results for mean fixation duration (A), average length of sequences (B), and average dwells on the legend (C) independent of legend type. Error bars indicate the 95% confidence interval.

3 Results

First, we calculated the average fixation duration (Figure 4 (A)). This is a measure commonly related to the task difficulty [7]. However, independent of the used legend type, a one-way ANOVA ($F(2,51) = .06$, $p = .940$) could not show a statistical significant difference between the three Maps.

From Figure 4 (B) we can see that in general, sequence length (i.e. number of dwells on an AOI) is shorter for Map 1 (15.1, $SD = 8.1$) followed by Map 2 (33.5, $SD = 6.0$) and Map 3 (39.0, $SD = 5.1$). A one-way ANOVA ($F(2,51) = 22.110$, $p < .001$) confirmed statistically significant differences between the Maps. All following results are Bonferroni adjusted ($\alpha = 0.017$). Post hoc analysis with a Tukey test resulted in a significant difference between Map 1 and Map 2 ($p < .001$), and Map 1 ($p < .001$) and Map 3 but not between Map 2 and Map 3 (textitp = .318). Furthermore, a one-way ANOVA showed no significant effect of legend types onto the length of the gaze sequence (Map 1: $F(2,15) = 1.261$, $p = .312$; Map 2: $F(2,15) = .400$, $p = .678$; Map 3: $F(2,15) = 3.530$, $p = .055$).

We also counted the number of dwells on the legend (Figure 4 (C)). As this data was not normally distributed we used a Kruskal-Wallis H which confirmed statistical differences ($\chi^2 = 29.832$, $p < .001$). Following the results of the Mann Whitney U post-hoc tests shows that participants dwelt significantly less often on the legend on Map 1 (mean = 3.94, $SD = 3.67$) compared to Map 2 (mean = 6.44, $SD = 3.05$, $U = 47.0$, $p < .001$) and Map 3 (mean = 10.22, $SD = 3.84$, $U = 16.0$, $p < .001$). Also the result between Map 2 and 3 is significant ($U = 56.0$, $p < .001$). If we compare these values with the number of different symbols in Table 1, we can see a correlation between number of symbols that differ between the two cities and participants' dwells on the legend. The ratio is between 0.51 and 0.62. Again, a Kruskal-Wallis H was applied to calculate the effect of legend types onto the number of dwells on the legend. However, legend type has no statistically significant effect with Map 1 ($\chi^2 = 4.258$, $p = .119$), but it has on Map 2 ($\chi^2 = 5.984$, $p = .050$) and on Map 3 ($\chi^2 = 7.645$, $p = .022$).

Furthermore, we analyzed the sequences before the legend was visited the first time. In average 5.8 switches between City A and City B have occurred before the gaze shifted to the legend. However, we could neither find statistical significance between the different maps ($\chi^2 = .917$, $p = .632$) nor did the legend ($\chi^2 = 7.743$, $p = .021$) seem to have an impact.

4 Discussion and Future Work

Although we could not find any significant differences in the fixation duration, evaluation of the sequence length indicates that more differences of symbols, which we take as an indicator for task difficulty, result in more focus switches between cities and legend. Furthermore, we could show that the number of dwells on the legend goes in line with the number of different

symbols between two cities. The fact that the number of dwells on the legend was always higher than the number of different symbols is particularly interesting, as this suggests, participants mostly evaluated one symbol at a time, when visiting a legend and needed some more to reassure their answer.

One reason could be that the symbols consist of arbitrary shapes and colors (Figure 1). It could be that more decisive or iconic symbols are easier to remember and require less re-visits of the legend. Future work has to inspect AOI sequences in more detail. For instance, can correlation between sequences of different participants contribute to find common patterns for certain tasks? This knowledge can help to create better maps or in the context of interactive systems, create better map interfaces, that adapt to the user's current task.

References

- 1 Alžběta Brychtová and Arzu Çöltekin. An Empirical User Study for Measuring the Influence of Colour Distance and Font Size in Map Reading Using Eye Tracking. *The Cartographic Journal*, 53(3), 2016.
- 2 Henry W. Castner and Ronald J. Eastman. Eye-Movement Parameters and Perceived Map Complexity - I. *Cartography and Geographic Information Science*, 11(2), 1984.
- 3 Arzu Çöltekin, Sara I. Fabrikant, and Martin Lacayo. Exploring the efficiency of users' visual analytics strategies based on sequence analysis of eye movement recordings. *International Journal of Geographical Information Science*, 24(10), 2010.
- 4 Ioannis Giannopoulos, Peter Kiefer, and Martin Raubal. GeoGazemarks: Providing Gaze History for the Orientation on Small Display Maps. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction (ICMI '12)*, New York, New York, USA, 2012. ACM Press.
- 5 Fabian Göbel, Ioannis Giannopoulos, and Martin Raubal. The importance of visual attention for adaptive interfaces. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct - MobileHCI '16*, New York, New York, USA, 2016. ACM Press.
- 6 Fabian Göbel, Peter Kiefer, Ioannis Giannopoulos, Andrew T. Duchowski, and Martin Raubal. Improving Map Reading with Gaze-Adaptive Legends. In *ETRA '18: 2018 Symposium on Eye Tracking Research & Applications*. ACM, 2018, (to appear).
- 7 Kenneth Holmqvist and Richard Andersson. *Eye Tracking: A comprehensive guide to methods, paradigms and measures*. CreateSpace, Lund, 1 edition, 2017.
- 8 Peter Kiefer and Ioannis Giannopoulos. Gaze Map Matching: Mapping Eye Tracking Data to Geographic Vector Features. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems, SIGSPATIAL '12*, New York, New York, USA, 2012. ACM.
- 9 Peter Kiefer, Ioannis Giannopoulos, Martin Raubal, and Andrew Duchowski. Eye Tracking for Spatial Research: Cognition, Computation, Challenges. *Spatial Cognition & Computation*, 17, 2017.
- 10 Alan M. MacEachren. *How Maps Work Representation, Visualization, and Design*. Guilford Press, 2004.
- 11 Daniel R. Montello. Cognitive Map-Design Research in the Twentieth Century: Theoretical and Empirical Approaches. *Cartography and Geographic Information Science*, 29(3), 2002.
- 12 Kristien Ooms, Philippe De Maeyer, Veerle Fack, Eva Van Assche, and Frank Witlox. Interpreting maps through the eyes of expert and novice users. *International Journal of Geographical Information Science*, 26(10), 2012.
- 13 Terry A. Slocum, Robert B. McMaster, Fritz C. Kessler, and Hugh H. Howard. *Thematic Cartography and Geovisualization*. Pearson, Upper Saddle River, New Jersey, USA, 3 edition, 2009.

Facilitating the Interoperable Use of Cross-Domain Statistical Data Based on Standardized Identifiers

Jung-Hong Hong

Department of Geomatics, National Cheng Kung University, Taiwan
junghong@mail.ncku.edu.tw

Jing-Cen Yang

Department of Geomatics, National Cheng Kung University, Taiwan
jingcen@mail.ncku.edu.tw

Abstract

In the big data era, the successful sharing and integration of data from various resources becomes an essential requirement. As statistical data serves as the foundation for professional domains to report the phenomena in the reality according to the selected administration units, its importance has been well recognized. However, statistical data is typically collected and published by different responsible agencies, hence the heterogeneity of how the data is designed, prepared and disseminated becomes an obstacle impeding the automatic and interoperable use in multidisciplinary applications. From a standardization perspective, this research proposes an identifier-based framework for modeling the spatial, temporal and thematic aspects of cross-domain statistical data, such that any piece of distributed statistical information can be correctly and automatically interpreted without any ambiguity for further analysis and exploration. The results indicate the proposed mechanism successfully enables a comprehensive management of indicators from different resources and enhances the easier data retrieval and correct use across different domains. Meanwhile, the interface design exemplifies an innovated improvement on the presentation and interpretation of statistical information. The proposed solution can be readily implemented for building a transparent sharing environment for the National Spatial Data Infrastructure (NSDI).

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases Cross-Domain, Statistical Data, Standardized Codes, Visualization

Digital Object Identifier 10.4230/LIPICs.GIScience.2018.31

Category Short Paper

Funding This paper is partial result from the research project (MOST 106-2627-M-006-004) granted by the Ministry of Science and Technology in Taiwan.

1 Introduction

The recent trends of open data and big data analytics have brought a new wave of information revolution, where a tremendous number of cross-domain data is available for uses in the Internet. Since the data may be acquired from various domains and stakeholders, it comes no surprise that users have to deal with unfamiliar or even unknown data structure produced by other domains [5]. In other words, big data are highly heterogeneous [1]. Despite the technology breakthrough in terms of Internet speed and storage has been remarkable, the lack of a comprehensive design, identification and encoding strategy of distributed data is impeding the successful sharing and interpretation of cross-domain applications. Failure to



© Jung-Hong Hong and Jing-Cen Yang;
licensed under Creative Commons License CC-BY

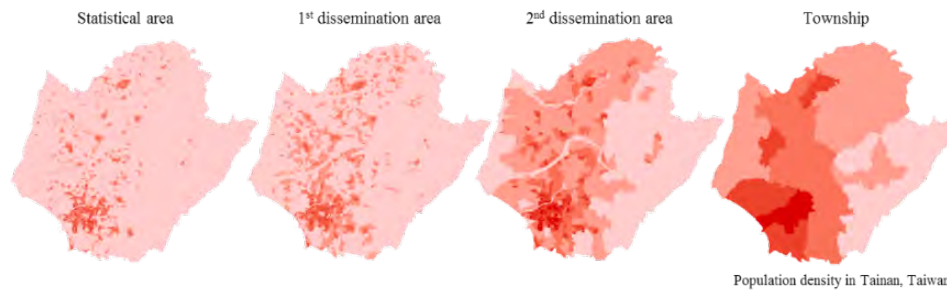
10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 31; pp. 31:1–31:7

Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



■ **Figure 1** Top four of the most meticulous level of TGSC framework.

overcome such barriers absolutely limits the feasibility of correct decision making and any further exploration. It is therefore necessary to examine how to improve the interoperability of distributed data and enhance the application intelligence of cross-domain data.

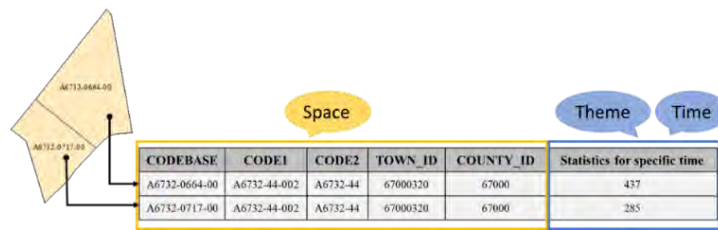
Statistics plays an indispensable role in the sustainable development for a nation. Often managed with respect to a particular level of administrative units, statistical data is typically recorded by tables or illustrated by choropleth maps. Various domains follow this space-partitioned framework to establish and update domain statistical data according to a selected frequency. The effective integration of cross-domain statistical data enables a better understanding about continuously changing reality and correct assessment of future action plans. Every country has their own space-partitioned framework for statistical units. For example, a 7-level system named Taiwan Geographical Statistical Classification (TGSC) was established in 2012 as the common references for domain agencies to publish different granularities of statistical data to suffice different application needs (Figure 1). With the development of GIS, the distribution of statistical data evolves from tables with fixed schema [3], Web-based GIS platform (<http://datashine.org.uk>) to open data [2]. The correct use of statistical data, regardless of the technology being used, requires an in-depth knowledge about the data being used and professional skill for correctly manipulating the GIS software. This requirement becomes a major obstacle after the statistical data is widely and easily available to novice users. Ignorance about the meaning behind the acquired data may easily lead to wrong decisions. Worst of all, users may not even notice they are making mistakes. An interoperable solution for correctly handling and integrating cross-domain statistical data is thus necessary. This paper proposes an identifier-based mechanism for the standardized representation of distributed cross-domain statistical data. It aims to not only simplify the interpretation and processing of statistical data, but also smartly enriches the service content with related indicators and visual aids.

2 Method

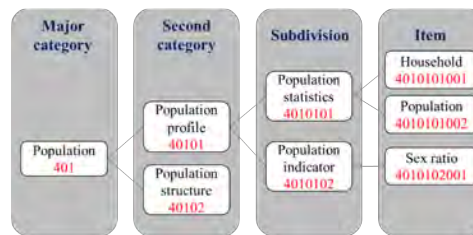
A necessary presumption when using statistical data shared by other domains is to correctly interpret its meaning. Four major approaches are adopted in this research to facilitate an interoperable sharing mechanism for overcome current exchange barriers and enrich the capability of decision making:

2.1 Standardized identifier framework

As statistical data typically uses quantitative measures to describe the phenomena for a selected geographic location (Where) from a particular theme consideration (What) at a



■ **Figure 2** Standardized identifier framework.



■ **Figure 3** Theme code structure.

■ **Table 1** Statistical method code list.

Code	Statistical method
TC	Total count
SUM	Summation
PC	Percentage
TH	Per mille
RAT	Rate
DEN	Density

given time (When), therefore these three aspects should be unambiguously modeled by unique identifiers to avoid confusion. We proposed to subdivide the attributes into two major parts, one for spatial identification and another for the temporal and thematic description of the statistical indicators (Figure 2). Every row consists of only one unique spatial attribute and a number of temporal/thematic attributes. The TGSC identifiers are directly used for representing the spatial identifiers and can be linked to its geometric representation. The theme codes from different domains are organized following a tree structure, so that every theme is given a unique identifier (Figure 3). The theme code is further extended to include the concept of the indicator (Table 1), such that 4010101001TC represents the indicator for the total count of household. The design of temporal coding system takes the time mode, time resolution, time instance and time range into consideration to ensure all temporal information can be unambiguously represented, interpreted and compared. Table 2 shows two examples. By definition, the population data of every month refer to the status at the end of the month, so we use “TI” to denote this is a time instant, “4010101002” and “TC” to indicate the data is about population and total count, and “02_201701_E” to imply the time is the last day of January, 2017. The number of deaths, on the other hand, is referred to the statistics of a period of time, so it is represented as “TP”.

■ **Table 2** Examples of standardized code.

	Population in Jan. 2017	Number of deaths in 2008
Time interpretation	Statistics at the end day of the month	Accumulated in a period
Time mode	Time instant	Time period
Standardized Code	TI_4010101002_TC_02_201701_E	TP_4010402001_TC_01_2010_0

■ **Table 3** Examples of related auxiliary indicators.

	Original indicator	Related auxiliary indicators
Statistical concept	Total population (4010101002_TC)	Average population of 2 nd dissemination area (4010101002_L4L2AVG) Standard deviation of total population of 2 nd dissemination area (4010101002_L4L2STD)
Domain knowledge	Crude mortality rate (4010406001_TH)	Mid-year population(4010101002_TC) Number of deaths(4010402001_TC)

2.2 Auxiliary indicators

For a chosen indicator, auxiliary indicators are developed for aiding the interpretation of statistical results, e.g., quality measures and spatial variation. Auxiliary indicators are automatically calculated according to the concept of the selected indicator. For example, standard deviation is automatically calculated for every indicator based on average concept; the Spatial Dispersion Index (SDI) proposed by Weng and Tsai in 2006[4] is calculated for every indicator based on the concept of total count. Every auxiliary indicator is also modeled by unique and standardized codes. The package of the chosen indicator and related auxiliary indicators enriches users' understanding about the different aspects of the acquire data without revealing the raw data. Domain providers can therefore flexibly package a set of related indicators either based on the statistical theories (e.g., average and standard deviation) or domain knowledge. Table 3 shows examples about how these two types of related indicators are designed and recorded.

2.3 Management mechanism

With the rules embedded in the coding system, the retrieval of data meeting specific requests can be easily completed by transforming the standardized identifiers. Two types of transformation rules respectively based on spatial and temporal perspectives are developed. The search for statistical data at finer or coarser levels is as easy as using the spatial transformation rule to replace the spatial identifier, while the search of time series data can be also easily completed by using temporal transformation rule to replace the temporal identifier. By registering the tables and the indicators in the data catalog, the search of requested data can be readily completed. Even if the requested data is not directly available, it still can be calculated if its formula is predefined and the required parameters are available (Figure 4).

2.4 Visualization technique

Users are prompted with an integrated interface that can simultaneously illustrate a number of related indicators with maps, tables or charts. The traditional choropleth maps are augmented by new visual aids like highlighted boundaries or spyglasses to make users aware

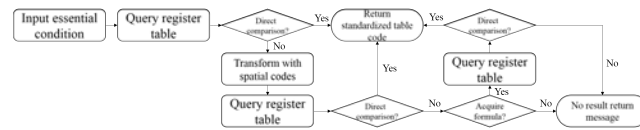


Figure 4 Searching mechanism.

Shape	CODE2	TOWN_ID	COUNTY	TP_4010406001_TH_01_2010_0	TP_4010402001_TC_01_2010_0	TI_4010101002_TC_01_2010_M
Polygon 2M	A6700-01	67000000	67000	9.10912	27	2084
Polygon 2M	A6700-02	67000000	67000	6.94844	12	1726
Polygon 2M	A6700-05	67000000	67000	9.88417	29	2084
Polygon 2M	A6700-06	67000000	67000	6.29587	20	3177

Figure 5 Subpart of mortality rate data in 2010.

Table 4 Query procedure.

Step 1	SELECT Table FROM registration table WHERE Attribute = 'TP_4010406001_TH_01_2010_0' AND Scope = '67000' AND LevelCodeVersion = 'U0202A' AND Time = '2010' Query result: Table = 'U0202A_67000_4010406_2008T2010'
Step 2	Acquire the 2 nd dissemination area of mortality rate in Tainan in 2010 SELECT TP_4010406001_TH_01_2010_0 FROM U0202A_67000_4010406_2008T2010

of the possible quality or geographic distribution issue that may otherwise not directly observable. According to users' selected indicators, the developed mechanism analyzes the results of auxiliary indicators and automatically prompts users with meaningful visual illustration.

3 Result

The yearly mortality data for the city of Tainan is chosen as the test data. Figure 5 shows a subpart of the data for the year of 2010. The search for a particular indicator starts with locating the table that includes the requested indicator from the registration table. As the example of table 4 shows, the specified constraints include "TP_4010406001_TH_01_2010_0" (the standardized code for the mortality rate in the year of 2010), "67000"(the spatial code of the Tainan city)," U0202A"(the level of 2nd dissemination area) and "2010"(time constraint). After locating the table ("U0202A_67000_4010406_2008T2010"), the system proceed to retrieve the requested data in step 2. Any statistical data stored in the database can be found in a similar way. For example, the data for one year earlier can be found by using the transformation rules to change the constraint to "TP_4010406001_TH_01_2009_0" and time constraint to "2009".

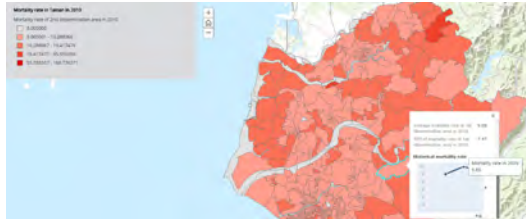
Assume that the data of the year 2011 is not directly available, it can be calculated according to the predefined formula by filling in the time constraint. As figure 6 shows, the formula for mortality rate requires the number of deaths (TP_4010402001_01_Year_0) and the mid-year population (TI_4010101002_TC_01_Year_M). The requested indicator of "TP_4010406001_TH_01_Year_0" can then be calculated accordingly (Table 5). Even if the data of the number of deaths and the mid-year population is provided by different

$$\text{Crude mortality rate} = \frac{\text{Number of deaths during a specified period}}{\text{Mid-year population}} \times 1000 \xrightarrow{\text{Standardized code}} TP_4010406001_TH_01_2010_0 = \frac{TP_4010402001_TC_01_Year_0}{TI_4010101002_TC_01_Year_M} \times 1000$$

■ **Figure 6** Use standardized codes to represent crude mortality rate.

■ **Table 5** Function of calculating crude mortality rate.

```
CalculateMortalityRate=(TP_4010402001_TC_01_2011_0/TI_4010101002_TC_01_2011_M)×1000
GenerateMortalityRate('67000','U0202A','2011')
```



■ **Figure 7** Historical mortality rate of 2nd dissemination area.

	Mortality rate	Number of deaths	Mid-year population	Average mortality rate within the next level of spatial unit	Standard deviation within the next level of spatial unit	SDI of deaths	Geometric center of deaths
CO021_TOWNS_4_CO011	TP_4010406001_TH_01_2010_0	TP_4010402001_TC_01_2010_0	TP_4010101002_TC_01_2010_M	TP_4010406001_L4L1AVG_01_2010_0	TP_4010406001_L4L1STD_01_2010_0	TP_4010402001_SDI_01_2010_0	TP_4010402001_PCX
CO021_TOWNS_4_CO011	1.12	1.12	1.12	1.12	1.12	1.12	1.12
CO021_TOWNS_4_CO011	1.12	1.12	1.12	1.12	1.12	1.12	1.12
CO021_TOWNS_4_CO011	1.12	1.12	1.12	1.12	1.12	1.12	1.12
CO021_TOWNS_4_CO011	1.12	1.12	1.12	1.12	1.12	1.12	1.12
CO021_TOWNS_4_CO011	1.12	1.12	1.12	1.12	1.12	1.12	1.12
CO021_TOWNS_4_CO011	1.12	1.12	1.12	1.12	1.12	1.12	1.12
CO021_TOWNS_4_CO011	1.12	1.12	1.12	1.12	1.12	1.12	1.12
CO021_TOWNS_4_CO011	1.12	1.12	1.12	1.12	1.12	1.12	1.12
CO021_TOWNS_4_CO011	1.12	1.12	1.12	1.12	1.12	1.12	1.12

■ **Figure 8** The package of related statistical indicators.

responsible agencies, the search mechanism can still easily find the required data as long as they are willing to comply with the rules of standardized identifiers.

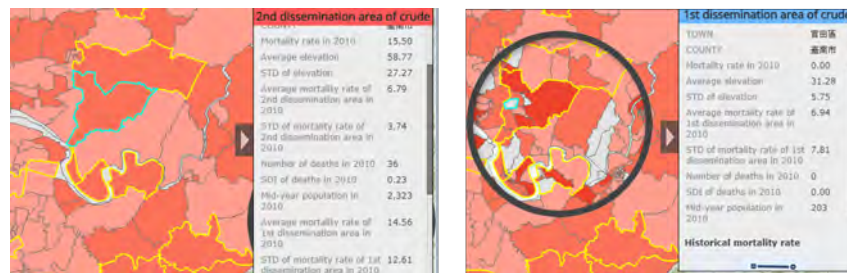
After acquiring the requested time-series data, the interface is designed simultaneously illustrate multiple aspects of indicators for easier visual inspection. Figure 7 shows the interface can show the mortality rate for the 2nd dissemination area for a single year and the historical status after users select a particular dissemination area.

In addition to the mortality rate data, auxiliary indicators related to mortality rate according to statistical model and domain demands are also available. The related auxiliary indicators include the standard deviation within the next level of spatial unit (TP_4010406001_L4L1STD_01_2010_0), SDI of deaths (TP_4010402001_SDI_01_2010_0), etc (Figure 8). Higher standard deviation usually implies a higher spatial variation within the dissemination area. The geometric center and SDI index number allow users to assess the geographic distribution of features within the dissemination area.

Based on the analysis of the auxiliary indicators, users can easily identify dissemination areas that require special attention. In figure 9, polygons with highlighted boundary imply the 2nd dissemination area with high spatial variation on mortality rate based on the analysis of its corresponding 1st dissemination area. Users can use the Spyglass tool to visually inspect the detailed geographic distribution.

4 Conclusion

In the cross-domain data sharing environment, the proposed standardized is capable of enabling the enrichment and interpretation of individual domain of statistical data, as well as the transformation, integration and visualization of cross-domain statistical data. Every



■ **Figure 9** Different levels of statistical data with spyglass interface.

individual piece of distributed statistical data in the proposed mechanism is standardized and self-described, which enables users to develop automatic processing mechanisms and reduce the tedious efforts for conquering the heterogeneity among different domains. In addition to the requested data, users are automatically provided with multiple auxiliary indicators based on the consideration of statistical theory or domain knowledge. In addition to the traditional illustration strategies of table and choropleth maps, users are prompted an innovated interface with awareness capabilities of explaining the illustrated results based on the auxiliary indicators. Based on the consensus identifier framework, the result can be further extended for distributing statistical data in the Internet in the future, e.g., data request via API-based service or Resource Description Framework (RDF).

References


- 1 Amir Gandomi and Murtaza Haider. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2):137–144, 2015.
- 2 Evangelos Kalampokis, Eftimios Tambouris, Areti Karamanou, and Konstantinos Tarabanis. Open statistics: The rise of a new era for open data? In *International Conference on Electronic Government and the Information Systems Perspective*, pages 31–43. Springer, 2016.
- 3 Corinna Koebnick, Annette M Langer-Gould, Michael K Gould, Chun R Chao, Rajan L Iyer, Ning Smith, Wansu Chen, and Steven J Jacobsen. Sociodemographic characteristics of members of a large, integrated health care system: comparison with us census bureau data. *The Permanente Journal*, 16(3):37, 2012.
- 4 Pei-Wen Weng and Bor-Wen Tsai. Spatial dispersion index: old conception, new formula. *Journal of Taiwan Geographic Information Science*, 4:1–12, 2006.
- 5 Yu Zheng. Methodologies for cross-domain data fusion: An overview. *IEEE transactions on big data*, 1(1):16–34, 2015.

Identification of Geographical Segmentation of the Rental Apartment Market in the Tokyo Metropolitan Area

Ryo Inoue

Graduate School of Information Sciences, Tohoku University, 6-6-06 Aramaki-Aoba, Aoba, Sendai, Miyagi 980-8579, Japan

rinoue@tohoku.ac.jp

 <https://orcid.org/0000-0002-0106-9777>

Rihoko Ishiyama

Graduate School of Information Sciences, Tohoku University, 6-6-06 Aramaki-Aoba, Aoba, Sendai, Miyagi 980-8579, Japan

rihoko.ishiyama.r5@dc.tohoku.ac.jp

Ayako Sugiura

Phronesis Inc., 1-14-9, Nishi-Shimbashi, Minato, Tokyo 105-0003, Japan

a.sugiura@phronesis.link

Abstract

It is often said that the real estate market is divided geographically in such a manner that the value of attributes of real estate properties is different for each area. This study proposes a new approach to the investigation of the geographical segmentation of the real estate market. We develop a price model with many regional explanatory variables, and implement the generalized fused lasso - a regression method for promoting sparsity - to extract the areas where the valuation standard is the same. The proposed method is applied to rental data of apartments in the Tokyo metropolitan area, and we find that the geographical segmentation displays hierarchal patterns. Specifically, we observe that the market is divided by wards, railway lines and stations, and neighbourhoods.

2012 ACM Subject Classification Applied computing → Economics

Keywords and phrases geographical market segmentations, rental housing market, sparse modelling, generalised fused lasso, Tokyo metropolitan area

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.32

Category Short Paper

Funding This study was supported by JSPS KAKENHI Grant Number JP18H01552 and the Tokyo Association of Real Estate Appraisers.

Acknowledgements The apartment rental data was provided to us by At Home Co, Ltd.

1 Introduction

The real estate market is segmented by many aspects, including consumer types, property types, and environmental factors. Above all, location plays a major part in market segmentation. People who prefer to live urban areas highly value accessibility to the city centre and proximity to convenient urban amenities, while people who prefer to live in suburbs value



© Ryo Inoue, Rihoko Ishiyama, and Ayako Sugiura;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 32; pp. 32:1–32:6

Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

property size and proximity to green spaces. As a result, the value of attributes of real estate properties is different in each area.

Geographic market segmentation in the real estate market has attracted much research interest, and attempts have been made to understand the area where valuation standards are the same (see Goodman and Thibodeau (2003) [1]). Previous studies presume a division structure according to specific geographical units, such as school districts, postal districts, and census tracts. However, since the real estate market has a hierarchical division structure from municipality to neighbourhood levels, they might have failed to extract the actual condition of geographic segmentation.

This study proposes a new approach to the investigation of the geographical segmentation of the real estate market. We construct a real estate price model with many regional explanatory variables that depend on different spatial resolutions, and implement the generalized fused lasso - a regression method for promoting sparsity - to extract areas where the valuation standard is the same. The proposed method is applied to the rent data of apartments in the Tokyo metropolitan area to confirm the applicability of the proposed approach.

2 Generalized Fused Lasso

The generalized fused lasso is one method of sparse modelling, which is the solution of a constrained optimisation problem that selects the substantial parameters from among many candidates.

2.1 Lasso

Lasso [2] is a method that minimises the residual sum of squares subject to a constraint on the sum of the absolute values of regression coefficients (excluding the intercept). Hence, lasso gives a solution to the constrained optimisation problem

$$\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 \quad \text{subject to} \quad \|\beta\|_1 \leq t \quad (1)$$

where \mathbf{y} is $n \times 1$ vector of the observations, \mathbf{X} is an $n \times k$ matrix of explanatory variables, β is a $k \times 1$ regression coefficient vector, and t is the positive lasso regularisation parameter. Equation (1) is equivalent to

$$\min_{\beta} \left[\frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\| \right] \quad (2)$$

where λ is a Lagrange multiplier. The optimal values of λ or t are usually determined through cross-validation.

2.2 Generalized fused lasso

Fused lasso [3] is a method to investigate the presence or absence of a difference between consecutive parameters. The optimisation problem of fused lasso imposes a new condition on the differences between consecutive parameters;

$$\min_{\beta} \left[\frac{1}{2} \left(y_i - \sum_{j=1}^k \beta_j x_i^{(j)} \right)^2 + \lambda \sum_{j=1}^k |\beta_{j+1} - \beta_j| + \gamma \lambda \sum_{j=1}^k |\beta_j| \right] \quad (3)$$

where y_i , $x_i^{(j)}$, and β_j are components of \mathbf{y} , \mathbf{X} , and β respectively. The hyperparameter γ determines the weight between the two regularisation terms.

■ **Table 1** Summary of variables.

Variable name	Mean	Standard deviation	Maximum	Minimum
Rent per square meter (yen / m^2)	3123.34	804.19	6999.50	1000
Apartment age (year)	21.63	11.99	69.17	0
Area of property (m^2)	36.33	19.03	440	10
Walking time to the nearest station (min)	6.75	4.18	60	0
Floor number	3.77	2.58	15	1
Number of rooms	1.41	0.66	8	1

Fused lasso estimates parameters whose difference to consecutive parameters tends to be zero; it can estimate common parameters. Generalised fused lasso [3] is a generalised form of fused lasso, in that it imposes constraints on differences between arbitrary neighbouring parameters. It is given by

$$\min_{\beta} \left[\frac{1}{2} \left(y_i - \sum_{j=1}^k \beta_j x_i^{(j)} \right)^2 + \lambda \sum_{(m,n) \in E} |\beta_m - \beta_n| + \gamma \lambda \sum_{j=1}^k |\beta_j| \right] \quad (4)$$

where E is a set of combinations of neighbouring parameters.

2.3 The Application of generalised fused lasso in geographical analysis

Generalised fused lasso can be applied to geographical analysis. Wang and Rodriguez (2014) [4] estimate the regional divisions of incidence rate of pediatric cancer, for example. The regularisation term that is imposed on the difference between parameters of neighbouring districts enable the authors to estimate a common parameter for them if the difference is not significant.

This study applies generalised fused lasso to the apartment rent data in the Tokyo metropolitan area to investigate the regions where the pricing of real estate properties is the same among neighbouring districts. By setting the explanatory variables that represent regions to different spatial resolutions (i.e. from a municipality level to a neighbourhood level), the analysis could identify the geographical segmentation of the market that was different to previously determined regional divisions

3 Analysis of the Rental Apartment Market in the Tokyo Metropolitan Area

3.1 Apartment rental data

This study utilises apartment rent data in the Tokyo metropolitan area for the years 2015 and 2016. It was collected by At Home Co., Ltd. High-rise condominiums whose number of floors exceed 15 are excluded as their rents have a different pricing structure compared other apartments. Consequently, the total number of records used in this study is 270,605. The data have many property attributes; the natural logarithm of rent per square meter is used as the dependent variable, and the other attributes shown in Tables 1 and 2 are set as explanatory variables.

■ **Table 2** Description of dummy variables.

Dummy name	Description	Number of variables
Railway line dummy	All railway lines are included, except dummies that are the same as some nearest station dummies	59
Nearest station dummy	All nearest stations that appear in data are included Reference: Heiwajima station	474
Cho dummy	All chos that appear in data are included Reference: Nansa-3	293

3.2 Apartment rent model

This study sets the following apartment rent model.

First, the five explanatory variables of apartment age, area of property, walking time to the nearest station, floor number, and number of rooms are used to estimate the ward (municipality)-level parameters. The Tokyo metropolitan area, which is the target area, consists of 23 wards. As such, 23 parameters are estimated for these five factors.

Next, another three different levels of location factors that affect the market are considered in this study: railway lines, nearest railway stations, and “cho” (neighbourhood). These location factors are represented by dummy variables in this model.

The apartment rent model is given by

$$\begin{aligned}
 y_i = & \beta_0 + \sum_{p_w \in P_{ward}} \sum_{w \in Ward} \beta_{p_w w}^{ward} x_{ip_w w}^{ward} + \sum_{l \in Line} \beta_l^{line} d_{il}^{line} \\
 & + \sum_{s \in Station} \beta_s^{station} d_{is}^{station} + \sum_{c \in Cho} \beta_c^{cho} d_{ic}^{cho} + \epsilon_i
 \end{aligned} \tag{5}$$

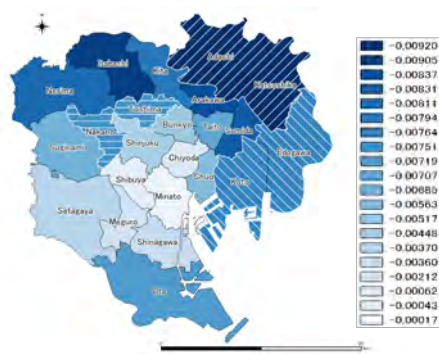
where β_0 denotes the intercept of the regression, $\beta_{p_w w}^{ward}$ denotes the ward-level regression coefficient for the explanatory variable p_w in ward w , β_l^{line} denotes the regression coefficient of the railway line dummy variable l , $\beta_s^{station}$ denotes the regression coefficient of the nearest station dummy variable s , β_c^{cho} denotes the regression coefficient of the cho dummy variable c , P_{ward} denotes a set of ward-level explanatory variables, $Ward$ denotes a set of wards in the target area, $Line$ denotes a set of railway lines, $Station$ denotes a set of railway stations, and Cho denotes a set of chos. Note that a station and a cho whose average rent per square meter are selected as the reference and dummy variables respectively, are not set for that station and cho.

The regularisation terms that impose weights on the differences between parameters of adjacent regions are set for ward-level parameters and parameters of cho dummies. If the differences between parameters of adjacent wards and chos are not significant, the common parameters would be estimated. The optimisation problem for this analysis is given by

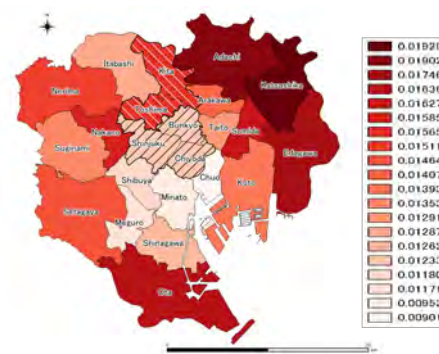
$$\begin{aligned}
 \min_{\beta} \left[\frac{1}{2} \sum_{i \in Trans} \left(y_i - \sum_{p \in P} \beta_p x_{ip} \right)^2 + \lambda \sum_{p_w \in P_{ward}} \sum_{(a,b) \in Neighbor_{ward}} |\beta_{p_w a}^{ward} - \beta_{p_w b}^{ward}| \right. \\
 \left. + \lambda \sum_{(c,d) \in Neighbor_{cho}} |\beta_c^{cho} - \beta_d^{cho}| + \gamma \lambda \sum_{p \in P} |\beta_p| \right]
 \end{aligned} \tag{6}$$

where $Trans$ is the set of all properties, $Neighbor_{ward}$ is a set of 55 combinations of adjacent wards, $Neighbor_{cho}$ is a set of 5006 combinations of adjacent chos, and λ and γ are the regularization parameters.

When solving Equation (6), numeric explanatory variables are standardised.



■ **Figure 1** Parameters of property areas.



■ **Figure 2** Parameters of floor numbers.

3.3 Results

Four settings of 0.001, 0.1, 1, and 10 for γ are tested, and the estimation with minimum AIC (Akaike's Information Criterion) value is selected. Consequently, when $\gamma = 1$, the model with 673 parameters was adopted. The adjusted coefficient of determination was 0.758.

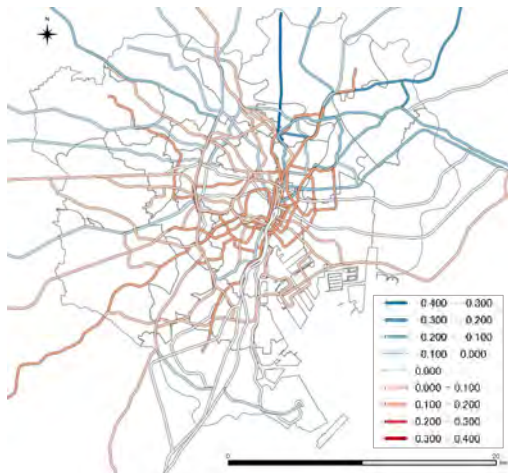
Figures from 1 to 6 show the estimated parameters. Figures 1 and 2 indicate the parameters of property areas and floor numbers. The shaded parts represent the areas with common parameters. They indicate that similar valuations for apartment attributes occur in some wards. Figures 3 and 4 illustrate that the railway lines and stations in the south-western area are valued higher than those in the north-eastern area. Above all, the apartment rents in Minato and Shibuya wards are high in central Tokyo.

Cho is set as the smallest geographical unit in this study. Figure 5 shows that many parameters are estimated to be zero. The proposed approach succeeds in selecting substantial parameters from many among candidates and reveals that apartment rents are locally homogeneous in most of areas. However, many non-zero parameters are estimated in the Minato and Shibuya wards. Figure 6 shows the Hiroo and Shirokane districts. The thick green lines indicate the ranges where the estimated parameters of cho dummies are the same. The Hiroo and Shirokane districts are famous for being two of the most exclusive residential districts in Tokyo. The results confirm that the cho-level local geographical segmentation occurs in these areas. Rent formation around Hiroo station is fragmented; different levels of rent are formed depending on the direction of properties from the station.

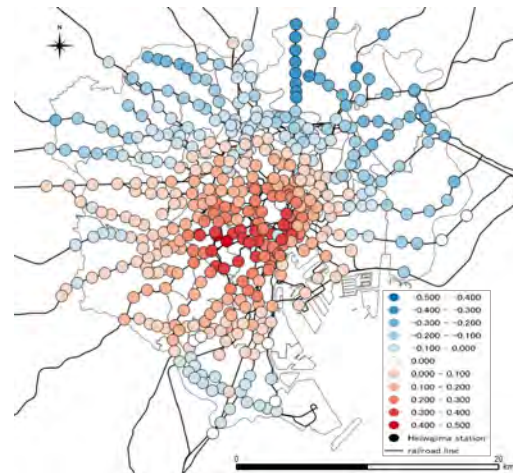
4 Conclusion

This study proposed a new approach to investigate the geographic segmentation of the real estate market. The approach consists of the price model with many regional parameters to represent the difference of price formation by region. Parameter estimation was performed by generalized fused lasso to extract substantial parameters (impose sparsity) and to search for common parameters in adjacent regions. The applicability of the approach is examined by the analysis of geographical segmentations of the rental apartment market in the Tokyo metropolitan area.

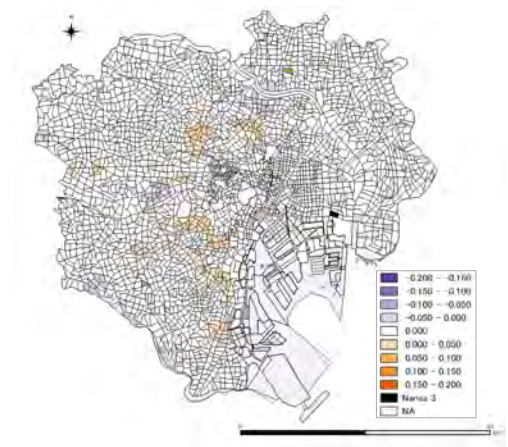
The estimated results confirmed the applicability of the proposed approach and revealed the following facts. Several adjacent wards had the same valuations for apartment attributes, the valuation on railway lines and stations was high in the south-western area, and cho-level geographic segmentation was observed, especially in the Minato and Shibuya wards.



■ **Figure 3** Parameters of railway lines.



■ **Figure 4** Parameters of railway stations.



■ **Figure 5** Parameters for chos.



■ **Figure 6** Parameters for chos around Hiroo and Shirokane.


References

- 1 Allen C. Goodman and Thomas G. Thibodeau. Housing market segmentation and hedonic prediction accuracy. *Journal of Housing Economics*, 12(3):181–201, 2003. doi:10.1016/S1051-1377(03)00031-7.
- 2 Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, 58(1):267–288, 1996. doi:10.1111/j.1467-9868.2011.00771.x.
- 3 Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fusedlasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, 67(1):91–108, 2005. doi:10.1111/j.1467-9868.2005.00490.x.
- 4 Hao Wang and Abel Rodríguez. Identifying pediatric cancer clusters in florida using loglinear models and generalized lasso penalties. *Statistics and Public Policy*, 1(1):86–96, 2014. doi:10.1080/2330443X.2014.960120.

Automatic Wall Detection and Building Topology and Property of 2D Floor Plan

Hanme Jang

Seoul National University, Department of Civil and environmental Engineering, GIS/LBS Laboratory, Seoul, Korea
janghanie1@snu.ac.kr

 <https://orcid.org/0000-0003-3895-4224>

Jong Hyeon Yang

Seoul National University, Department of Civil and environmental Engineering, GIS/LBS Laboratory, Seoul, Korea
yangjonghyeon@snu.ac.kr

Yu Kiyun

Seoul National University, Department of Civil and environmental Engineering, GIS/LBS Laboratory, Seoul, Korea
kiyun@snu.ac.kr

Abstract

Recently, indoor space construction information has been actively carried out primarily in large buildings and in underground facilities. However, the building of this data was done by only a handful of people, and it was a time- and money-intensive task. Therefore, the technology of automatically extracting a wall and constructing a 3D model from architectural floor plans was developed. Complete automation is still limited by accuracy issues, and only a few sets of floor plan data to which the technology can be applied exist. In addition, it is difficult to extract complicated walls and their thickness to build the wall-junction structure of indoor spatial information, which requires significant topological information in the automation process. In this paper, we propose an automatic method of extracting the wall from an architectural floor plan suitable for the restoration of the indoor spatial information according to the indoor spatial information standard.

2012 ACM Subject Classification Information systems → Information extraction, Computing methodologies → Image segmentation

Keywords and phrases Image Segmentation, Indoor space, Adjacency matrix, Wall thickness

Digital Object Identifier 10.4230/LIPICs.GIScience.2018.33

Category Short Paper

Funding This research was supported by a grant(18NSIP-B135746-02) from National Spatial Information Research Program (NSIP) funded by Ministry of Land, Infrastructure and Transport of Korean government.

1 Introduction

Currently, indoor spatial information is constructed for large facilities such as subways and shopping malls. However, according to [10], indoor space construction work uses a mixture of manual and automatic methods, and requires adequate financial resources and time. It is difficult to construct indoor spatial information for general buildings and facilities. To



© Hanme Jang, Jong Hyeon Yang, and Yu Kiyun;
licensed under Creative Commons License CC-BY

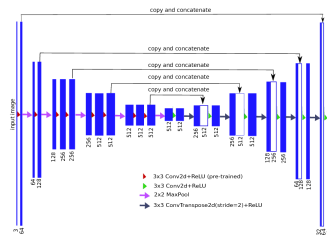
10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 33; pp. 33:1–33:5

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



■ **Figure 1** U-net.

overcome this, many studies automatically vectorize walls using 2D scanned floor plan images. In these studies, CAD files in DWG format, which are printed on paper, lose some of their attributes, and topological data are used. To express and restore these data from floor plans, studies have developed a standard format of indoor spatial data. OGC has introduced CityGML and IndoorGML, which are indoor spatial information presentation standards. Among them, CityGML has been developed for 3-dimensional modeling of urban space, while IndoorGML was proposed for indoor spatial information representation. According to [6], IndoorGML supports modeling of various viewpoints of indoor space using multi-layer and space division concepts, and it is essential to construct the node-link structure of space in order to compensate for the limitations of CityGML [5]. Therefore, the authors extracted the topological information of the wall and its thickness according to the indoor spatial information standard in this study. Lastly, the data used was provided by Korea's building information integration system.

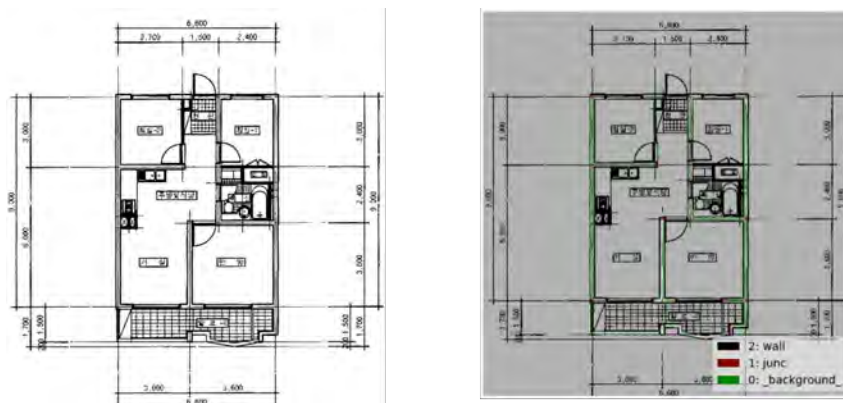
2 Related Work

Wall detection studies are based on image processing and consist of four steps. First, preprocessing is performed, where the noise of the drawing is removed. Noise, which is an auxiliary part of the data, includes numerical lines, titles, legends, etc. In the past, various filters were used to remove noise by [2], and simple neural networks have also been tested. Second, OCR is also a very important part of pre-processing, recognizing characters and replenishing the information contained in the floor plan or removing characters that may interfere with wall detection. The third step is the vectorizing process. Most algorithms deal with only straight or arc-shaped walls. Typically, [7] constructed attribute and topological information using nodes and semantic data, and only nodal points of a right angle were considered. [8] automatically generated vector drawings by applying various filters using the vertical and horizontal characteristics of the wall. [9] assumed that all the walls are straight and divided the space into rectangles of various sizes and shapes, and combined them to represent the walls, thus all the walls are represented by straight lines. The fourth is symbol recognition and is excluded from the scope of this study. At this time, it is more difficult to detect free-form walls than straight-line walls. In addition, the wall detection study shows a significant difference in performance, depending on which data are used in [1]

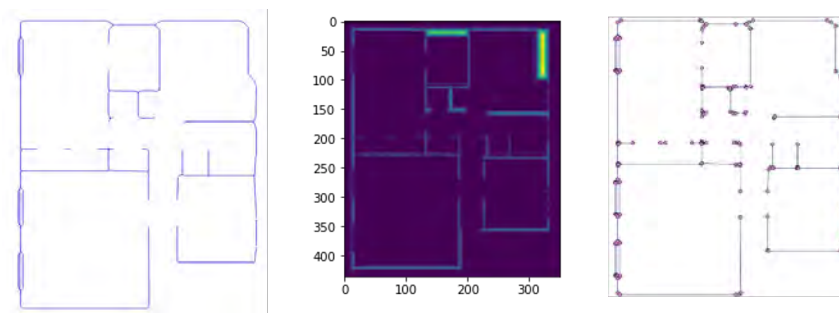
3 Method & Result

3.1 preprocessing with image segmentation

This study is divided into preprocessing, segmentation, and vectorization steps of the drawing, and preprocessing is accomplished using a deep neural network. The network used is U-net,



■ **Figure 2** Drawing and label.



■ **Figure 3** Skeleton, width of wall and node.

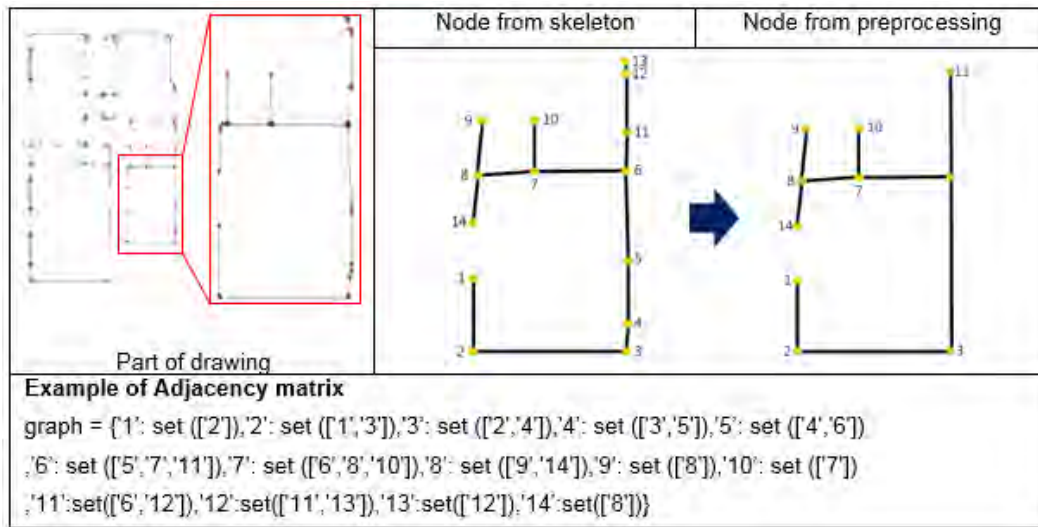
which is efficient for data augmentation; it uses context information efficiently and exhibits very accurate localization performance. U-net was selected due to its advantage of high speed and very high performance with very little data according to [3]. Since annotating floor plan data is time-consuming, few data were used. U-net was determined to be suitable for this study, and its structure is described in figure 1. To train U-net, labels 0-2 were applied to the floor plan as shown in 2. (0: wall, 1: node, 2: background)

3.2 recovering topological information

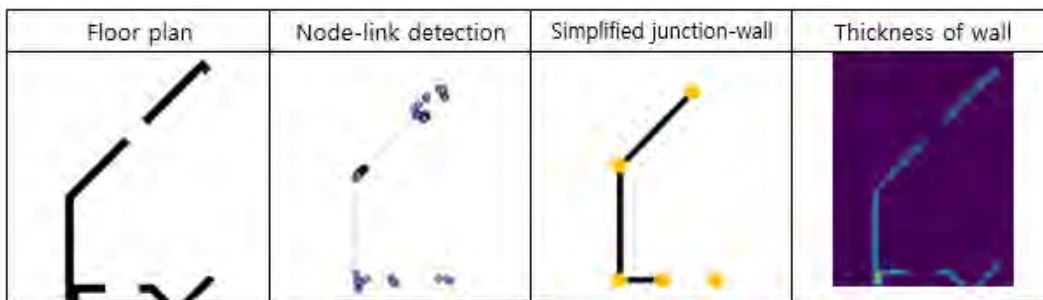
The thickness of the walls was obtained by the method of [4], and thinning was performed using the Zhang-Suen algorithm with the same data. In [11], the Zhang-Suen algorithm preserves the topological information of the wall because it provides information on the connectivity clearly, and each node of the skeleton obtained through thinning can be used as a candidate for real nodes existing on the wall. Results are described in 3.

3.3 building adjacency matrix

However, the number of extracted nodes from the skeleton tends to be overestimated compared to the actual intersections of the wall entities. Therefore, the nodes nearest to each junction were extracted separately from preprocessing as the positions of actual junctions. An adjacency matrix was constructed between junction and link, and a depth-first search was performed to simplify the graph in 4. Finally, the wall thickness value assigned to the pixels facing the detected wall was input to construct the vector data.



■ **Figure 4** Node-link simplification.



■ **Figure 5** Whole process.

Finally, the wall thickness value assigned to the pixels facing the detected wall is input to construct the vector data.

4 Conclusion

In this study, we designed an automation process that can extract information from printed architectural floor plans with missing geometric and topological information as vectors. For this purpose, image preprocessing using U-net was performed, and characters, various numerical lines, and other shapes were removed. Next, in addition to extraction of the wall thickness, skeletonization was performed to obtain connectivity information of walls and nodes as candidates of real junctions. Although the skeletonization result is composed of the skeleton link and nodes, it is difficult to identify them as the precise junction of the building. Therefore, the junctions extracted during preprocessing are considered as a guideline of the real edge of the drawing, and an adjacency matrix was created. Lastly, the thickness of the wall was added to the graph, and the link-node connectivity information of the floor plan was finally recovered. This process is described in 5. This study aimed to deal with walls placed at arbitrary angles that are not covered by existing research and is characterized by restoring wall thickness using image processing and an adjacency matrix.

5 Future Work

In this study, we constructed an adjacency matrix using links and nodes and utilized it to determine the direction of the walls and connectivity. However, the position of each node may be horizontally or vertically mispositioned. As a result, there is a disadvantage in that the rooms recovered by our method do not form rectangles (i.e., do not have four right angles). Therefore, in order to create a room in the graph with the same shape as in the actual building, it is necessary to locate each node at the correct position. In addition, in the process of inputting the thickness of the wall as an attribute of the link and the problem of changing the wall thickness while using the mode of the near pixels must be solved in future work.

References

- 1 Lluís-Pere de las Heras, Sheraz Ahmed, Marcus Liwicki, Ernest Valveny, and Gemma Sánchez. Statistical segmentation and structural recognition for floor plan interpretation. *International Journal on Document Analysis and Recognition (IJ DAR)*, 17(3):221–237, 2014.
- 2 Samuel Dodge, Jiu Xu, and Björn Stenger. Parsing floor plan images. In *Machine Vision Applications (MVA), 2017 Fifteenth IAPR International Conference on*, pages 358–361. IEEE, 2017.
- 3 Hongyang Gao, Hao Yuan, Zhengyang Wang, and Shuiwang Ji. Pixel deconvolutional networks. *arXiv preprint arXiv:1705.06820*, 2017.
- 4 Ilsoo Jeon, Namsik Won, and Kidong Bu. A preprocessing scheme of thinning capable of lines' thickness recognition for the automated vectorizing of maps. *Journal of the Korean Association of Geographic Information Studies*, 2(2):1–8, 1999.
- 5 Ki-Joune Li, Tae-Hoon Kim, Hyung-Gyu Ryu, and Hae-Kyong Kang. Comparison of citygml and indoorgml-a use-case study on indoor spatial information construction at real sites. *Journal of Korea Spatial Information Society*, 23(4):91–101, 2015.
- 6 Ki Joune Li and Ji Yeong Lee. Basic concepts of indoor spatial information candidate standard indoorgml and its applications. *Journal of Korea Spatial Information Society*, 21(3):1–10, 2013.
- 7 Chen Liu, Jiajun Wu, Pushmeet Kohli, and Yasutaka Furukawa. Raster-to-vector: Revisiting floorplan transformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2195–2203, 2017.
- 8 Rui Tang, Yuhan Wang, Darren Cosker, and Wenbin Li. Automatic structural scene digitalization. *PloS one*, 12(11):e0187513, 2017.
- 9 Wenming Wu, Lubin Fan, Ligang Liu, and Peter Wonka. Miqp-based layout design for building interiors. *Computer Graphics Forum*, 2018.
- 10 Xuetao Yin, Peter Wonka, and Anshuman Razdan. Generating 3d building models from architectural drawings: A survey. *IEEE computer graphics and applications*, 29(1), 2009.
- 11 TY Zhang and Ching Y. Suen. A fast parallel algorithm for thinning digital patterns. *Communications of the ACM*, 27(3):236–239, 1984.

Mapping Wildlife Species Distribution With Social Media: Augmenting Text Classification With Species Names

Shelan S. Jeawak¹

Cardiff University, School of Computer Science and Informatics, Cardiff, UK
JeawakSS@cardiff.ac.uk

Christopher B. Jones

Cardiff University, School of Computer Science and Informatics, Cardiff, UK
JonesCB2@cardiff.ac.uk

Steven Schockaert²

Cardiff University, School of Computer Science and Informatics, Cardiff, UK
SchockaertS1@cardiff.ac.uk

Abstract

Social media has considerable potential as a source of passive citizen science observations of the natural environment, including wildlife monitoring. Here we compare and combine two main strategies for using social media postings to predict species distributions: (i) identifying postings that explicitly mention the target species name and (ii) using a text classifier that exploits all tags to construct a model of the locations where the species occurs. We find that the first strategy has high precision but suffers from low recall, with the second strategy achieving a better overall performance. We furthermore show that even better performance is achieved with a meta classifier that combines data on the presence or absence of species name tags with the predictions from the text classifier.

2012 ACM Subject Classification Computing methodologies → Machine learning, Information systems

Keywords and phrases Social media, Text mining, Volunteered Geographic Information, Ecology

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.34

Category Short Paper

1 Introduction

The value of social media to assist in mapping and predicting geospatial phenomena has been demonstrated in areas including the occurrence of disease, social unrest, natural disasters, levels of wellbeing and characteristics of the man-made and natural environment [7, 8]. In the fields of environmental monitoring and wildlife observation there is clearly strong potential for exploiting social media, reflected in the fact that searching for named species on photo-sharing websites such as Flickr often reveals thousands of results, many of which are associated with coordinates and almost all with time stamps. It can be envisaged that these observations could complement the many effective citizen science campaigns that record aspects of the natural environment and assist environmental scientists in understanding the

¹ Shelan S. Jeawak has been sponsored by HCED Iraq.

² Steven Schockaert has been supported by ERC Starting Grant 637277.



occurrence and behaviour of animals and plants [4]. Although many mentions of species names in social media might not correspond to records of actual occurrences, several studies have confirmed the validity of significant numbers of species observations in social media [1, 2]. While these studies highlight the potential value of such data, little progress has been made to date on developing reliable automated methods for exploiting all the textual content of social media postings for tasks such as mapping species distributions.

Here we present the results of experiments to predict species distribution based on geocoded social media postings from the Flickr website. As a baseline approach we study the performance of a method that predicts the occurrence of a species in a given region if there is at least one photograph on Flickr from that region which has been tagged with the name of the species (using either its common name or scientific name). This method is then compared with a standard machine learning based text classification approach, in which all Flickr tags are used, and in which a species may be predicted to occur in a region even if no photographs in that region have been tagged with its name. For the text classifier, we follow the method from [6]. In particular, we show that the best results are obtained by a meta-classifier, which combines the prediction of the text classifier with information about the occurrence of the species name in or near the given region. These results clearly show that better distribution models can be found by taking explicit account of the occurrence of the species name as a tag, in combination with exploiting all other tags.

2 Related Work

An overview of the potential for exploiting social media in conservation and biodiversity was provided by Di Mini et al [3], who conducted a study of the use of social media platforms for posting observations of nature. The most commonly used platforms were, in order of level of sharing of nature related content: Facebook, Instagram, Twitter, Youtube, Flickr and LinkedIn. The potential of Flickr for mapping wildlife observations was illustrated by Barve [1] who mapped geotagged postings that included the scientific or common names for the Monarch Butterfly and the Snowy Owl, although that study did not conduct any systematic evaluation of the quality of the retrieved data. Daume [2] performed a manual evaluation of a sample of Twitter postings that named three invasive species (using associated photos for validation). They identified factors correlated with valid observations, such as the presence of a linked photo and tags that describe the environment (e.g. ‘leaves’ and ‘tree’). The present work exploits such associated tags in predicting species distribution. An approach to validating individual observations in Flickr was described by ElQadi et al [5] who used Google’s reverse image-search service to find photos similar to those in Flickr postings. The tags of the Google photos were then compared with those in Flickr in an attempt to filter out non-wildlife images. In our work we learn an association between all Flickr tags and the presence of particular species at a location.

The methods presented here build on the work of [6] which exploited weighted values of all tags to train an SVM (support vector machine) classifier to predict the presence of various environmental phenomena including species. In looking at species distribution no distinction was made in [6] between whether the species name was present or not and the focus was on the additional value that Flickr tags provide relative to scientific data such as climate and landcover.

3 Methodology

The objective of this paper is to find a method that can use Flickr tags for predicting the occurrence of wildlife species. To this end, we split the target spatial area into grid cells

$C = \{c_1, \dots, c_x, \dots, c_m\}$ and associate each cell with all the georeferenced Flickr tags that occur within the cell. Following [6], we use Positive Pointwise Mutual Information (PPMI) to weight how strongly tag t is associated with cell c . In particular, PPMI compares the actual number of occurrences with the expected number of occurrences (given how many tags occur overall in c and how common the tag t is). Let $f(t, c)$ be the number of times tag t (from the set of all tags T) occurs in the cell c . Then the weight $PPMI(t, c)$ is given by $\max\left(0, \log\left(\frac{P(t, c)}{P(c)P(t)}\right)\right)$ where:

$$P(t, c) = \frac{f(t, c)}{N} \quad P(t) = \frac{\sum_{c' \in C} f(t, c')}{N} \quad P(c) = \frac{\sum_{t' \in T} f(t', c)}{N} \quad N = \sum_{t' \in T} \sum_{c' \in C} f(t', c')$$

Each cell c is now represented as a sparse vector V_p , encoding the PPMI weight of all the tags in c . We assume that a training set $K \subset C$ is available which contains cells with known ground truth species observations and a testing set $U \subset C \setminus K$ containing cells whose species presence our method will try to estimate.

Our method of estimating the presence of a particular species s in cell c involves learning two classifiers *SVM1* and *SVM2*. The aim of the first classifier *SVM1* is to make initial predictions for the cells in the testing set U using the feature vector representation V_p . To give a higher confidence to tags that correspond to the name of the species, we combined the output of *SVM1* (i.e. classifier confidence score value) with information about the presence or absence of the *Common Name* or the *Scientific Name* of that species in the cell c or the neighboring cells. In particular, the cell c is now represented as a feature vector V_m which contains three features: the confidence value predicted by *SVM1*, the presence of the species actual name in c as a binary feature (being 1 if the c contains the actual name and 0 otherwise), and the percentage of neighbours that contain the species name (again as a common or scientific name) as tag. The second classifier *SVM2* is learned using the feature vector V_m to give the final estimation.

4 Experimental Evaluation

4.1 Data Acquisition

In this work we use two datasets: the ground truth species distribution from the National Biodiversity Network Atlas (NBN Atlas)³ and the geocoded social media postings from the photo sharing website Flickr⁴. The NBN is a collaborative project committed to making biodiversity information available via the NBN Atlas. This dataset covers the UK and Ireland. We used the Flickr API to collect approximately 12 million georeferenced Flickr photographs within the UK and Ireland in September 2015. However, our analysis in this paper will focus only on the tags associated with these photographs. The NBN Atlas dataset contains a total of 302 birds with at least 1000 observations, of which 200 have a name that occurs in at least 100 Flickr photographs. Among these, we have considered a random sample of 50 birds for our experiments. Note that even species with a large number of occurrences may possibly only occur in a few cells.

³ NBN Atlas occurrence download at <http://nbnatlas.org>. Accessed 19 April 2018.

⁴ <http://www.flickr.com>



■ **Figure 1** Training, Tuning, and Testing regions.

4.2 Experimental Settings and Baselines

In the experiments, we consider a binary classification problem for each of the selected birds. Specifically, the task we consider is to predict in which of the grid cells the bird occurs (i.e. for which grid cells the NBN Atlas data contains at least one observation). We test our method at three levels of granularity, considering grid cells of size 10, 20 and 30 kilometers. The set of cells C was split into two-thirds for training, one-sixth for testing, and one-sixth for tuning the SVM parameters. It is known that the quality of any supervised model is strongly affected by the way in which the data are divided. Therefore, we split the study area into geographically separated regions, as shown in Figure 1, to test the ability of our method to make predictions about geographic regions for which no observation records are given. This makes the task more challenging than choosing the cells randomly, due to possible differences between the training and testing regions. Finally, for formal evaluation we compared the results of three different methods: “Species Names” which predicts that the species occurs if its common or scientific name appears in at least one Flickr photo in the test cell, “All Flickr Tags” (*SVM1*) which uses the PPMI-based feature vector modelling all Flickr tags to train an SVM classifier using the cells in the training set and predict labels for the cells in the testing cells, and finally “Meta features” (*SVM2*) which is our proposed method, as described in Section 3.

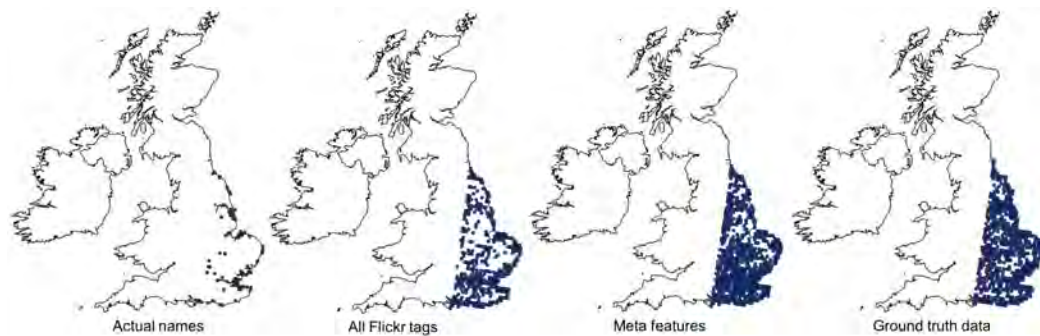
4.3 Results and Discussion

The results of predicting species distribution are reported in Table 1 in terms of the average accuracy, average precision, average recall, average F1 score, and average Area Under the ROC Curve (AUC) over the 50 birds. The results clearly show that “All Flickr Tags” significantly outperforms “Species Names”. However, the proposed meta-classifier leads to the best results overall, especially in terms of F1 score.

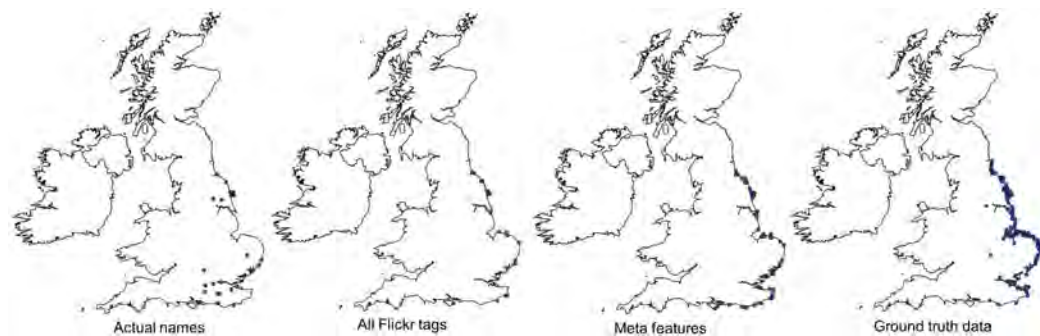
While the “All Flickr Tags” approach works well overall, we found a few cases where using only the species names led to better performance. Perhaps unsurprisingly, this is mostly the case when the number of NBN records (i.e. True labels) in the training region is low, as there may not be enough training data to effectively learn an SVM classifier in such cases. To illustrate such issues, Table 2 shows the F1 scores of 5 individual species. As can be seen, for common species such as Mallard, Dunlin, and Green Sandpiper, the “All Flickr Tags” method performs rather well. In contrast, for some less common species (or species which only occur in particular geographic contexts), such as Atlantic Puffin and Nightingale, we found better results when using the “Species name” method. Interestingly, our proposed meta classifier, which takes account of both the species presence data and the

■ **Table 1** Results for predicting the distribution of 50 species across the testing area.

Dataset	Cell Size	Accuracy	Precision	Recall	F1 Score	AUC
Species Names	10 km	0.520	0.876	0.109	0.183	0.550
All Flickr Tags	10 km	0.779	0.787	0.500	0.560	0.801
Meta features	10 km	0.825	0.820	0.603	0.637	0.850
Species Names	20 km	0.501	0.943	0.241	0.355	0.613
All Flickr Tags	20 km	0.784	0.852	0.639	0.705	0.893
Meta features	20 km	0.870	0.907	0.811	0.832	0.917
Species Names	30 km	0.567	0.970	0.384	0.515	0.684
All Flickr Tags	30 km	0.831	0.868	0.758	0.795	0.943
Meta features	30 km	0.919	0.943	0.896	0.905	0.952



■ **Figure 2** Prediction of the Dunlin distribution across the testing area with 10km grid cells.



■ **Figure 3** Prediction of the Atlantic Puffin distribution across the testing area with 10km grid cells.

all tags classification for nearby regions, outperforms both of the other methods for almost all the considered species.

Figures 2 and 3 visually illustrate the performance of our method. Note that these species (like most of the considered birds) occur in fewer than 50% of the cells, which is intuitively why the “All Flickr Tags” method is more cautious in predicting occurrence (i.e. in absence of any reason to predict occurrence, it is safer for a classifier to predict non-occurrence).

5 Conclusions and Future Work

In this paper we have presented a method for mapping the location of wildlife species occurrence using the evidence of tags from the photo sharing web site Flickr. We have shown

■ **Table 2** F1 scores for predicting the distribution of individual species using different methods.

	No.NBN records	No.Flickr photos	Cell size	Species Names	All Flickr Tags	Meta features
Mallard (<i>Anas platyrhynchos</i>)	1718823	11831	10 km	0.640	0.978	0.985
			20 km	0.899	0.974	0.986
			30 km	0.955	0.988	0.992
Dunlin (<i>Calidris alpina</i>)	278872	796	10 km	0.196	0.630	0.744
			20 km	0.346	0.920	0.969
			30 km	0.553	0.980	0.996
Green Sandpiper (<i>Tringa ochropus</i>)	103295	187	10 km	0.077	0.610	0.806
			20 km	0.195	0.849	0.955
			30 km	0.367	0.906	0.980
(Common) Nightingale (<i>Luscinia megarhynchos</i>)	24437	383	10 km	0.128	0.0	0.401
			20 km	0.326	0.0	0.705
			30 km	0.512	0.0	0.835
(Atlantic) Puffin (<i>Fratercula arctica</i>)	11551	2512	10 km	0.152	0.136	0.367
			20 km	0.173	0.359	0.518
			30 km	0.264	0.476	0.630

that while a method based simply on the presence or absence of the species name provides good precision, much better overall accuracy, with similar precision, can be achieved with a machine learning classifier that combines the presence-absence data with predictors based on all the textual tags of the photos.

One line of future work is to investigate the use of a text classifier to estimate confidence in observations of wildlife species in individual social media postings. This could be of particular value when considering postings that mention a species name but in a context that might be unrelated to its occurrence in nature.

References


- 1 Vijay Barve. Discovering and developing primary biodiversity data from social networking sites: A novel approach. *Ecological Informatics*, 24:194–199, 2014.
- 2 Stefan Daume. Mining twitter to monitor invasive alien species? An analytical framework and sample information topologies. *Ecological Informatics*, 31:70–82, 2016.
- 3 Enrico Di Minin, Henrikki Tenkanen, and Tuuli Toivonen. Prospects and challenges for social media data in conservation science. *Frontiers in Environmental Science*, 3:63, 2015.
- 4 Janis L. Dickinson, Benjamin Zuckerberg, and David N. Bonter. Citizen science as an ecological research tool: Challenges and benefits. *Annual Review of Ecology, Evolution, and Systematics*, 41:149–172, 2010.
- 5 Moataz Medhat ElQadi, Alan Dorin, Adrian Dyer, Martin Burd, Zoe Bukovac, and Mani Shrestha. Mapping species distributions with social media geo-tagged images: Case studies of bees and flowering plants in australia. *Ecological Informatics*, 39:23–31, 2017.
- 6 Shelan S. Jeawak, Christopher B. Jones, and Steven Schockaert. Using flickr for characterizing the environment: An exploratory analysis. In *13th International Conference on Spatial Information Theory, COSIT 2017, September 4-8, 2017, L'Aquila, Italy*, pages 21:1–21:13, 2017.
- 7 Philip Lei, Gustavo Marfia, Giovanni Pau, and Rita Tse. Can we monitor the natural environment analyzing online social network posts? a literature review. *Online Social Networks and Media*, 5:51–60, 2018.
- 8 Anthony Stefanidis, Andrew Crooks, and Jacek Radzikowski. Harvesting ambient geospatial information from social media feeds. *GeoJournal*, 78(2):319–338, 2013.

Multimodal-Transport Collaborative Evacuation Strategies for Urban Serious Emergency Incidents Based on Multi-Sources Spatiotemporal Data

Jincheng Jiang¹

Shenzhen University, Shenzhen Key Laboratory of Spatial Smart Sensing and Service, Smart City Research Institute, School of Architecture and Urban Planning, China

j.jiang@szu.edu.cn

 <https://orcid.org/0000-0001-5522-6910>

Yang Yue

Shenzhen University, Shenzhen Key Laboratory of Spatial Smart Sensing and Service, Smart City Research Institute, School of Architecture and Urban Planning, China

yueyang@szu.edu.cn

Shuai He

Sichuan University, Institute for Disaster Management and Reconstruction, China

shuaihe@scu.edu.cn

Abstract

When serious emergency events happen in metropolitan cities where pedestrians and vehicles are in high-density, single modal-transport cannot meet the requirements of quick evacuations. Existing mixed modes of transportation lacks spatiotemporal collaborative ability, which cannot work together to accomplish evacuation tasks in a safe and efficient way. It is of great scientific significance and application value for emergency response to adopt multimodal-transport evacuations and improve their spatial-temporal collaboration ability. However, multimodal-transport evacuation strategies for urban serious emergency event are great challenge to be solved. The reasons lie in that: (1) large-scale urban emergency environment are extremely complicated involving many geographical elements (e.g., road, buildings, over-pass, square, hydrographic net, etc.); (2) Evacuated objects are dynamic and hard to be predicted. (3) the distributions of pedestrians and vehicles are unknown. To such issues, this paper reveals both collaborative and competitive mechanisms of multimodal-transport, and further makes global optimal evacuation strategies from the macro-optimization perspective. Considering detailed geographical environment, pedestrian, vehicle and urban rail transit, a multi-objective multi-dynamic-constraints optimization model for multimodal-transport collaborative emergency evacuation is constructed. Take crowd incidents in Shenzhen as example, empirical experiments with real-world data are conducted to evaluate the evacuation strategies and path planning. It is expected to obtain innovative research achievements on theory and method of urban emergency evacuation in serious emergency events. Moreover, this research results provide spatial-temporal decision support for urban emergency response, which is benefit to constructing smart and safe cities.

2012 ACM Subject Classification Computing methodologies → Modeling and simulation

Keywords and phrases evacuation, multimodal-transport, path planning, disaster system modeling, time geography

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.35

¹ [National Natural Science Foundation of China:[Grant Numbers 41701452, 41671387, 41401444 and 91546106]]



© Jincheng Jiang, Yue Yang and He Shuai;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 35; pp. 35:1–35:8

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Category Short Paper

1 Introduction

Serious emergency events mainly refer to major natural disasters, fire disaster, explosive outbreaks, production safety accident, terrorist attack, explosion events, and so on. Tens of thousands of people died in such disasters every year. These events cause serious damages to personal and property safety, and they are principal threat to urban security. The top priority task after serious emergency events is to evacuate the crowd from the sites of accident [3]. As an important core of response plan for emergency management, emergency evacuation for serious sudden events has become a hot topic in our society [9], and it is very benefit to construct safety cities.

Emergency evacuation for serious sudden events in bustling city has its own characteristics: (1) High population density. The density of population in city is very high and the aggregation effect works during evacuation process. For example, over 300 thousands people gathered in Shanghai's Bund area for the arrival of the new year and caused stampede event in 2014. (2) Pedestrian take occupation of vehicular road. As a result, regular traffic rules don't work anymore. (3) Wide spreading. As large-scale crowd-gathering, the traffic congestion spreads out and the evacuation distance generally reaches several kilometers. (3) Evacuated objects may change their modes of transportation. (4) The evacuation statuses are highly dynamical [7]. Under this scenario, if there is no scientific and reasonable unified guidance, the crowd must be mingled with the traffic vehicles and the evacuation efficiency is very low. Thus, single modal-transport cannot meet the requirements of quick evacuations under serious sudden events [11]. A new theory is in urgent need to efficiently and synergistically invoke various transportation tools [8]. So that, the pedestrian and vehicles can be scheduled in a scientific and reasonable way to ensure the high-efficiency, safe, ordered emergency evacuation system. Undoubtedly, multimodal-transport evacuation strategies for urban serious emergency event are very meaningful and urgent required, but also full of great challenges.

Existing multimodal-transport evacuations generally indicate the pedestrian-vehicle mixture evacuation [10] [1], [5], [4]. They focused on analyzing the behavior characteristics under mixed statuses, but lack of spatiotemporal collaborative capacity. In general, the challenges of developing the efficient multimodal-transport evacuations come from two aspects: (1) For dynamic distributed people and vehicles, all walk, road vehicle and rail transit are used to evacuate the pedestrian and traffic flows in a collaborative way; (2) Under time-geography environment, multimodal-transport evacuations are constrained by the limited road resource and dynamic conditions. Traffic control, road resource allocation and route planning should be considered to minimize obstructions, maximize evacuation efficiency, minimize traffic conflicts between vehicle and pedestrian and ensure safety. These two challenges become the development bottleneck of urban emergency evacuation. With spatiotemporal dynamic evacuation task and time geography-constrained environment, multimodal-transport collaborative evacuation strategies are a difficult issue to be resolved for emergency response in serious incidents.

This paper focuses on the multimodal-transport collaborative evacuation strategies considering walk, road vehicle and rail transit under dynamic distributions of people and vehicles and time geography constraints. The research achievements could improve the spatiotemporal collaborative capacity for various transportation tools, provide space-time decision supports for emergency response in serious sudden incidents.

2 Research framework

This paper aims at constructing an effective multimodal-transport collaborative evacuation optimization model under space-time evacuated objects and time geography-constrained environment for urban serious sudden incidents. This optimization model should satisfy two requirements: in global scale, the entire evacuation system must be operated in a high-efficiency and safe way under scientific guidance; in local scale, personalized escape path and transportation modes should be provided for individuals. For this goal, we propose a research framework in which Four main parts are contained:

- Modeling the emergency environment;
- Constructing multimodal-transport collaborative evacuation optimization model;
- Solving the model and separating evacuation strategies;
- Experimental testing and assessing. More detailed contents embed in each part and their association among them are shown Fig. 1.

Generally speaking, this paper utilizes multi-source spatiotemporal data to construct static and dynamic emergency, uses multi-commodity network flow model to build multi-objectives multi-constrained emergency evacuation optimization model considering multi transport modals. Moreover, our solution contains multiple strategies, such as routing planning [6], [2], road resources distribution, dynamic flow control and transportation tool conversion, etc. As considering relative comprehensive factors, this proposed model is expected to achieve satisfactory effects and this will be tested by empirical data.

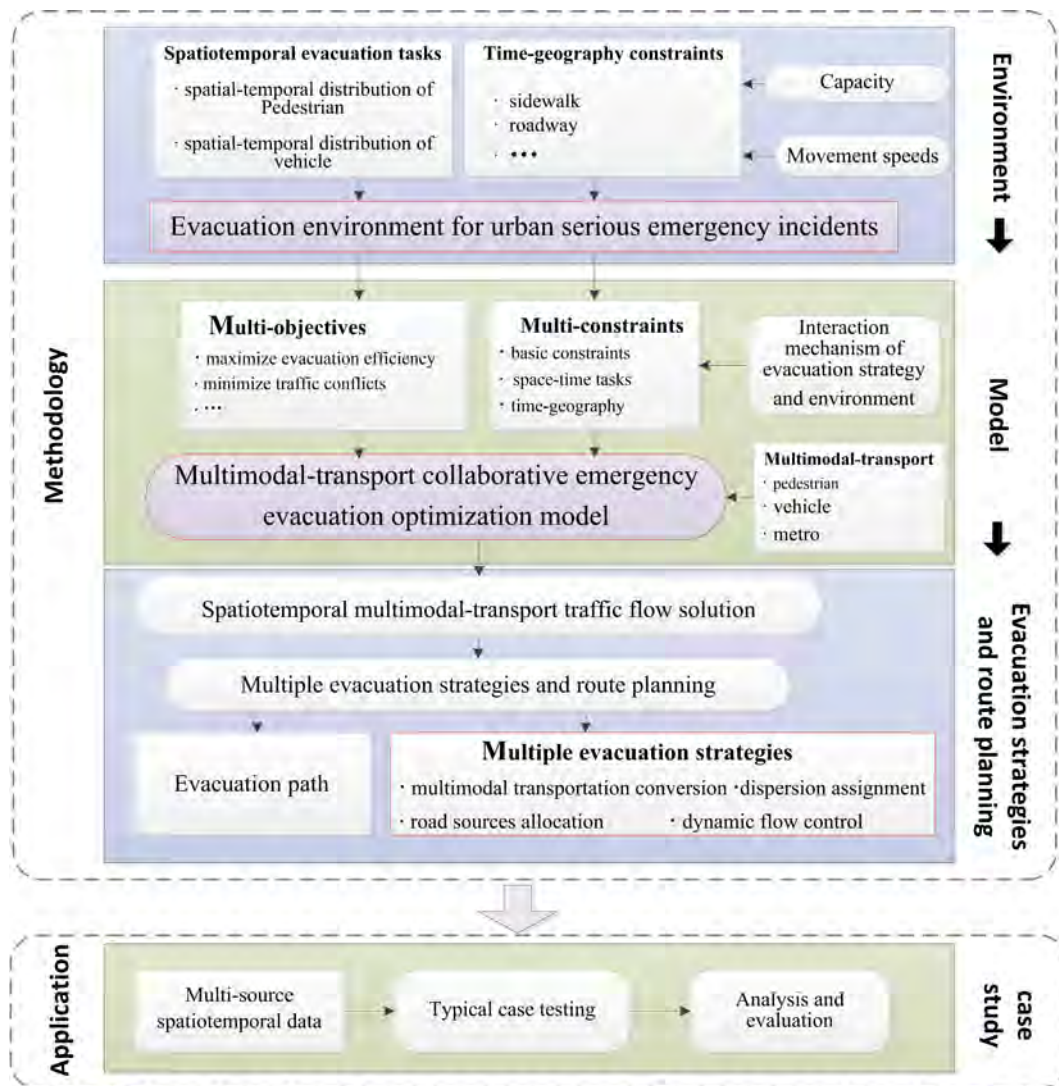
3 Study area and data source

Our study area is an entertainment Center, which is located in Shenzhen, one of the biggest cities in China. Many large-scale public events, such as concerts, sports, exhibition, were held here. Large crowds gather, and the large traffic jam ultimately emerges and last a long time. Furthermore, complex geographical environment involves 2 metro stations, 7 bus stops, four main routes and lots of bypass, squares, bridges, and so on. Thus, this study is of representativeness in metropolis. Multi-source spatiotemporal data used in this paper mainly includes:

- Foundational geographic data. This dataset can accurately static 3D geographic model including buildings, roads, overpasses, underground passages, hydrographic net, vegetation-covered area, metro stations, and so on.
- Phone cellular signaling data. For the goal in this paper, this data is mainly used to estimate dynamic population distribution.
- GPS data of bus and taxi. With the real position information of bus, we can infer how many available buses at arbitrary time moment. Massive amounts of taxi's GPS are widely used to calculate traffic congestion status by many software. Besides of traffic congestion status, this paper attempt to further figure out dynamic traffic flows.
- Smart card records for buses and subways. Smart cards record the location and time of passengers to get on and off the bus or subway. From these, we can know the maximum passenger capacity and residual available capacity.

4 Methodology

With above known emergency evacuation environment, an effective multi-objective multi-dynamic-constraints optimization model for multimodal-transport collaborative emergency evacuation is constructed in this section. As the core of the entire framework, the objective functions and constraint conditions in this model are analyzed from macro-perspective.



■ **Figure 1** The framework to study multimodal-transport collaborative evacuation strategies

Objective functions. All evacuation efficiency, security, evacuation distance for various transportation tools and many other objectives should be optimized in an integrated way.

- As the basic requirements for emergency response, the evacuation efficiency is the primary goal.
- The mixture of pedestrian and vehicles would cause chaos. It is very necessary to separate pedestrian and vehicle flows and minimize vehicle-pedestrian conflicts.
- In order to make full use of the advantages of various transportation modes, the conversion among different transportation modes should be considered. So that, short-, moderate-, and long-distance evacuations are assumed respectively by walk, road transport and rail transit. In such way, a kind of spatial-layered evacuation phenomenon emerges.
- Besides, more other objectives should also be considered, and different objectives play various roles or even contradict each other.



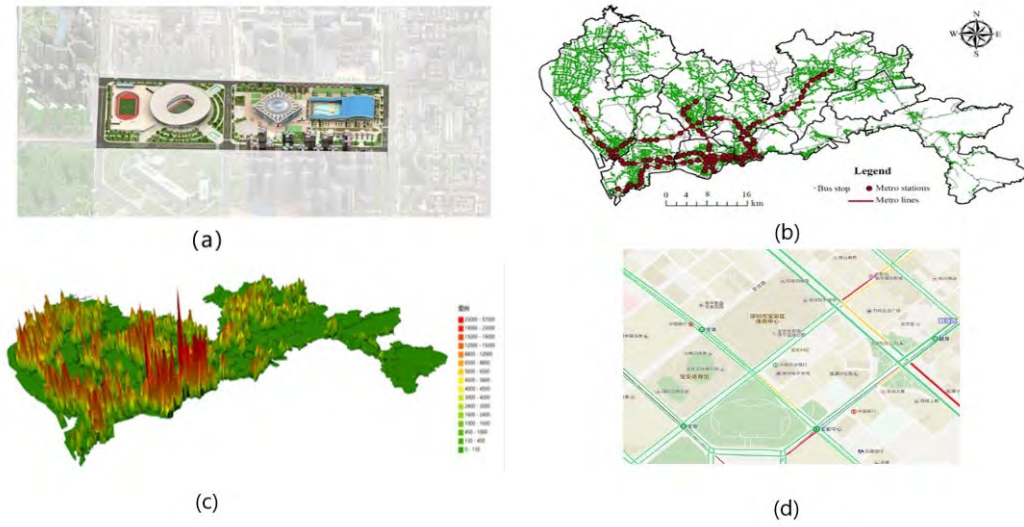
■ **Figure 2** The study area: an entertainment center in Shenzhen, China.

Constraint conditions. Numerous constraint conditions can be classified into three major categories:

- **Basic constraints.** Evacuation strategy and status (or environment) influence each other. Their interaction mechanisms can be described in mathematical forms, which is treated as a kind of constraint. These constraints include limited flow capacity of pedestrian, vehicle and metro, and the conservation of total population.
- **Constraints from space-time evacuated tasks.** At the scene of an accident, dynamic evacuation tasks are assigned to walk, car, bus and metro, etc. in real time, so that stranded population at arbitrary time and site is always less than a threshold; in transit, under the conservation of total population, conversion of various transportation modes is allowed to control traffic and pedestrian flows. This reflects the task-cooperation relations among multimodal transportations.
- **Constraints from time-geography.** The movement speeds of escapers are affected by surrounding environment, which is a kind of constraints. The capacity of a metro station is limited to receive people within a specific time interval. On the sidewalks and motorways, pedestrian and vehicle flows are not allowed to exceed their respective capacities. But, pedestrian can occupy motorways in such emergency scenario and this requires to reasonably allocate the road resources. Constraints from time-geography implies the resource competition relations among multimodal transportations.

Model Solution. Above multi-objective multi-dynamic-constraints optimization model simultaneously optimize numerous factors. As a result, lots of evacuation strategies are hidden in the solution of aforementioned emergency evacuation optimization model. Two main steps are needed to be executed to obtain the final evacuation strategies:

- **Spatiotemporal multimodal-transport traffic flow solution**
Pedestrian and vehicle can respectively occupy different lanes in the same road, but not allowed to be mixed up in a same lane. Due to the limited capacity of road, the flows of



■ **Figure 3** Emergency environment modeling from multi-source spatiotemporal data: (a) 3D emergency scenario model; (b) Spatial configuration of bus stops and the metro system; (c) population distribution; (d) dynamic traffic states.

pedestrian and vehicle are complementary:

$$f_{i,j}^p(t) \leq C_{i,j}^p(t) \times u_{i,j}(t) \quad (1)$$

$$f_{i,j}^v(t) \leq C_{i,j}^v(t) \times [1 - u_{i,j}(t)] \quad (2)$$

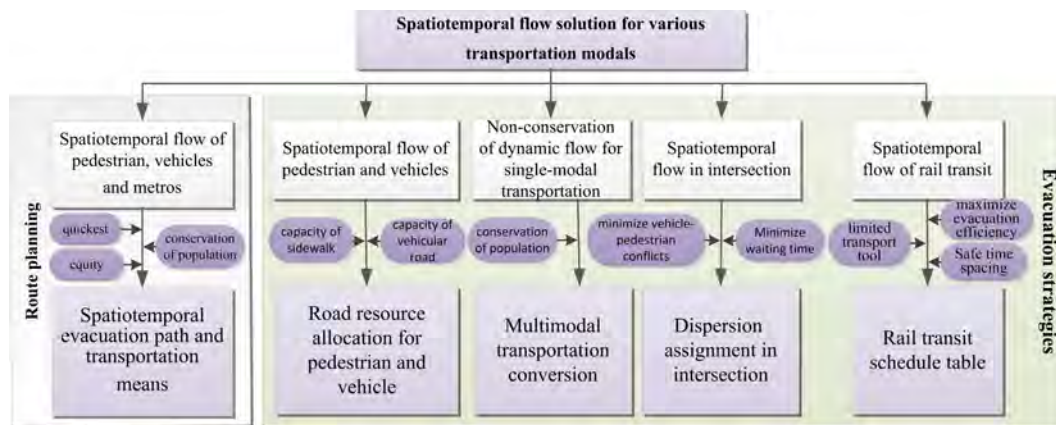
Where $f_{i,j}^p(t)$ and $f_{i,j}^v(t)$ are respectively dynamic pedestrian and vehicle flows; $C_{i,j}^p(t)$ and $C_{i,j}^v(t)$ are the road capacity for pedestrian and vehicle; $u_{i,j}(t)$ the percentages of lane occupied by pedestrian, which suggest road resource allocation. As for rail transit, it is an independent system, but share small number of nodes with sidewalks.

If pedestrian, vehicles and metros are respectively treated as three kinds of commodity flows with different behavior characteristics, then multi-commodity network flow model can be used to solve above multi-objective multi-dynamic-constraints optimization model. After these endeavors, the optimized spatiotemporal flow of pedestrian, vehicles and metros would be obtained.

- Multiple evacuation strategies and route planning In order to consider the interaction mechanism of evacuation strategy and evacuation environment, above model simultaneously optimize multiple strategies and their solutions are mixed together. These strategies mainly include road sources allocation, dynamic flow control and route planning for pedestrian and vehicles, orbital traffic scheduling, multimodal transportation conversion, dispersion assignment in intersection, and so on. In order to make evacuation strategies is more clear and available for managers to operate, it is necessary to extract them as shown in Fig. 4.

5 Conclusion

Multimodal-transport collaborative evacuation system for urban serious emergency incidents is a complex dynamic system: on one hand, the pedestrian and vehicles from large-scale public places, building, parking lots, etc. are dynamic. In order to reduce threats of sudden incidents, the pedestrian and vehicles in the site of incident should be evacuated in a quickest time; on other hand, emergency evacuation is carried out in special place and time. Restricted by the



■ **Figure 4** Multiple evacuation strategies and route planning.

limited road resources, superabundant pedestrian or vehicles could cause congestion and low efficiency of emergency evacuation. For management departments, it is necessary to figure out the optimized task allocation for multimodal transportation, road resources distribution and other evacuation strategies under both dynamic tasks and time geography constraints. The aim of improving the spatiotemporal cooperative capability of multimodal transportation is to minimize the safety risk of evacuated objects and maximize the evacuation efficiency. Thus, multimodal-transport collaborative evacuation mechanism considering dynamic tasks and time geography constraints is a key scientific issue. This paper comprehensively takes into account the external environmental impacts and interactions of internal multiple evacuation strategies to solve above issue.

For evacuated individuals, they all want to escape from dangerous places along shortest path and in a quickest way. If everyone does this, some areas could be heavily-crowded and the evacuation efficiency must be very low; while global optimal paths would sacrifice the interests of some individuals and increase their safety risks. It is necessary to find a balance point between global and individual optimum. This paper firstly figures out the global optimal spatiotemporal flows based on optimization theory, and then the individual escape path is obtained following the principle of risk-sharing.

Above two key scientific issues are respectively to ensure the efficiency, safety and equity. This paper aims at solving the challenge of spatiotemporal collaborative capacity for multimodal transportations. Its achievement can improve the theory and method of emergency response for urban serious incidents, and safety of smart city.

References

- 1 S. I. Bingfeng, Ming Zhong, and G. A. O. Ziyou. Link resistance function of urban mixed traffic network. *Journal of Transportation Systems Engineering and Information Technology*, 8(1):68–73, 2008.
- 2 Z. Fang, X. Zong, Q. Li, Q. Li, and S. Xiong. Hierarchical multi-objective evacuation routing in stadium using ant colony optimization approach. *Journal of Transport Geography*, 19(3):443–451, 2011.
- 3 Michael Frank Goodchild. *Data modeling for emergencies*. The Geographical Dimensions of Terrorism, 2003.
- 4 Muhammad Moazzam Ishaque and Robert B. Noland. Trade-offs between vehicular and pedestrian traffic using micro-simulation methods. *Transport Policy*, 14(2):124–138, 2007.


- 5 Rui Jiang and Qing-Song Wu. Interaction between vehicle and pedestrians in a narrow channel. *Physica A: Statistical Mechanics and its Applications*, 368(1):239–246, 2006.
- 6 Q. Li, Z. Fang, Q. Li, and X. Zong. Multiobjective evacuation route assignment model based on genetic algorithm. In *In Geoinformatics, 2010 18th International Conference on*, pages 1–5, 2010.
- 7 W. Li, Y. Li, P. Yu, J. Gong, and S. Shen. The Trace Model: A model for simulation of the tracing process during evacuations in complex route environments. *Journal of Transport Geography*, 60:108–121, 2016.
- 8 P. Murray-Tuite and B. Wolshon. Evacuation transportation modeling: An overview of research, development, and practice. *Transportation Research Part C: Emerging Technologies*, 27:25–45, 2013.
- 9 S. Shekhar, K. S. Yang, and V. M. V Gunturi et al. Experiences with evacuation route planning algorithms. *International Journal of Geographical Information Science*, 26(12):2253–2265, 2012.
- 10 Xin Zhang and Gang len Chang. The multi-modal evacuation system (mes) for baltimore metropolitan region. In *Intelligent Transportation Systems (ITSC), 2012 15th International IEEE Conference on. IEEE*, 2012.
- 11 X. Zheng, T. Zhong, , and M. Liu. Modeling crowd evacuation of a building based on seven methodological approaches. *Building and Environment*, 44(3):437–445, 2009.

A New Map Symbol Design Method for Real-Time Visualization of Geo-Sensor Data

Donglai Jiao¹

Nanjing University of Posts and Telecommunications, WenYuan Road/Nanjing, China

jiaodonglai@njupt.edu.cn

 <https://orcid.org/0000-0003-4578-2715>

Jintao Sun

Nanjing University of Posts and Telecommunications, WenYuan Road/Nanjing, China

sunjintao183@163.com

Abstract

Maps are an excellent way to present data with spatial components. For the large-scale geo-sensors being utilized in recent years, the map-based management and visualization of geo-sensor data have become ubiquitous. Without a doubt, managing and visualizing geo-sensor data on maps will have vastly more future applications. However, current maps typically do not support real-time communication in the Internet of Things (IoT), and it is difficult to implement real-time visualization of sensor data on a map. Map symbols are the language of maps. In this paper, we describe a new map symbol design method for geo-sensor data acquisition and visualization on maps. We refer to the sensor data visual method in supervisory control and data acquisition system (SCADA) and apply it to the design process of map symbols. Based on the traditional vector map symbol, the mapping relationship between the sensor data and the graphic element is defined in the map symbol design process. When the map symbol is rendered in the map, the map symbol is integrated into the map layer. The communication module in the map that communicates with the sensor device receives real-time sensor data and triggers a refresh of the map layer according to the mapping profile. All the methods and processes shown herein have been verified in *GeoTools*.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases Sensor, real-time visualization, Internet of Things, map symbols

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.36

Category Short Paper

1 Introduction

The Internet of Things (IoT) is emerging as a major trend shaping the development of the Information and Communication Technologies (ICT) sector[12]. The possibility of seamlessly merging the real and the virtual world through the massive deployment of embedded devices opens up new exciting directions for both research and business [5]. With the development of sensors and the gradual maturity of sensing technology, the IoT is being widely applied in industrial process monitoring, production chain management, material supply chain management, product quality control, equipment maintenance and other production processes [7]. Since the IoT is becoming an increasingly trendy topic for individuals, businesses and governments, the needs for easy-to-understand visualization focused on different sensor

¹ This work was supported by National Natural Science Foundation of China (Grant No. 41471329, 41101358).



© Donglai Jiao and Jintao Sun;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 36; pp. 36:1–36:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

state are increasing as well. Meaningful presentation and visualization are critical for IoT applications as more information is provided to consumers. These methods will also enable policy makers to convert data into knowledge, a process which is critical for helping the end user make decisions quickly [13].

Visualizing the geo-sensor data while regularly updating the presentation of the location is necessary [10]. A good way to show the information is on a map. There are many applications of sensor data visualization based on maps [9, 15, 18]. Although some GIS technologies are able to visualize real-time data[2], there is no sensor data exchange between the map and communication server. Although periodically refreshing the map is a way to visualize changing sensor data, a frequent refresh rate increases the burden of the system. In addition, an infrequent refresh rate will cause some changes to the sensor data to be ignored. Hence, there is no single, well-defined way to provide sensor data for real-time visualization through maps. The following discussion describes some of the most important design choices made in mapping between the data models of map symbols and the models required for real-time geo-sensor data visualization. In this paper, we propose to compensate for deficiencies in the methods by incorporating sensor data transmission protocol into the map symbol architecture.

2 Sensor data acquisition and visualization in IoT

Traditionally, most sensor data acquisition and visualization has been built around SCADA, which is a system for remote monitoring and control that operates with coded signals over communication channels [8, 1, 14]. In basic SCADA architectures, information from sensors is sent to RTUs (remote terminal units), which then send that information to SCADA software. SCADA software analyzes and displays the data in a Human Machine Interface (HMI) in which all the elements, such as buttons, text arrays and other objects, are represented graphically in visualization screens. In recent years, large-scope sensor arrays that are produced worldwide have been utilized. The location of the sensor data, which is commonly handled by the Geographic Information System (GIS), appears to be increasingly important, and the implementation of geographical schematics in SCADA systems has been widely accepted. Ten [16] proposed a framework to migrate a GIS database to a SCADA system in which spatial data is converted to a SVG format to appear in an HMI. Back S employ international standards from both domains to enable information exchange between the SCADA and GIS systems and then present new concepts for bridging these systems [6]. The above studies focused on how to transfer the spatial information from a GIS to a SCADA system and present it via an HMI but focused less on how to collect sensor data and perform visualization in the GIS.

For visualization of geographic objects, the map in the GIS is a “special” HMI. Cartographers design and use symbols to represent geographic features. The procedure of a map for spatial features is similar to an HMI in a SCADA system. The geographic object is abstracted to a map symbol, which is composed of graphic elements, and then the symbol is rendered on the map. The key to visualize real-time sensor data on a map is the mapping profile between the sensor data and the graphic elements in the map symbol, just as with a SCADA system.

3 Mapping profile definition between sensor data and map symbol

The traditional design principles of map symbols are based on the visual variable system[4]. Map symbols describe the different characteristics of geographical entities by the visual

variables, such as size, hue, orientation, shape, location, texture and density. According to the process in SCADA, building the mapping between the sensor data and the graphic elements in a point map symbol is the key to visualize real-time sensor data on the map. Therefore, we incorporate this mapping into the traditional point symbol model. The data collected by the geo-sensor are periodic, so the sensor data in the system is presented in the form of discrete data. According to the principle of data visualization, different data types correspond to different visualization methods. For example, finite discrete data can be directly matched to different visual variables, and infinite discrete data can be divided into limited intervals, with each interval corresponding to different visual variables.

Production rules are widely used for representing knowledge in system[17]. We examine methods for expressing the mapping as a succinct collection of production rules of the form

```
IF conditions THEN outcomes
```

There is at least one set of logical expressions in *conditions*; a logical expression defines the relationship between a parameter representing sensor data and a threshold (e.g., $Gas < 5$), and different expressions are joined by logical operators (*not, and, or*). Outcomes are defined as *visual variable = value*. As an example, consider:

```
Parameter Gas
IF Gas >5 THEN TY1.fill=rgb(255,255,255)
IF Gas <=5 THEN TY1.fill= rgb(255,0,0)
end parameter
```

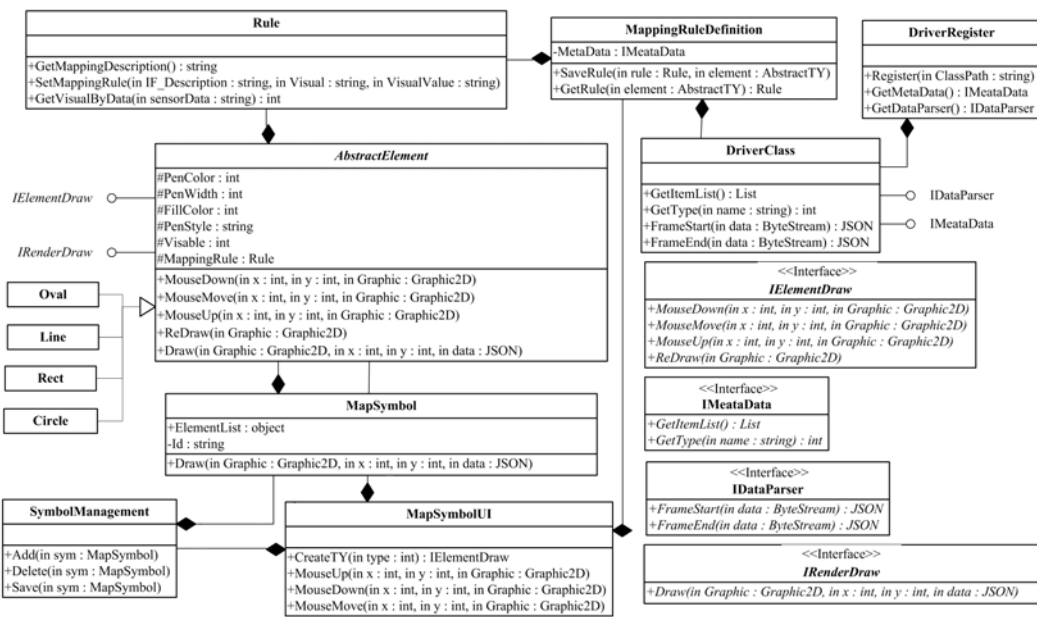
In this sample, the *rect* graphic element whose *id* is *TY1* in gas station map symbol notation corresponds to the sensor *Gas*, and the fill color will change to $rgb(255,0,0)$ if *Gas* is less than 5 ton.If *Gas* is more than 5, the fill color will change to $rgb(255,255,255)$.

4 The Map Symbol Architecture for sensor data real-time visualization

4.1 Driver Interface oriented sensor data transmission protocol

Through the network, sensor data is transmitted from the sending side to the server side. At the transmitter, sensor data is serialized into a data stream (a frame data) according to a certain sequence or organization mode. After receiving the data stream on the server, the data were deserialized in the same sequence or organization mode. The agreement of data organization is called the data transmission protocol[11]. In the design process of map symbols, establishing the mapping relationship between the sensor data in protocol and the visual variable is the key step. Therefore, the user needs to obtain the metadata information of the sensor data in the process of map symbol design, such as data type, data name, data length, data precision and so on. In the map render process, the sensor data transmitted from sending side should be converted into an open data format for data visualization. We define the metadata interface (*IMeataData*) and data-parsing interface (*IDataParser*) for data transmission protocol. The metadata interface can obtain the name and type of the item in the sensor data that is used for the design of the map symbol. The data-parsing interface takes action on server side, and transform the data stream from private format into public format. JSON is a lightweight text data exchange format[3]. We take JSON as a public format data description.

The protocol designers program the driver class, which implements the two interfaces (*IMeataData*, *IDataParser*). On one hand, the map symbol designer does not care about



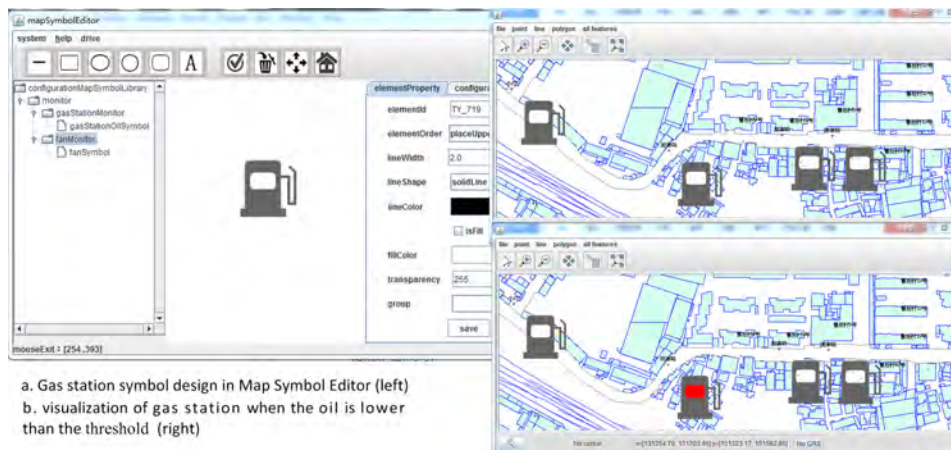
■ **Figure 1** Outline of the model of the map symbol.

the structure of the transmission protocol. The metadata information of the sensor data can be obtained by *IMeataData* and used for mapping definition. On the other hand, in the process of the map render, the geo-sensor data frame converted to JSON format data by *IDataParser*, and then the JSON format data is used for real-time visualization.

4.2 Model of map symbol for sensor data real-time visualization

Traditionally, a sensor device was abstracted into a point symbol (graphics block) shown on a map. The graphic element is the basic component of a map symbol. From the point of view of object-oriented modeling (programming), each type of graphic element includes visual variables as properties. The functions of the graphic element can be generalized into two types: graphic design and map symbol render in map visualization. We design the two type functions separately into two interfaces (*IElementDraw*, *IRenderDraw*). *IElementDraw* contains the methods needed for the graphic design, such as mouse up, mouse move, mouse down and redraw. *IRenderDraw* is mainly for map rendering, which includes the method to invoke when the map is rendered. The abstract class of graph element (*AbstractElement*) which implements the two interfaces (*IElementDraw*, *IRenderDraw*) is defined in the model. All properties of each type of graphic element in a map symbol are inherited from the abstract class. In the process of designing the map symbol, *IMeataData* in the driver class show the sensor metadata information to the map symbol designer. The designer defines the mapping of the sensor data item and visual variables, and saves it. In the process of map rendering, the graphic rendering function (*IRenderDraw*) maps the sensor data into visual variables by the *Rule* class.

MapSymbolUI binds the *IElementDraw* interface and the mouse operation in the drawing area, which makes the symbol model and UI integrated. Users can choose different types of graphic elements, and use the mouse event in the drawing area to draw the symbol element, and save it into the current symbol data model. The outline of the model show as Figure 1.



■ **Figure 2** Map symbol design and application (Take the gas station as an example).

5 The application of geo-sensor data real-time visualization by map symbol

The current GIS is a component-based system, and the different components are coupled together through an interface. The components in GIS associated with map visualization are the layer component and the symbol component, which are coupled through a rendering interface. We add the real-time sensor data acquisition module in the layer component when the map symbols are combined with it. The data acquisition module connects with the sensor through a “long-polling” connection. When the module receives data from the sensor, the module calls the parsing interface (*IDataParser*) in the driver class to transform the received sensor data into JSON format data, and then the JSON data is forwarded to the symbol render interface (*IRenderDraw*) via a layer component. According to the mapping profile, the symbol-rendering interface changes the visual variables and then realizes the real-time sensor data visualization based on the map symbol.

In this paper, we developed a new map symbol editor in the JAVA language (Figure 2a) based on the model (Figure 1) and use *GeoTools* to verify it. We use a gas station as an example, the new map symbol editor designs a gas station symbol (Figure 2a). When the oil of the gas station is below the threshold, the rectangle box of the map symbol is changed to a red filled circle (Figure 2b).

6 Conclusions

We have described the design and implementation of the map symbol for real-time sensor data visualization on the map. We have identified aspects in the map symbol that are needed to implement real-time visualization of sensor data. These aspects include the following:

- Define how the sensor data can be mapped to the visual variable of map symbols.
- Develop a new map symbol design system oriented real-time geo-sensor data visualization.
- Verify the real-time visualization by map symbol in *GeoTools*.

The present research focuses on how to achieve real-time visualization of sensor data on a map. At present, there are only a few types of graphic elements in the symbol system, and the change of graphic elements is relatively simple. In the future, we hope to design a variety of graphic elements and design more diverse graphic elements that change according to the mapping profile.

References


- 1 Qiu B.Gooi H B. Web-based SCADA display systems (WSDS) for access via Internet. *Ieee Transactions on Power Systems*, 15(2):681–686, 2000. doi:10.1109/59.867159.
- 2 ESRI. Mapping The Internet Of Things, 2018. Online; accessed 29 January 2018. URL: <https://learn.arcgis.com/en/arcgis-book/chapter9/>.
- 3 Soliman M.Abiodun T.Hamouda T.Zhou J.Lung C H. Smart Home: Integrating Internet of Things with Web Services and Cloud Computing. In *IEEE 5th International Conference on Cloud Computing Technology and Science, CloudCom 2013, Bristol, United Kingdom, December 2-5, 2013, Volume 2*, pages 317–320, 2013. doi:10.1109/CloudCom.2013.155.
- 4 Garlandini S.Fabrikant S I. Evaluating the Effectiveness and Efficiency of Visual Variables for Geographic Information Visualization. In *9th International Conference on Spatial Information Theory, Aber Wrac'h, FRANCE, SEP 21-25, 2009*, pages 195–211, 2009.
- 5 Miorandi D.Sicari S.De P F.Chlamtac I. Internet of things: Vision, applications and research challenges. *Ad Hoc Networks*, 10(7):1497–1516, 2012. doi:10.1016/j.adhoc.2012.02.016.
- 6 Back S.Kranzer S B.Heistracher T J.Lampoltshammer T J. Bridging SCADA Systems and GI Systems. In *1st IEEE World Forum on Internet of Things, WF-IoT 2014*, pages 41–44, 2014.
- 7 Bandyopadhyay D.Sen J. Internet of Things: Applications and Challenges in Technology and Standardization. *Wireless Personal Communications*, 58(1):49–69, 2011. doi:10.1007/s11277-011-0288-5.
- 8 Molina F J. Barbancho J. Luque J. Automated Meter Reading and SCADA application for wireless sensor network. In *2nd International Conference on Ad-Hoc Networks and Wireless, Montreal, Canada, October 8-10, 2003*, pages 223–234, 2003. doi:10.1007/978-3-540-39611-6_20.
- 9 Simek M.Mraz L.Oguchi K. *SensMap: Web Framework for Complex Visualization of Indoor and Outdoor Sensing Systems*. IEEE, 2013.
- 10 Stampach R.Kubicek P.Herman L. Dynamic Visualization of Sensor Measurements: Context Based Approach. *Quaestiones Geographicae*, 34(3):117–128, 2015. doi:10.1515/-quageo-2015-0020.
- 11 Al-Fuqaha A.Guizani M.Mohammadi M.Aledhari M.Ayyash M. Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications. *Ieee Communications Surveys and Tutorials*, 17(4):2347–2376, 2015. doi:10.1109/comst.2015.2444095.
- 12 Gubbi J.Buyya R.Marusic S.Palaniswami M. Internet of Things (IoT): A vision, architectural elements, and future directions. *Future Generation Computer Systems-the International Journal of Escience*, 29(7):1645–1660, 2013.
- 13 Gubbi J.Buyya R.Marusic S.Palaniswami M. Internet of Things (IoT): A vision, architectural elements, and future directions. *Future Generation Computer Systems-the International Journal of Grid Computing and Escience*, 29(7):1645–1660, 2013.
- 14 Aydogmus Z.Aydogmus O. A Web-Based Remote Access Laboratory Using SCADA. *Ieee Transactions on Education*, 52(1):126–132, 2009. doi:10.1109/te.2008.921445.
- 15 Herman L.Reznik T. *Web 3D Visualization of Noise Mapping for Extended INSPIRE Buildings Model*. Springer-Verlag Berlin, 2013.
- 16 H B Ten C.Wuergler E.Diehl H J.Gooi. Extraction of Geospatial Topology and Graphics for Distribution Automation Framework. *Ieee Transactions on Power Systems*, 23(4):1776–1782, 2008. doi:10.1109/tpwrs.2008.2004835.
- 17 Stefanuk V L.Zhozhikashvili A V. Productions and rules in artificial intelligence. *Kybernetes*, 31(5-6):817–826, 2002. doi:10.1108/03684920210432790.
- 18 Liang S H L.Huang C Y. GeoCENS: A Geospatial Cyberinfrastructure for the World-Wide Sensor Web. *Sensors*, 13(10):13402–13424, 2013. doi:10.3390/s131013402.

How Do Texture and Color Communicate Uncertainty in Climate Change Map Displays?


Irene M. Johannsen

Department of Geography, University of Bonn, Meckenheimer Allee 172, D-53115 Bonn, Germany
irene.johannsen@uni-bonn.de

Sara Irina Fabrikant

Department of Geography, University of Zurich, Winterthurerstr. 180, CH-8057 Zurich, Switzerland
sara.fabrikant@geo.uzh.ch
 <https://orcid.org/0000-0003-1263-8792>

Mariele Evers

Department of Geography, University of Bonn, Meckenheimer Allee 166, D-53115 Bonn, Germany
mariele.evers@uni-bonn.de
 <https://orcid.org/0000-0001-7767-6058>

Abstract

We report on an empirical study with over hundred online participants where we investigated how texture and color value, two popular visual variables used to convey uncertainty in maps, are understood by non-domain-experts. Participants intuit denser dot textures to mean greater attribute certainty; irrespective of whether the dot pattern is labeled certain or uncertain. With this additional empirical evidence, we hope to further improve our understanding of how non-domain experts interpret uncertainty information depicted in map displays. This in turn will allow us to more clearly and legibly communicate uncertainty information in climate change maps, so that these displays can be unmistakably understood by decision-makers and the general public.

2012 ACM Subject Classification Information systems → Geographic information systems, Human-centered computing → User centered design, Human-centered computing → Contextual design, Human-centered computing → Empirical studies in visualization, Human-centered computing → Visualization design and evaluation methods

Keywords and phrases uncertainty visualization, empirical study, visual variables, climate change

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.37

Category Short Paper

Funding This work is based on the unpublished MSc thesis by the first author advised by the subsequent authors. It was partially supported by the Canton of Zurich, Switzerland.

Acknowledgements We would like to thank the Geographic Information Visualization and Analysis (GIVA) group at the Geography Department of the University of Zurich for their feedback and methodological support. Special thanks go to Annina Bruegger for her expertise and time to prepare this manuscript in LaTeX. I would particularly like to thank Prof. Evers for supporting this research and Prof. Sara Irina Fabrikant for guiding my empirical study, providing me with the opportunity to be a part of the GIVA group, and to publish my MSc thesis research. Finally, we are grateful for the many people who participated in our study without whom we would not have been able to write this paper.



© I. Johannsen, S. I. Fabrikant, and M. Evers;
licensed under Creative Commons License CC-BY

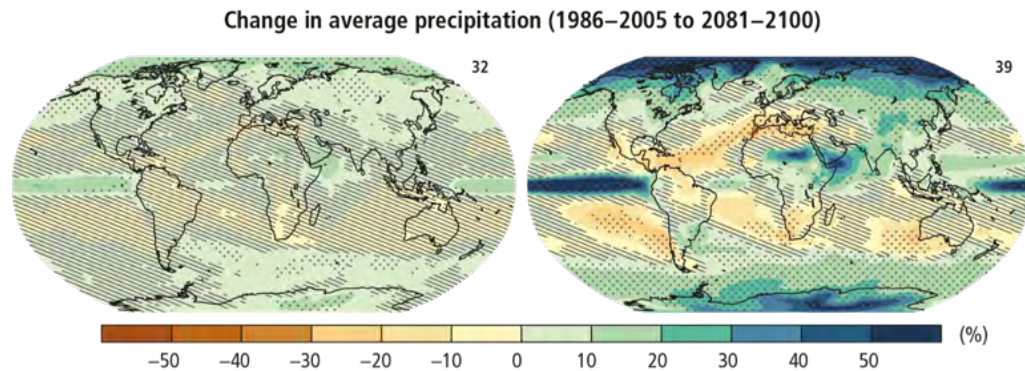
10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 37; pp. 37:1–37:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



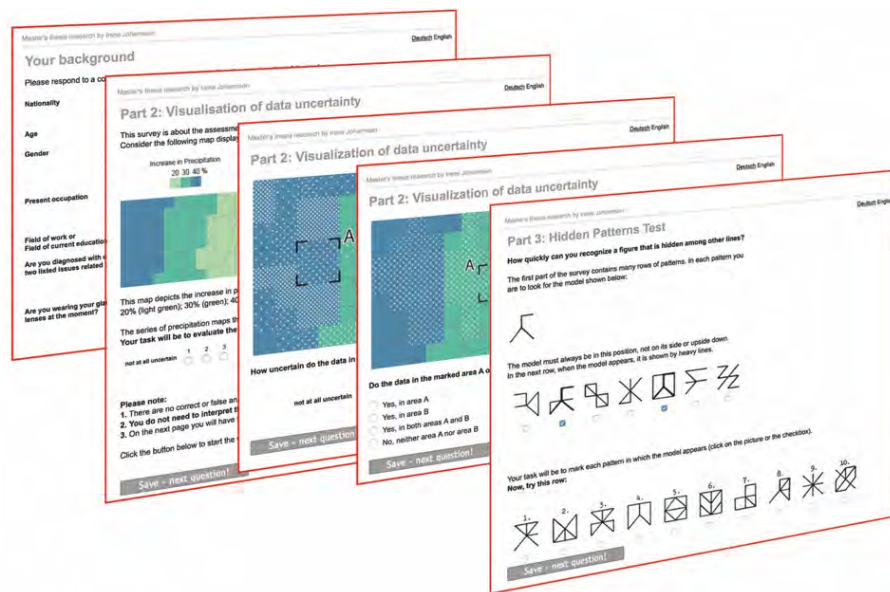
■ **Figure 1** Thematic map conveying climate change predictions using color value combined with color hue to communicate average changes in precipitation. The visual variable texture, including stippling (black dots) and hatching (diagonal lines) visualizes prediction uncertainties (Source: [5]: Figure SPM.7).

1 Introduction

Maps are a popular means to inform decision-makers and the general public about climate change. For example, well-known and highly cited reports produced by the Intergovernmental Panel on Climate Change [5], the European Environment Agency (EEA 2017), and the US National Climate Assessment (e.g., [14]) contain on average at least one thematic map every dozen pages to make climate change visible and tangible to everyone (Figure 1). Important decisions on climate change mitigation and adaptation are often made with the help of such maps [15]. Climate change predictions contain various sources and types of uncertainties. This information is also visualized in the earlier mentioned climate change reports, as to alert decision-makers and the public of the inherent prediction uncertainties (Figure 1). For instance, the numbers printed in the upper right corner above the two maps in Figure 1 describe the number of model outcomes used to compute the depicted average change in precipitation over the depicted period. The stippling texture (dot pattern) in these maps indicate regions where the projected change is large compared to natural internal variability (i.e., greater than two standard deviations of internal variability in the 20-year averages), and where 90% of the models agree on the sign of change. The hatching texture (diagonal line pattern) in Figure 1 shows regions where the projected change is less than one standard deviation of the natural internal variability in the 20-year averages (WGI Figure SPM.8, 3Figure 1.20, Box 12.1). The visualization of complex and difficult to interpret climate change statistics, including the inherently difficult to comprehend concept of uncertainty can lead to uninformed (at worst, wrong) decisions and respective harmful consequences. It is therefore critically important that climate change maps clearly and legibly communicate the information, so that these displays can be unmistakably understood by the decision-makers.

2 Background

The visualization of uncertainty has been empirically studied by a diverse visualization community for over 20 years [13]. GIScientists, for instance, have investigated the suitability of various visual variables for the communication of uncertainty in maps [8]. Particular attention has been paid, for instance, to how color value ([12, 16]) and texture [10, 9] might intuitively communicate uncertainty information in thematic maps. Empirical study results

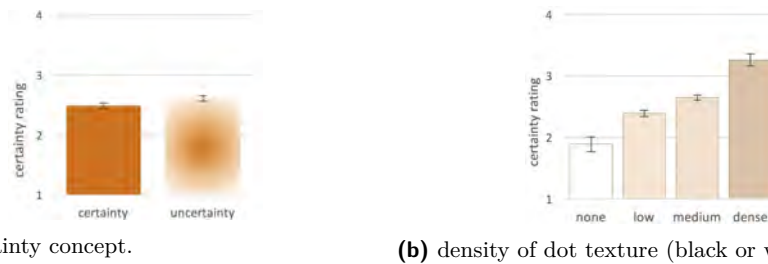


■ **Figure 2** Entire procedure of the online study, and respective sequence and style of test stimuli (administered in English and in German), showing both black and white textures, two question types, and the response box.

to date suggest that the graphic variable color value is particularly intuitively understood and associated with uncertainty [12]. This research also provides empirical evidence that the graphic variable texture, as shown in Figure 1, is particularly easy to read [11, 10, 16]. However, empirical findings are contradictory on how color value and texture intuitively communicate uncertainty, when the concept is labeled differently, i.e., uncertainty or certainty [10, 12], herein labeled un|certainty. In the following, we report on an empirical, online study that aims to narrowing mentioned research gaps.

3 Empirical Study

We systematically examine how the visual variables color value and texture [1] are intuitively understood by non-domain-expert map readers to convey un|certainty information in climate change maps. We also wish to further develop GIScience theory, focusing on the widely known, but little empirically evaluated cartographic principles “darker-is-more” and “denser-is-more”, typically used to convey increasing data magnitudes, and how these principles apply to the intuitive understanding of the visualization of un|certainty. For this, we specifically developed a new uncertainty visualization method which simulates color value by means of regularly spaced white and black dot textures of varying dot densities. This method not only combines the intuitively understood properties of the graphic variable color value to convey uncertainty [12], but it is also directly based on the graphic variable texture, which previous empirical uncertainty visualization research suggests to be highly legible [12, 16]. We were inspired by the halftone technique, a classic reprographic method to simulate continuous tone by means of a dot pattern, varying either in dot size or in dot spacing [17]. We thus employed black and white dot patterns of various densities to lighten or darken areas in the classed, univariate precipitation change maps used as stimuli in our study (Figure 2). Our developed map stimuli were directly inspired by the maps



■ **Figure 3** Main Effects: un|certainty concept (a) and visual variable texture density (b).

available in the IPCC Report 2014, as shown in Figure 1. To control for potential perceptual confounds, we carefully checked map stimuli against color deficient viewing simulations [7] and by running biologically inspired vision models [6] to assure consistent center-surround contrasts across all stimuli. The online study had three sections (Figure 2); comprising of a background questionnaire (Part 1), two types of map-based questions (Part 2), and the Hidden Patterns Test (Part 3), a standardized spatial abilities test [4], deployed via an online survey (i.e., onlineumfragen.com). We collected data during July 14-27, 2017, targeting various international GIScience/cartography, geography and geomatics lists. Participants could choose to complete the test either in German or in English. We retained 104 participants for data analysis (52 females and males), because they completed the entire test (Total $N=799$, completion rate of 13%). Based on the background questionnaire, our participants have mostly a geography, cartography, and geomatics background (approx. 40% of the total sample), but are considered non-climate-domain experts. After completing the background questionnaire and a warm up trial, participants were then asked to rate on a 4-step response scale matching the four depicted dot densities (within subject factor: density) how un|certain (between-subject factor: question type) the labeled zones highlighted on a series of maps, looked to them. In the second map-based portion of the study, participants were also asked to compare two precipitation maps that differed in dot color black|white (within-subject factor: color) of the newly developed uncertainty visualization method. Finally, participants completed the Hidden Patterns test to assess their visuo-spatial abilities.

4 Results

To compare ratings across the un|certainty conditions, we assigned “not at all un|certain” to rating 1 and “very” un|certain to rating 4. We then linked the word pairs “very uncertain” with “not at all certain” to compare the ratings across un|certainty conditions. We ran mixed ANOVAs on the ratings, and where data assumptions were violated, we relied on the Aligned Rank Transform (ART) [18]. Interestingly, textures that are labeled uncertain (Figure 3a), on average, receive significantly higher certainty ratings, compared to those that are labeled certain ($F(1,102) = 8.877$, $p < .01$, partial $\eta^2 = .08$).

Participants associated the increased density of the dot textures (Figure 3b) with increased certainty ($F(3,306) = 40.026$, $p < .001$, partial $\eta^2 = .28$). All textured zones are rated more certain compared to the non-textured zones ($\bar{x} = 2.39$, $F(1,102) = 49.13$, $p < .001$, partial $\eta^2 = .33$). The differences between the increasing texture densities are all significant ($p < .001$). There were no significant differences comparing the color of the dot texture (white vs. black dots). We also did not find any significant differences between participants’ expertise with climate change mapping and their spatial abilities relating to the Hidden Patterns test.

5 Discussion

In contrast to our hypotheses, based on above cited uncertainty visualization research, the response pattern shown in Figure 3b is the same, whether participants rate uncertainty or certainty. In other words, we find empirical support for the basic cartographic principle the more (denser) the (dot) texture, the more participants associate this with more certainty in precipitation change maps. In doing so, we replicate similar studies using different types of textures (hatching, dot size, dot arrangement, and color) to convey uncertainty [12, 3, 16]. This is somewhat in contradiction with the cartographic principle “the darker-is-more”, assumed with color value. The denser (more certainty) a white (dot) texture on the dark map background, the lighter it appears. However, [12] found that the progression from a light color shade or from light appearing fuzziness (i.e., more uncertainty) to a dark or solid color shade (i.e., more certainty) was amongst the top three most intuitively understood visual variables to convey uncertainty. To our surprise, the color of the dots (white vs. black) did not make a significant difference in our collected certainty ratings. One explanation for this unexpected result is that the developed dot textures possibly appeared too coarse as to produce distinguishable (just noticeable) differences in color value across the white and the black dot conditions. Looking into participants’ open responses in the comments response box, it seems that a significant portion of them interpreted the dots in the textures to mean precipitation measurement locations. With an increase of the precipitation sampling locations within a zone, a plausible conclusion could thus mean an increase in data certainty.

6 Conclusion and Outlook

We set out to empirically assess whether the well-known cartographic principles “darker-is-more” and “denser-is-more” also applied to the intuitively understood visualization of data un|certainty. Our empirical findings suggest that the increase of regularly spaced dot textures in precipitation change maps are indeed associated with perceived increase in data certainty. This association pattern is stable whether or not the term uncertainty or certainty are used to label the textures in the map displays. However, certainty ratings increase significantly when the term uncertainty is used in the maps, compared to when the texture is labeled certainty. Unexpectedly, the color of the dot texture has no significant influence on un|certainty ratings in our study. While we varied the spacing of the regular dot textures, others have also varied the arrangement of textures (e.g., [2, 8]). This invites like-minded researchers to further systematically investigate dot arrangements in future empirical studies. In closing, we hope to have shed further light on how the popular visual variables texture and color value might be employed to clearly and legibly communicate uncertainty information in climate change maps, so that these displays can be unmistakably understood by decision-makers and the general public.

References

- 1 J. Bertin. *Sémiologie Graphique: Les Diagrammes – les Réseaux – les Cartes*, Mouton, Paris (English translation of the second French edition by William J. Berg, 1983). Editions de l’Ecole des Hautes Etudes en Sciences Sociales, 1967.
- 2 N. Boukhelifa, A. Bezerianos, T. Isenberg, and J. D. Fekete. Evaluating sketchiness as a visual variable for the depiction of qualitative uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2769–2778, Dec 2012. doi:10.1109/TVCG.2012.220.


- 3 L. Cheong, S. Bleisch, A. Kealy, K. Tolhurst, T. Wilkening, and M. Duckham. Evaluating the Impact of Visualization of Wildfire Hazard upon Decision-making under Uncertainty. *International Journal of Geographical Information Science*, 30(7):1377–1404, 2016. doi:10.1080/13658816.2015.1131829.
- 4 R. B. Ekstrom, J. W. French, H. H. Harman, and D. Diran. *Manual for Kit of Factor-Referenced Cognitive Tests*. Educational Testing Service, Princeton, N.J, 1976.
- 5 IPCC. *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. IPCC, Geneva, Switzerland, 2014.
- 6 L. Itti, C. Koch, and E. Niebur. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.
- 7 B. Jenny and Nathaniel V. Kelso. Designing Maps for the Colour-vision Impaired. *Bulletin of the Society of Cartographers*, 40(1):9–12, 2007.
- 8 C. Kinkeldey, A. M. MacEachren, and J. Schiewe. How to Assess Visual Communication of Uncertainty? A Systematic Review of Geospatial Uncertainty Visualisation User Studies. *The Cartographic Journal*, 51(4):372–386, 2014. doi:10.1179/1743277414Y.0000000099.
- 9 M. Kunz, A. Grêt-Regamey, and L. Hurni. Visualization of Uncertainty in Natural Hazards Assessments Using an Interactive Cartographic Information System. *Natural Hazards*, 59(3):1735–1751, 2011. doi:10.1007/s11069-011-9864-y.
- 10 M. Leitner and B. P. Buttenfield. Guidelines for the Display of Attribute Certainty. *Cartography and Geographic Information Science*, 27(1):3–14, 2000. doi:10.1559/152304000783548037.
- 11 A. M MacEachren. Visualizing Uncertain Information. *Cartographic Perspective*, 13:10–19, 1992. doi:10.1.1.62.285.
- 12 A. M. MacEachren, R. E. Roth, J. O’Brien, B. Li, D. Swingley, and M. Gahegan. Visual Semiotics & Uncertainty Visualization: An Empirical Study. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2496–2505, 2012. doi:10.1109/TVCG.2012.279.
- 13 J. Mason, A. Klippel, S. Bleisch, A. Slingsby, and S. Deitrick. Special issue introduction: Approaching Spatial Uncertainty Visualization to Support Reasoning and Decision Making. *Spatial Cognition and Computation*, 16(2):97–105, 2016. doi:10.1080/13875868.2016.1138117.
- 14 J.M. Melillo, T.C. Richmond, and G. W. Yohe. *Climate Change Impacts in the United States: The Third National Climate Assessment*. U.S. Government Printing Office, Washington D.C., USA, 2014.
- 15 I. Neverla and M. S. Schäfer. Einleitung: Der Klimawandel und das Medien-Klima. In I. Neverla and M.S. Schäfer, editors, *Das Medienklima: Fragen und Befunde der Kommunikationswissenschaftlichen Klimaforschung*, pages 9–25. Springer, Wiesbaden, Germany, 2012. doi:10.1007/978-3-531-94217-9.
- 16 D. P. Retchless and C. A. Brewer. Guidance for Representing Uncertainty on Global Temperature Change Maps. *International Journal of Climatology*, 36(3):1143–1159, 2016. doi:10.1002/joc.4408.
- 17 A.H. Robinson. *Elements of Cartography*. John Wiley & Sons, New York, NY, 1995.
- 18 J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins. The Aligned Rank Transform for Nonparametric Factorial Analyses Using only ANOVA Procedures. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, pages 143–146, New York, NY, 2011. ACM. doi:10.1145/1978942.1978963.

An Analytical Framework for Understanding Urban Functionality from Human Activities

Chaogui Kang¹

School of Remote Sensing and Information Engineering, Wuhan University, 129 Luoyu Road, Wuhan, China


cgkang@whu.edu.cn

 <https://orcid.org/0000-0002-0122-9419>

Yu Liu²

Institute of Remote Sensing and Geographical Information Systems, Peking University, 5 Yiheyuan Road, Beijing, China

liuyu@urban.pku.edu.cn

 <https://orcid.org/0000-0002-0016-2902>

Abstract

The intertwined relationship between urban functionality and human activity has been widely recognized and quantified with the assistance of big geospatial data. In specific, urban land uses as an important facet of urban structure can be identified from spatiotemporal patterns of aggregate human activities. In this article, we propose a space, time and activity cuboid based analytical framework for clustering urban spaces into different categories of urban functionality based on the variation of activity intensity (*T*-fiber), mixture (*A*-fiber) and interaction (*I*- and *O*-fiber). The ability of the proposed framework is empirically evaluated by three case studies.

2012 ACM Subject Classification General and reference → General literature

Keywords and phrases Urban functionality, Human activity, STA cuboid, Spatiotemporal distribution, Clustering

Digital Object Identifier 10.4230/LIPICs.GIScience.2018.38

Category Short Paper

1 Introduction

Human activities and urban functionality are strongly intertwined. As stated in [3], “Land use typically refers to the distribution of activities across space, including the location and density of different activities, where activities are grouped into relatively coarse categories, such as residential, commercial, office, industrial and other activities”. It implies that different land use types inherently demonstrate distinct patterns of activity density and intensity [12], which are their most intuitive characteristics and can be both aggregate and temporal.

The interconnection between land use and urban activity, on the one hand, enables the generation, allocation and prediction of urban activities in space and time. For instance,

¹ National Key Research and Development Program of China (No. 2017YFB0503604), National Natural Science Foundation of China (No. 41601484), China Postdoctoral Science Foundation (No. 2015M580666 and 2017T100569), Fundamental Research Funds of the Central Universities of China (No. 2042016kf0055) and Open Research Fund of State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing of Wuhan University, China (No. 15S01).

² National Key Research and Development Program of China (No. 2017YFB0503604) and National Natural Science Foundation of China (No. 41625003).



© Chaogui Kang and Yu Liu;
licensed under Creative Commons License CC-BY

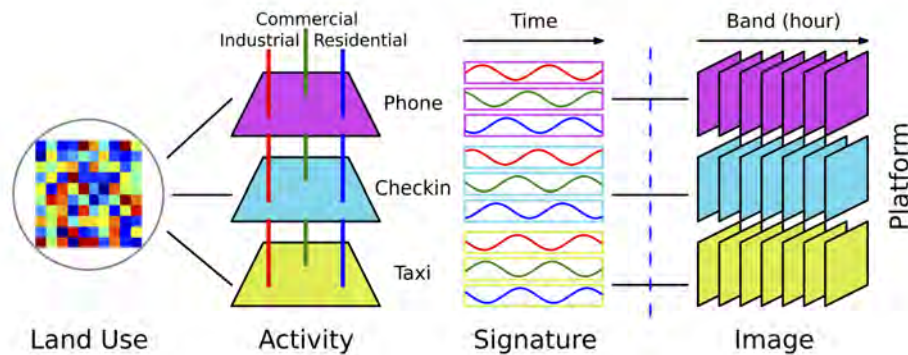
10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 38; pp. 38:1–38:8

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



■ **Figure 1** Urban land use inference as an analog to remote sensing.

there exists an abundant body of literature in urban modeling relying on the link between the two [1]. On the other hand, urban activity and its spatiotemporal dynamics can be regarded as good proxies of urban land use distributions. The usage of a specific urban space depends on those who occupy it and when, and what they do. These factors, in turn, constitute of a unique signature of the given urban space or land use zone [10]. Yet, an universal analytical framework for unraveling the relationship between urban land use types and their associated signatures of human activity is still missing.

Earlier research attempts show that utilizing urban activities to identify land use types can be analog to remote sensing for geographical classification [7]. As shown in Figure 1, the signature of changes of activity intensity along time can be taken as spectral characteristics of remote sensing for differentiating geographical objects. Recent advances in land use inferences and urban space segmenting based on urban activities follow this scheme in general. The procedure can be summarized as: (1) building feature vectors of urban spaces (e.g., places and regions) based on the variations of human activities in a predefined temporal granularity (e.g., hours of weekdays and weekends); (2) classifying urban spaces into different land use types (e.g., residential, commercial, leisure) based on the similarity between their feature vectors using mainstream clustering algorithms. Due to this fact, traditional classification approaches used in remote sensing are naturally adapted for the purpose, for instance Principal Component Analysis [10], K-means [8], Supervised Classification [11], and just name a few. However, it is still an open and interesting question that how to build the signature from the spatiotemporal dynamics of human activity to inform stakeholders the characteristic of the underlying urban spaces.

In this article, we will show readers a proposed space, time and activity cuboid based analytical framework for understanding the functionality of urban spaces based on human mobility data. With the cuboid, different activity signatures are derived for urban land use inference from the perspectives of the variation of activity intensity, mixture and interaction. The proposed framework is applied in three case studies based on three types of activity datasets (i.e., bus ridership, taxicab ridership and metro ridership) in different geographical regions. For each of the three types of activities, we build, normalize and cluster the signatures of each urban zone using different approaches. The results are accessed by the ground truth land use map as an evaluation of the capabilities of each combination of activity, feature, normalization and clustering algorithm.

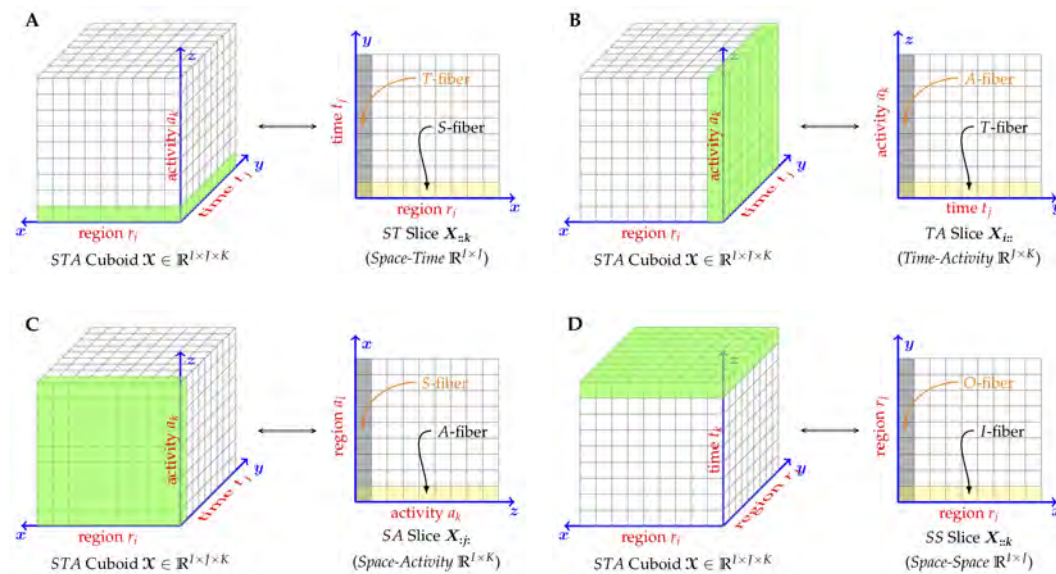


Figure 2 The STA cuboid of human activity in space and time. In A, B and C, the 3-dimension tensor \mathcal{X} consists of I regions, J time slots and K features (or activities). In addition, S -norm, T -norm and A -norm define the normalizations of fibers $\mathbf{x}_{:,jk}$ (i.e., fixing time and activity), $\mathbf{x}_{i,:k}$ (i.e., fixing space and activity), $\mathbf{x}_{ij,:}$ (i.e., fixing space and time), respectively. Intuitively, S -fiber $\mathbf{x}_{:,jk}$ quantifies the spatial distribution of a given activity a_k at the given time t_j , T -fiber $\mathbf{x}_{i,:k}$ the temporal signature of a given activity a_k in the given region r_i , and A -fiber $\mathbf{x}_{ij,:}$ the mixture of distinct activities in a give region r_i and at a given time t_j . In principle, S -norm captures the relative intensity of activities in different regions (**volume**), T -norm the fluctuations of activity intensity along time (**shape**), and A -norm the component of activities of a region (**texture**). Additionally, if the interaction between spatial regions can be observed, a new tensor \mathcal{X} consists of I regions ($I = J$) and K time slots is built in D. Under this scenario, SS -norm captures the flow patterns between each pair of regions along time (**network**), and I -fiber $\mathbf{x}_{:,jk}$, O -fiber $\mathbf{x}_{i,:k}$ quantify the inflow and outflow of human mobility in the region, respectively.

2 A space, time and activity cuboid based analytical framework

For urban land use inference, we concentrate on three dimensions as *Space*, *Time* and *Activity* (STA) and propose a cuboid representation of the three dimensions as shown in Figure 2. In the cuboid, the *Space* dimension denotes the I distinct regions which are usually regular grids across space; the *Time* dimension represents the J different time slots; and the *Activity* dimension contains the K types of activities. Therefore, the proposed STA cuboid quantifies the distributions of human activities in space and time.

In practice, we usually observe individuals' diverse activities (large K) in fine spatial and temporal granularities (large I and J) with the assistance of the increasing availability of user-centric geospatial data. If fixing activity a_k , we can obtain a ST slice demonstrating the spatial distributions of the given activities along with time (Figure 2A). In the ST slice, each row is a S -fiber of the distribution of the given activity a_k in space at the given time slot t_j . Whereas, each column is a T -fiber of the signature capturing the fluctuations of intensities of activity a_k in region r_i at time t_j . In a similar way, we can obtain a TA slice if fixing the location (region) of interest (Figure 2B). The TA slice delineates the intensities of various types of activities $\{a_1, \dots, a_K\}$ and their fluctuations along time $\{t_1, \dots, t_J\}$ within the given region r_i . In the TA slice, each row is a T -fiber while each column is a A -fiber of the

component of different types of activities in the given region r_i and time slot t_j . Fixing time t_j , a SA slice demonstrates how the different types of activities distribute across space, and its rows are A -fibers and columns S -fibers. Additionally, if the interaction between regions can be observed, a new tensor \mathcal{X} consists of I regions ($I = J$) and K time slots is built (Figure 2D). Therefore, SS -norm captures the flow patterns between each pair of regions along time. Fixing time t_k , a SS slice demonstrates how the different regions interact with each other in space, and its rows are inflow I -fibers and columns outflow O -fibers.

Based on the fibers (i.e., activity signatures) derived from the space-time-activity tensor, we then relate urban land use and human activity from three distinctive perspective. Considering that the signatures are organized as time series, the clustering approach is adopted to assign urban spaces into different categories of urban functionality based on the similarity of their signatures in terms of the variation of activity intensity (i.e., the T -fiber), the component of activity type (i.e., A -fiber) and the pattern of spatial interaction (i.e., I - and O -fiber). Note that, in addition to the signature, different normalization method and clustering algorithm can result in different classification of urban land use types. To be concise, hereafter we will concentrate on the activity signature and discuss the normalization and clustering method briefly.

3 Applications of the framework for urban functionality inference

3.1 Clustering based on the variation of activity intensity (T -fiber)

Using a seven-day taxi trajectory data set collected in Shanghai, we investigate the temporal variations of both pick-ups and drop-offs, and their association with different land use features. For each hour in the seven days, we compute the numbers of pick-ups and drop-offs for each $1 \text{ km} \times 1 \text{ km}$ cell in the study area as the activity signatures. Two T -dimensional vectors, denoted by \mathbf{V}^{pickup} and $\mathbf{V}^{dropoff}$, can be constructed to represent the temporal variations of trips for each pixel i in the study area as

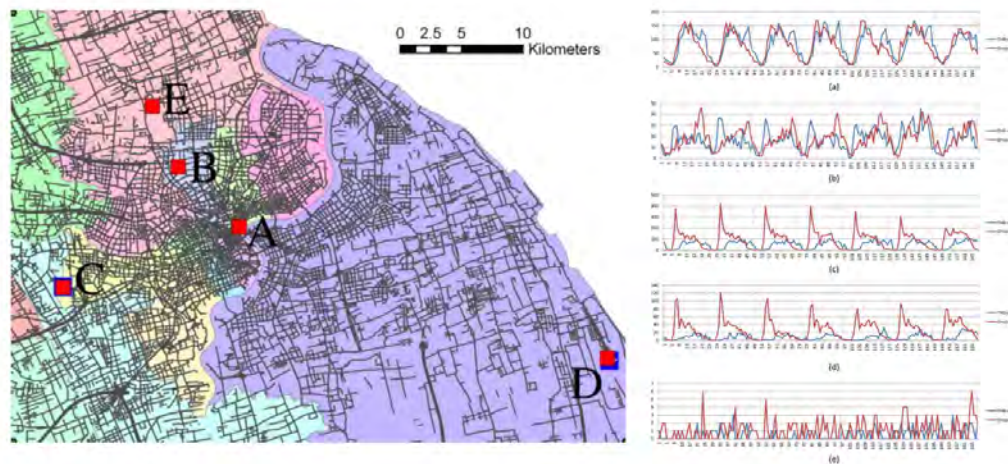
$$\mathbf{V}_i^{pickup} = [V_i^1, V_i^2, \dots, V_i^T] \quad (1)$$

$$\mathbf{V}_i^{dropoff} = [V_i^1, V_i^2, \dots, V_i^T] \quad (2)$$

Based on the balance between the numbers of drop-offs and pick-ups and its distinctive temporal patterns $\mathbf{V}_i^{dropoff} - \mathbf{V}_i^{pickup}$ for each pixel i at time t ($= 1, \dots, T$), the study area is classified into six traffic ‘source-sink’ areas using the K -means clustering method. These areas are closely associated with various land use types (commercial, industrial, residential, institutional and recreational) as well as land use intensity. Five sample points are selected from the study area to represent various locations (land uses), and their corresponding \mathbf{V}^{pickup} and $\mathbf{V}^{dropoff}$ are depicted in Figure 3. Their temporal patterns differed significantly. For example, the average numbers of pick-ups and drop-offs were roughly equal for cells A and B. In either cell C or D, however, the average number of pick-ups was much fewer than the average number of drop-offs. Cell E had far lower numbers of pick-ups and drop-offs than the other four locations. It confirms that the temporal patterns of pick-ups and drops-offs vary a great deal from place to place, and are manifest of the function of the place.

3.2 Clustering based on the component of activity type (A -fiber)

Leveraging a comprehensive data collection of bus, metro and taxi ridership from Shenzhen, China, we further unveil the spatio-temporal interplay between the mixed use of transport modes and the underlying urban land use. For each spatial analysis unit (SAU), we build



■ **Figure 3** Temporal variations of pick-ups and drop-offs of 5 sample points in Shanghai, China (A: downtown; B: residential; C: Hongqiao Airport; D: Pudong Airport; E: suburban) [8].

a volume signature capturing the temporal fluctuations of ridership of mass transit modes during a day. Taking 15-minutes as the temporal granularity, the signature \mathbf{V}_i of a SAU i is denoted as a $1 \times T$ vector quantifying the ridership of bus, metro or cab within the n th time slot. Targeting to compare three distinct mass transit modes, we therefore obtain the signatures of volume $\{ \mathbf{V}_i^{bus}, \mathbf{V}_i^{metro}, \mathbf{V}_i^{cab} \}$ for bus, metro and cab ridership within each SAU i , as well as the signatures of ratio $\{ \mathbf{R}_i^{bus}, \mathbf{R}_i^{metro}, \mathbf{R}_i^{cab} \}$ of ridership of different mass transit modes over time:

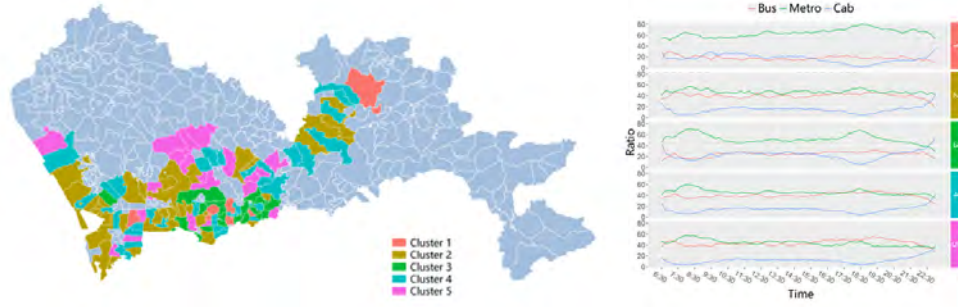
$$\mathbf{R}_i^{bus} = \mathbf{V}_i^{bus} / (\mathbf{V}_i^{bus} + \mathbf{V}_i^{metro} + \mathbf{V}_i^{cab}) \quad (3)$$

$$\mathbf{R}_i^{metro} = \mathbf{V}_i^{metro} / (\mathbf{V}_i^{bus} + \mathbf{V}_i^{metro} + \mathbf{V}_i^{cab}) \quad (4)$$

$$\mathbf{R}_i^{cab} = \mathbf{V}_i^{cab} / (\mathbf{V}_i^{bus} + \mathbf{V}_i^{metro} + \mathbf{V}_i^{cab}) \quad (5)$$

where \cdot / \cdot represents the itemwise division between two input vectors.

Applying a novel spectral clustering on the proposed signatures of the ratio of ridership, we obtain 5 clusters of SAUs that demonstrate distinct patterns of bus, metro and cab ridership dynamics as shown in Figure 4. In Cluster 1, metro rails play the most important role within these SAUs. During morning and evening commuting periods, metro ridership increase significantly. In comparison, the ratios of bus ridership and cab ridership are relatively low. Besides, the temporal fluctuations of bus ridership and cab ridership are also distinct. In Cluster 2, bus ridership and metro ridership are at a comparative level, which is significant higher than cab ridership. It indicates passengers have easy access to bus and metro at the same time. However, during morning and evening commuting periods, bus and metro ridership show no significant increase to that of working time periods. In Cluster 3, metro ridership demonstrate substantial increase during the morning and the evening commuting periods. On the contrary, bus ridership show no peaks during the commuting periods and its ratio is very low. In Cluster 4, bus ridership and metro ridership are very similar to that of Cluster 2. However, within these SAUs, increase of ridership during the morning commuting period are high while that during the evening commuting period is low. In Cluster 5, bus and metro compete for the dominant mass transit mode during different time regimes. During the morning commuting period, metro rails are the dominant mass transit mode. Whereas, during the evening commuting period, buses become the dominant mass transit



■ **Figure 4** Temporal variations of ridership patterns of mass transit modes associated with different clusters of SAUs in Shenzhen, China (Cluster 1: business and commercial; Cluster 2: rich residential; Cluster 3: mixed-use; Cluster 4: middle-income residential; Cluster 5: recreational) [13].

mode. Over the entire day, cab ridership is always relatively low. This phenomena reveals the transmission of passengers' preference of different mass transit modes over time. In general, different categorized urban spaces are associated with different accessibility levels (such as high-, medium-, and low-ranked) and different urban functionalities (such as residential, commercial, leisure-dominant, and home-work balanced). The results indicate that the demographic and socioeconomic attributes of the underlying urban environments can be revealed by the ridership dynamics of different mass transit modes.

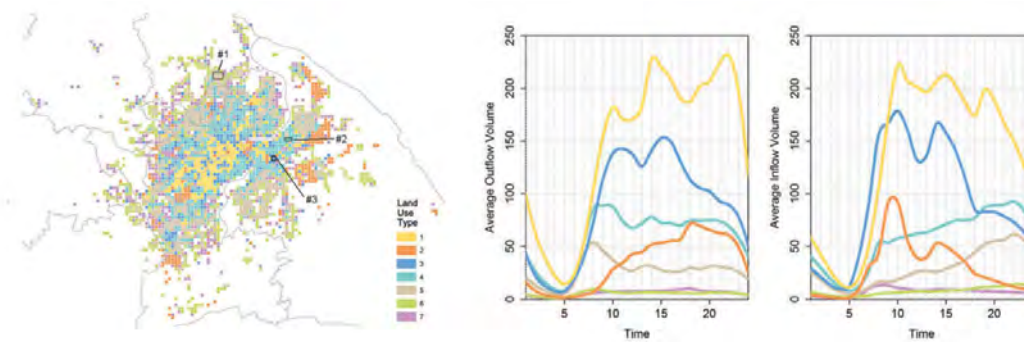
3.3 Clustering based on the pattern of spatial interaction (*I*- and *O*-fiber)

Based on the observation that spatial interaction patterns between places of two specific land uses are similar, we derive a new type of place signature to infer urban land uses from a perspective of connections. The method is validated with a case study using taxi trip data from Shanghai. Assuming that intra-city spatial interactions between N different places represented by travel flows can be extracted from the massive data sets, in each hour of a day, an $N \times N$ origin-destination (OD) matrix M^t ($t \in [1, 2, \dots, T]$) of population movements can be constructed, denoting the population moving from place i to j at time slot t as $m_{i,j}^t$ ($i, j \in [1, 2, \dots, N]$). Using $m_{i,j}^{t,k}$ and $m_{j,i}^{t,k}$ to denote the outflows from place i to j and inflows from j to i at time slot t , while the assumed land use type of j is K , we build the grouped interaction signature \mathbf{V}_i^{group} for place i as

$$\mathbf{V}_i^{group} = \begin{pmatrix} m_{i,\cdot}^{1,1} & \dots & m_{i,\cdot}^{T,1} & m_{\cdot,i}^{1,1} & \dots & m_{\cdot,i}^{T,1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ m_{i,\cdot}^{1,K} & \dots & m_{i,\cdot}^{T,K} & m_{\cdot,i}^{1,K} & \dots & m_{\cdot,i}^{T,K} \end{pmatrix} \quad (6)$$

which represents the flow patterns of a place with a trade-off between aggregated patterns and individual spatial interactions.

Inspired by the expectation-maximization (EM) algorithm, we use an iterative algorithm combined with the K -means clustering method to link the clustered parcels to their corresponding land uses. Figure 5 illustrates the classification result based on the grouped interaction signature of parcels. By interpreting the mean temporal signature curves and referring to Google Map information, we assign the roughly corresponding land uses to each parcel cluster. Type 1 have inflow peaks in the morning, afternoon and early evening, representing residents coming for work, business and eating/shopping/entertaining, respectively.



■ **Figure 5** Temporal variations of spatial interaction patterns between different categories of land uses in Shanghai, China [6].

Whereas, the two outflow peaks in the afternoon and night represent people's travels for business and going home, respectively. Therefore, these parcels are urban commercial and business area. Type 2 covers business and industrial area. Type 3 covers civic facilities, such as railway stations, hospitals and museums, and their inflow and outflow peaks are in the daytime. For Type 4, 5 and 6, the normalized mean temporal signatures of them all show that people leave these regions in the morning and return to these areas in the evening, which is consistent with the way people use residential areas. According to their spatial distributions, we name them urban residential area, outskirts urban residential area and suburban residential area, respectively. Type 7 are considered to be other land use area with few taxi trips. These results confirm that urban functionality can be better understood by analyzing the interaction patterns between different land uses.

4 Conclusion and Discussion

In this article, we proposed a space, time and activity cuboid based analytical framework for understanding the functionality of urban spaces. The core contribution is how to organize the human activity data into the cuboid for building meaningful and informative activity signatures. Applied in three case studies with the derived signatures of the variation of activity intensity, the component of activity type and the pattern of spatial interaction, the ability of the proposed analytical framework is confirmed. Note that directly following the remote sensing paradigm surely shows promising potentials for understanding and analyzing urban spaces. However, there are also several pitfalls should be aware of by researchers and practitioners as listed below.

- **Activity:** Different activities have substantially distinct spatiotemporal characteristics. Particularly, the big data revolution has been producing plentiful geo-data associated with individuals and their activities in space-time. Much more urban phenomenon are accessible and identifiable by this new and rich data source. For instance, spatial distributions of mobile phone and taxicab usages are observed to be quite different in many cities [4]. It is of critical importance to choose what kind of activity to analyze.
- **Feature selection:** Even for a single type of activity, the feature or feature combination selected also can result in inconsistent results. For instance, the combination of features related to taxi pick-up/set down dynamics significantly influence the recognition accuracy of urban land uses in a Chinese city [9]. Therefore, the feature should be carefully selected based on the research context.

- **Normalization:** Unlike the spectral curve of remote sensing, which is typically consistent for the same type of geographical objects and invariable with object size, the signatures of same land use types can be very different in magnitude order, in that socioeconomic activities change superlinearly with urban area size [2]. To cope with this issue, normalization is thus usually conducted before clustering the signatures.
- **Clustering:** In the context of using urban activity for land use classification, many classical clustering algorithms can be used because the signatures of urban zones can be simply regarded as time series. A comprehensive survey of time series clustering algorithms can be found in the literature [5]. The main challenge lies in the way to measure the similarity between the activity signatures.


References

- 1 Michael Batty. *Urban modelling*. Cambridge University Press Cambridge, 1976.
- 2 Luís MA Bettencourt. The origins of scaling in cities. *Science*, 340(6139):1438–1441, 2013.
- 3 Susan L Handy, Marlon G Boarnet, Reid Ewing, and Richard E Killingsworth. How the built environment affects physical activity. *American Journal of Preventive Medicine*, 23(2):64–73, 2002.
- 4 Chaogui Kang, Stanislav Sobolevsky, Yu Liu, and Carlo Ratti. Exploring human movements in singapore: a comparative analysis based on mobile phone and taxicab usages. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, pages 1–7. ACM, 2013.
- 5 T Warren Liao. Clustering of time series data—a survey. *Pattern Recognition*, 38(11):1857–1874, 2005.
- 6 Xi Liu, Chaogui Kang, Li Gong, and Yu Liu. Incorporating spatial interaction patterns in classifying and understanding urban land use. *International Journal of Geographical Information Science*, 30(2):334–350, 2016.
- 7 Yu Liu, Xi Liu, Song Gao, Li Gong, Chaogui Kang, Ye Zhi, Guanghua Chi, and Li Shi. Social sensing: A new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers*, 105(3):512–530, 2015.
- 8 Yu Liu, Fahui Wang, Yu Xiao, and Song Gao. Urban land uses and traffic ‘source-sink areas’: Evidence from gps-enabled taxi data in shanghai. *Landscape and Urban Planning*, 106(1):73–87, 2012.
- 9 Gang Pan, Guande Qi, Zhaohui Wu, Daqing Zhang, and Shijian Li. Land-use classification using taxi gps traces. *IEEE Transactions on Intelligent Transportation Systems*, 14(1):113–123, 2013.
- 10 Jonathan Reades, Francesco Calabrese, and Carlo Ratti. Eigenplaces: analysing cities using the space–time structure of the mobile phone network. *Environment and Planning B: Planning and Design*, 36(5):824–836, 2009.
- 11 Jameson L Toole, Michael Ulm, Marta C González, and Dietmar Bauer. Inferring land use from mobile phone activity. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, pages 1–8. ACM, 2012.
- 12 Mao Ye, Krzysztof Janowicz, Christoph Mülligann, and Wang-Chien Lee. 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*, pages 102–111, New York, NY, USA, 2011. ACM.
- 13 Mengxue Yue, Chaogui Kang, Clio Andris, Yu Liu, Kun Qin, and Qingxiang Meng. Understanding the interplay between bus, metro and cab ridership dynamics in shenzhen, china. *Transactions in GIS*, 2018. doi:10.1111/tgis.12340.

Application of Style Transfer in the Vectorization Process of Floorplans

Seongyong Kim

Department of Civil and Environment Engineering, Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 151-744, South Korea
syoi@snu.ac.kr

 <https://orcid.org/0000-0002-0774-6791>

Seula Park

Department of Civil and Environment Engineering, Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 151-744, South Korea
seula90@snu.ac.kr

Kiyun Yu

Department of Civil and Environment Engineering, Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 151-744, South Korea
kiyun@snu.ac.kr

Abstract

As the market for indoor spatial information burgeons, the construction of indoor spatial databases consequently gain attention. Since floorplans are portable records of buildings, they are an indispensable source for the efficient construction of indoor environments. However, as previous research on floorplan information retrieval usually targeted specific formats, a system for constructing spatial information must include heuristic refinement steps. This study aims to convert diverse floorplans into an integrated format using the style transfer by deep networks. Our deep networks mimic a robust perception of human that recognize the cell structure of floorplans under various formats. The integrated format ensures that unified post-processing steps are required to the vectorization of floorplans. Through this process, indoor spatial information is constructed in a pragmatic way, using a plethora of architectural floorplans.

2012 ACM Subject Classification Applied computing → Graphics recognition and interpretation, Computing methodologies → Scene understanding

Keywords and phrases Floorplan, Vectorising, Style Transfer, Generative Adversarial Networks

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.39

Category Short Paper

Funding This research was supported by a grant (18NSIP-B135746-02) from the National Spatial Information Research Program (NSIP) funded by the Ministry of Land, Infrastructure and Transport of Korean government

1 Introduction

Recently, the development of information technology has made it possible to expand location-based services such as position tracking and indoor navigation. Consequently, the indoor spatial information market has burgeoned and various studies on indoor spaces are being conducted. The accomplishment of several services and research on indoor spaces requires a database that contains information of the geometry, topology, and semantics of indoor



© Seongyong Kim, Seula Park, and Kiyun Yu;
licensed under Creative Commons License CC-BY

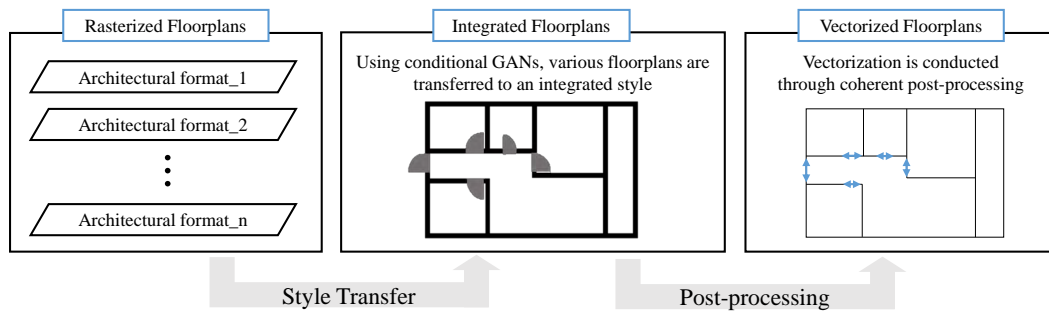
10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 39; pp. 39:1–39:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



■ **Figure 1** Study flow chart.

cells. The construction of indoor spatial information has been based on aerial photographs, 3D laser scanning, 2D floorplans, and CAD plans [7]. Among these, the 2D floorplan is pragmatically focused on because it is obviously present in existing buildings, being usually open source and simply and effectively available compared to other methods. With such features, OpenStreetMap and Google Maps provide a plug-in that allows users to build their own indoor maps using floorplans. In addition to these tools, several systems are used in the construction of indoor spatial information from floorplans, such as open sources like QGAR [13] and commercial vectorization software. These systems, however, have some limitations: the accuracy of outputs depends on the level of information represented by the floorplans, such as grid lines, layouts, symbols, and electric wiring, thus requiring heuristic revisions.

This paper proposes a method of refining various types of floorplans and levels of information in a consistent form. Even with the dramatic advances in computer vision, retrieving information on floorplans is a challenge due to the number of rooms in different houses, different formats of symbols and walls, and different levels of information. Despite all these difficulties, humans can still recognize the structures of houses from floorplans. The goal of this paper is to materialize this "perception" through deep networks by learning many types of floorplans.

2 Related work

Retrieving leaking information from raster floorplans follows sequential steps [4, 7, 8]. First, textual and graphical data are separated in a preprocessing step [14]. Then, lines in vector are extracted from the graphical data [1]. The next step is a pattern recognition that assigns semantic information, such as walls and openings, to the extracted lines [3, 6, 8]. Finally, room space is detected through the use of geometry and semantic information, including textual data [3, 12]. The studies of construction of indoor information from floorplans are, in a broad view, all or parts of these steps. Although the issue of these research is automation, they work partially or conditionally for practical datasets, demanding additional manual processes.

The main reason for handwork is that a model does not handle a wide spectrum of architectural floorplans. Previous research focus mainly on a consistent form of floorplan datasets [5] or, after classifying floorplans by wall and symbol formats, apply tuned algorithms respectively [3]. In order to ensure versatility for previous models of retrieved floorplans, the goal of this study is to convert several floorplans into an integrated and unified form. Figure 1 shows the whole workflow. Various rasterized floorplans are converted to an integrated format, and being vectorized by a coherent post-processing.



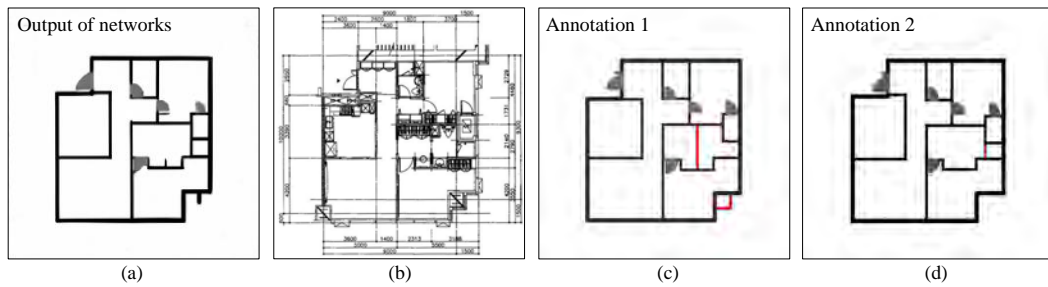
■ **Figure 2** Different floorplan formats: (a) format with simplified walls, (b) format with bearing walls, pillars and the interior symbols, (c) format with the electronic wiring, (d) format with different representative of walls.

We approached the problem by integrating diverse formats of floorplans with a style transfer that converts domains of data while maintaining their features. Recently, image style transfer has improved remarkably with the development of generative network models. Based on Generative Adversarial Networks (GANs), deep networks, such as Conditional GANs [9], CycleGANs [15], and DiscoGANs [10], have gained great reputation on style transfers. Conditional GANs and CycleGANs transfer images into different styles with preserving the underlying structure, while DiscoGANs focuses primarily on the texture of them. Conditional GANs works in condition that labeled pairs exist, while Cycle GANs and DiscoGANs aims to convert domains even when images in each domain are not paired. We propose an integrated format with a strength in vectorization and convert floorplans to this format. Given the characteristics of each networks, we use Conditional GANs because we prepare pairs of floorplans in the integrated format and underlying structure of floorplans that conveys geometry information is important on our goal.

3 Style for integrated floorplans

In order to integrate different floorplan formats into a unified format, the following should be considered: 1) the represented information shared in diverse floorplans, 2) the meaningfulness of the level of information extracted as material for indoor spatial data. The architectural floorplan spectrum that can be used for the construction of indoor spatial data is variable (Figure 2). In Figure 2 (a), the wall structure is simplified and detailed information, such as equipment, is omitted. Figure 2 (b) represents the bearing walls and the pillars, as well as several interior symbols. Figure 2 (c) represents the electronic wiring on the topology of the walls. Figure 2 (d) represents walls with different formats and some noises is present. Floorplans of such varied purposes, however, preserve a structure of cells made of walls and openings. In the integrated format, the simplified walls and openings are targeted as representative of the floorplan. Homogeneous walls and openings containing original information represent the structure of cells. Figure 1 shows example of the integrated format. Research on indoor information, such as matching indoor position with photo, are based on the extrusion of simplified walls with openings [2, 11]. In other words, indoor research generally uses only the topology of walls and openings, which means that the integrated floorplan style can be used significantly.

A floorplan dataset was provided by the E-AIS (Electronic Architectural Administration Information System), which is an architectural floorplans management system maintained by Korea's Ministry of Land and Transport. Approximately 400 floorplans in various formats were used and manually annotated to the integrated format.



■ **Figure 3** (a) output, (b) original floorplan, [(c) annotation 1 and (d) annotation 2] are annotations by different annotators. In relation to the original floorplan, annotation 1 represents the boundary column and the sliding door as a wall, while annotation 2 disregards this information.

4 Style transfer via conditional GANs

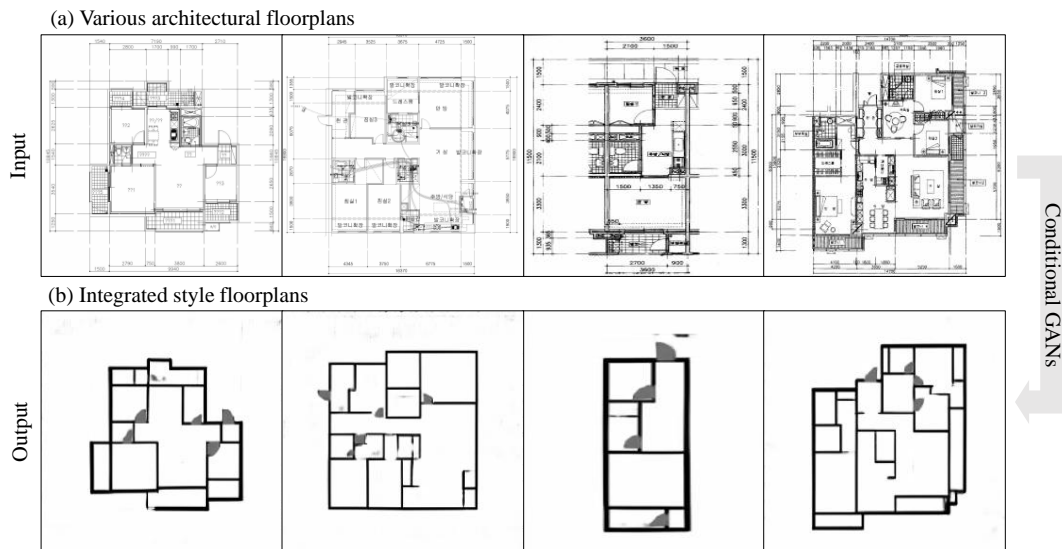
Conditional GANs proposed by Isola *et al.* [9] are practical and effective networks that transfer style preserving intrinsic features when data pairs are given. The generating networks are type U-net, which have the advantage of keeping the underlying structure, and the discriminating networks are type PatchGANs classifier, which discriminate generated images by summing up score of each patch implicitly. The integrated format and its correspondents were fed simultaneously, and both networks improve competitively. The network structure and the parameters were modified for the purpose of ours.

The goal of this study is to transfer floorplans to the integrated style. To maintain the underlying structure of floorplans, tuning is performed in a direction that emphasizes the sharp edges and the position accuracy of the simplified walls and openings. Regarding hyperparameters, 1) the L_1 error was increased by 1.5 times compared to that in the study on conditional GANs, and 2) the patch size of the discriminator was adjusted to 16×16 . The lower the ratio of L_1 error to generative error, the sharper the extracted walls, although they tend to be cut off. The smaller the patch size, the neater the induced outputs without noise, even if computation becomes larger. The aforementioned parameters were mutually determined for the networks.

The advantage of using generative models rather than pixel-based classification models such as Convolutional Neural Networks(CNNs) is that the structure of cells and the shape of the homogeneous walls are “selected and generated” by networks. As seen in Figure 2, many architectural floorplans are neither neat nor homogeneous, which means that the outputs of the classification models require demanding post-processing steps.

5 Qualitative evaluation

Given that the purpose of this work is to transfer floorplans from various formats to the integrated one, this study inevitably aimed at ambiguous criterion. The annotating depends on individuals while prioritizing a representation of the cell structure, thus yielding multiple annotations match with one floorplan (Figure 3). Since the networks were trained with these pairs, we do not ensure their incorrectness even when the outputs do not match with the annotations. Figure 3 (a) shows this. An output (Of course, this pairs is only used in test set) does not match with both annotations but represent the cell structure quite well. Therefore, an evaluation should take account of preserving structure of cells, which was inappropriate for raster output images. For this reason, qualitative evaluation is performed for the style transfer.



■ **Figure 4** Results (a) Various architectural floorplans as inputs, (b) Transferred floorplans in the integrated format as outputs

Figure 4 shows the results of the style transfer. Figure 4 (a) represents various floorplan spectra used as inputs, while Figure 4 (b) represents the integrated style floorplans operated by deep networks. Despite the diverse formats and levels of representative information, the floorplans were transferred into the coherent, integrated form. The networks generate walls with uniform shape and clear boundaries even in conditions where the original walls were crooked, or edges were blurred. However, in the case of openings, the positions were extracted correctly, but the boundary was blurred. In particular, the networks perform well at points where the information was overlapping (e.g. doors on tile patterns, pillars in walls) which was identified as a difficulty in previous studies.

6 Conclusion

When applying style transfer via conditional GANs, diverse floorplans were transferred into the integrated format. The deep networks is suitable for this problem, we confirm that it works well. This ensures that single unified post-processing steps are required to the consummation of vectorizing floorplans. Through this process, it is possible to construct indoor spatial information in a pragmatic way, using a plethora of architectural floorplans.

For further study, we will perform additional evaluation. This paper covers suggestion of the style transfer in the vectorization of floorplan, thus only the qualitative evaluation is performed. In the field of retrieving information from floorplans, a match table is a common evaluation method for vector results, that is proper forms for representing cell structures. Thus, after whole vectorization that is specific to the integrated format, we will perform the matching-based assessment as a quantitative evaluation.

References

- 1 Sheraz Ahmed, Marcus Liwicki, Markus Weber, and Andreas Dengel. Improved automatic analysis of architectural floor plans. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 864–869. IEEE, 2011.
- 2 Hang Chu, Dong Ki Kim, and Tsuhan Chen. You are here: Mimicking the human thinking process in reading floor-plans. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2210–2218, 2015.
- 3 Lluís-Pere de las Heras, Sheraz Ahmed, Marcus Liwicki, Ernest Valveny, and Gemma Sánchez. Statistical segmentation and structural recognition for floor plan interpretation. *International Journal on Document Analysis and Recognition (IJDAR)*, 17(3):221–237, 2014.
- 4 Lluís-Pere de las Heras, David Fernández, Ernest Valveny, Josep Lladós, and Gemma Sánchez. Unsupervised wall detector in architectural floor plans. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 1245–1249. IEEE, 2013.
- 5 Lluís-Pere de las Heras, Oriol Ramos Terrades, Sergi Robles, and Gemma Sánchez. Cvc-fp and sgt: a new database for structural floor plan analysis and its groundtruthing tool. *International Journal on Document Analysis and Recognition (IJDAR)*, 18(1):15–30, 2015.
- 6 Lluís-Pere de las Heras, Ernest Valveny, and Gemma Sánchez. Combining structural and statistical strategies for unsupervised wall detection in floor plans. In *Proceedings of the 10th IAPR International Workshop on Graphics Recognition*, pages 123–128, 2013.
- 7 Lucile Gimenez, Jean-Laurent Hippolyte, Sylvain Robert, Frédéric Suard, and Khaldoun Zreik. reconstruction of 3d building information models from 2d scanned plans. *Journal of Building Engineering*, 2:24–35, 2015.
- 8 Lucile Gimenez, Sylvain Robert, Frédéric Suard, and Khaldoun Zreik. Automatic reconstruction of 3d building models from scanned 2d floor plans. *Automation in Construction*, 63:48–56, 2016.
- 9 Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.
- 10 Taeksoo Kim, Moonsoo Cha, Hyunsoo Kim, Jungkwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017.
- 11 Chenxi Liu, Alexander G Schwing, Kaustav Kundu, Raquel Urtasun, and Sanja Fidler. Rent3d: Floor-plan priors for monocular layout estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3413–3421, 2015.
- 12 Sébastien Macé, Hervé Locteau, Ernest Valveny, and Salvatore Tabbone. A system to detect rooms in architectural floor plan images. In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, pages 167–174. ACM, 2010.
- 13 Jan Rendek, Gérald Masini, Philippe Dosch, and Karl Tombre. The search for genericity in graphics recognition applications: Design issues of the qgar software system. In *International Workshop on Document Analysis Systems*, pages 366–377. Springer, 2004.
- 14 Karl Tombre, Salvatore Tabbone, Loïc Péliissier, Bart Lamiroy, and Philippe Dosch. Text/-graphics separation revisited. In *International Workshop on Document Analysis Systems*, pages 200–211. Springer, 2002.
- 15 Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017.

Estimating Building Age from Google Street View Images Using Deep Learning

Yan Li

Department of Infrastructure Engineering, University of Melbourne/Melbourne, Australia
li17@student.unimelb.edu.au

Yiqun Chen¹

Department of Infrastructure Engineering, University of Melbourne/Melbourne, Australia
yiqun.c@unimelb.edu.au

Abbas Rajabifard

Department of Infrastructure Engineering, University of Melbourne/Melbourne, Australia
abbas.r@unimelb.edu.au

Kourosh Khoshelham

Department of Infrastructure Engineering, University of Melbourne/Melbourne, Australia
k.khoshelham@unimelb.edu.au

Mitko Aleksandrov

Department of Infrastructure Engineering, University of Melbourne/Melbourne, Australia
mitko.aleksandrov@unimelb.edu.au

Abstract

Building databases are a fundamental component of urban analysis. However such databases usually lack detailed attributes such as building age. With a large volume of building images being accessible online via API (such as Google Street View), as well as the fast development of image processing techniques such as deep learning, it becomes feasible to extract information from images to enrich building databases. This paper proposes a novel method to estimate building age based on the convolutional neural network for image features extraction and support vector machine for construction year regression. The contributions of this paper are two-fold: First, to our knowledge, this is the first attempt for estimating building age from images by using deep learning techniques. It provides new insight for planners to apply image processing and deep learning techniques for building database enrichment. Second, an image-base building age estimation framework is proposed which doesn't require information on building height, floor area, construction materials and therefore makes the analysis process simpler and more efficient.

2012 ACM Subject Classification Computing methodologies → Supervised learning by regression

Keywords and phrases Building database, deep learning, CNN, SVM, Google Street View

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.40

Category Short Paper

Funding The first author would like to acknowledge the financial support of China Scholarship Council (CSC).

¹ Corresponding author



1 Introduction

Building databases have been widely used for urban planning. New construction and renovation works require comprehensive building databases for analysis and decision-making. However, the application of such databases has been hampered by data integrity and accuracy issues [14].

Many image databases are available online. For example, Google Street View, which updates an online street image database periodically, provides advanced APIs for accessing building images by the location. Google street view based applications have been implemented in urban planning, such as estimating the demographic makeup of the cities [6], studying the relationships between city appearance and the health of its residents [5]. Another example is Google search, when appropriate keywords are provided, hundreds of relevant images will show in the results. Several public image classification databases are created from this data source, such as ImageNet [3] and CIFAR-10 [9].

Image processing using deep learning methods has shown great performance in many applications, such as image classification [10] and object segmentation [2]. Several popular deep learning methods have been developed for image analysis. Convolutional neural network (CNN) is the state-of-art method for feature extraction [10]. Support vector machine (SVM) has shown remarkable performance in regression problems, especially for high dimensional data [4]. Despite the rapid development of image processing techniques, building age estimation from images has not been studied in the research community. Existing methods such as [1] require additional building attributes (e.g. building height, floor area, etc.) for decision-making. Collecting these attributes is time-consuming and usually ends up with incomplete information. This paper proposes a novel method based on deep learning approach for direct building age estimation, using the CNN for image feature extraction and SVM for building age estimation.

The contributions of this research are two-fold:

- this is the first attempt for estimating building age from Google Street View images by using deep learning techniques. It provides new insight for planners to apply image processing and deep learning techniques for building database construction.
- the proposed image-based building age estimation framework is independent of building information, such as height, floor area, construction materials; therefore, it makes the analysis process simpler and more efficient.

2 Related work

2.1 Building age estimation

While building age is an important parameter in building specifications, the data is not always available or complete. Little research has been done for the building age estimation. [1] proposed an estimation method which adopts random forest regression and infers the building construction age from other attributes, such as ceiling height, footprint area, shape complexity and so on. However the accuracy of this method largely depends on the completeness of these attributes. This limitation motivates us to seek alternative solutions for overcoming the native incompleteness of existing database attributes. So looking for the available dataset is critical for our research and open data is ideal for this purpose. Several large Internet companies such as Google, Facebook, Instagram provide free APIs to access their image sources, in particular, Google Street View API provides house images based on a given location and hence it meets our requirement.

2.2 Image regression

This research treats the building age estimation as a regression problem. Image regression, which builds a regression model based on extracted image features, has shown the state-of-art performance in many tasks.

The general procedure of image regression contains two steps. The first step is to retrieve features from images, and the second step is to construct a regression model using the extracted features as inputs. The CNN [15, 11] is adopted in this research. Since the training of a CNN requires large training datasets and computing resources, many pre-trained models have been made publicly available. For example, the place365 dataset [16], trained by 8 million images, is used for scene recognition. Pre-trained CNN models are provided including AlexNet [10], ResNet18, ResNet50 [7], DenseNet161 [8], which are high performance CNN structures. As for the regression model, the support vector regression (SVR) can capture main features that characterize the algorithm (maximal margin). It is particularly suitable for high dimensional regression problems with a limited number of training samples.

3 Methodology

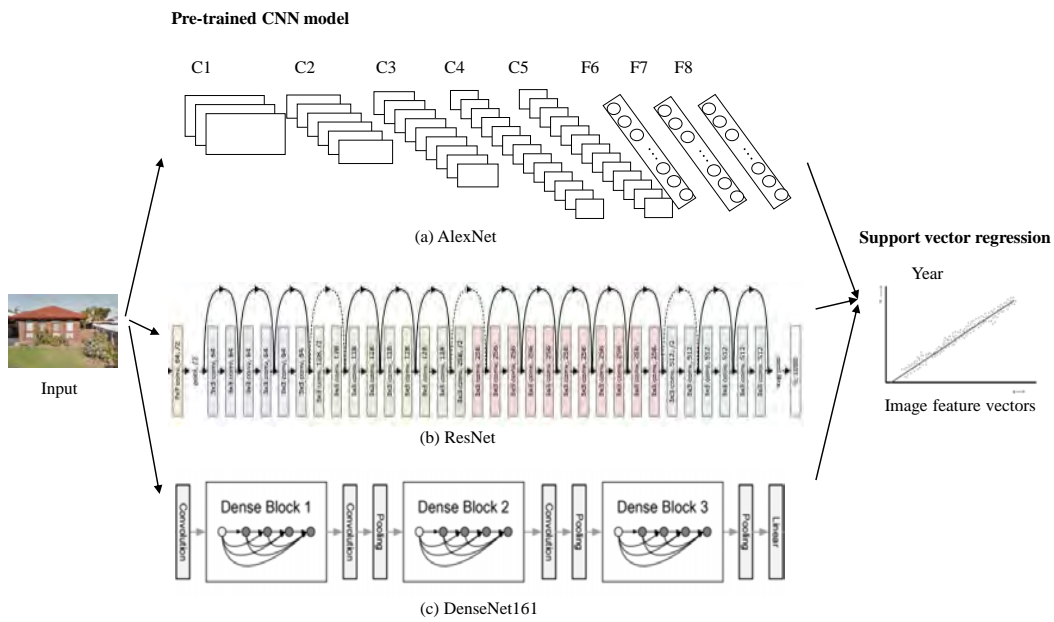
Our approach includes three major steps: data collection, feature extraction, and building age regression. First, the house images of each address are obtained from Google Street View API. Second, image features are extracted using a pre-trained CNN. At last, a SVR model is built by taking image feature vectors as inputs and building age as outputs.

3.1 Google Street View images download

Using the Google Street View Image API, we directly submit a list of addresses, for example, “172 Bouverie St, Carlton VIC 3053”, and then store the retrieved house images locally. This process avoids the potential accuracy problems introduced by geocoding procedure and successfully obtains all the house images except invalid street addresses. As the images are shot from streets, they usually contain the target house in the mid as well as parts of adjacent buildings on two ends. We tune the API parameters to obtain the exact region of the target house. The parameters include heading, pitch (the horizontal and vertical rotation of the camera respectively) and fov (field of view, controlling the width of the street view images). Based on our experiments, setting heading as 180, pitch as 0, and fov as 50 degrees yields the best image results. In particular, the fov should be assigned appropriately because a wide view will introduce neighbour buildings and a narrow view will only capture partials of the target building. Each retrieved building image is in 600x400 pixels, which is the largest size that Google Street View API provides.

3.2 Feature extraction by Convolutional Neural Network

In this paper, we choose the largest scene recognition database and three pre-trained CNN models including AlexNet, ResNet and DenseNet for image feature extraction, as shown in Figure 1. These models are different in network structures. AlexNet won the 2012 Imagenet competition. Compared with modern network structures, AlexNet is simple and consists of 5 convolutional layers, maxpooling layers, drop-out layers and three fully-connected layers. It is specially designed for classification with 1000 categories. ResNet won the 2015 ImageNet and COCO competitions, and it allows for effectively training deeper neural networks. DenseNet, proposed in 2016, is based on the hypothesis that convolutional networks can be substantially



■ **Figure 1** Different Convolutional neural networks are applied to extract image features. The convolutional networks are getting deeper from AlexNet to DenseNet.

deeper, more accurate and efficient to train if they build shorter connections between each layer and every other layer. All the experiments of extracting image features are implemented using the deep learning framework Pytorch [12].

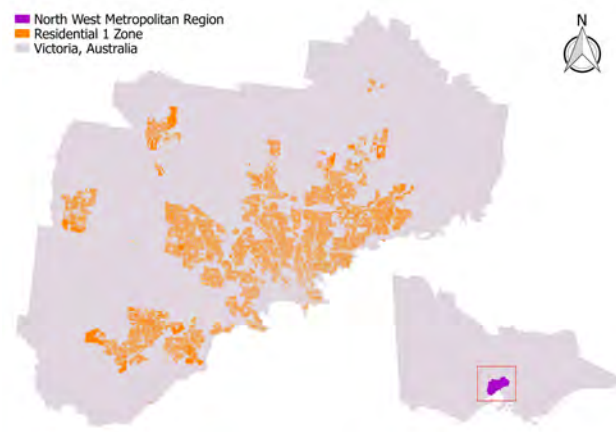
3.3 Support Vector Regression

The support vector regression (SVR) [4] is advanced in high dimensionality space because SVR optimization doesn't depend on the dimensionality of the input space, and provides different kernel functions for the decision function. This research chose Scikit-learn library [13] to build the SVR model by taking image vectors as inputs and building age (construction year) as outputs. 80% of data are used to train the regression model and the best fit SVR model is decided according to the training data. Then we perform regression on test data based on the trained model.

4 Experiment results

4.1 Dataset

As shown in Figure 2, the North and West Metropolitan Region (NWMR) is chosen as the case study area, which is the most populous and diverse region extending from the Melbourne CBD to the outer northern and western suburbs in Victoria, Australia. It has 2981 square kilometres, 14 local government areas and around one third (33.1%) of the population of Victoria (2011 Census). The building attributes for NWMR are extracted from Valuer-General Victoria valuation dataset which contains the location, street address, zoning type, construction year, building material and valuation prices for both land and property across the entire Victoria. We assume that the building images from Google Street View for different zoning types may vary significantly and weigh down the model performance, hence



■ **Figure 2** Case study area: North and West Metropolitan Region (NWMR), Victoria, Australia

■ **Table 1** Accuracy of each CNN structure

CNN structure	AlexNet	ResNet18	ResNet50	DenseNet161
MAE	10.749	10.996	10.722	10.689
RMSE	12.210	12.423	12.154	12.121

the dataset is further narrowed down to Residential 1 Zone (R1Z) which contains 520,694 (69.5%) buildings in NWMR. It also should be noted that among these R1Z buildings, 21,830 (4.19%) of them miss construction year (i.e., building age) information. The key motivation of this work is to estimate the missing values.

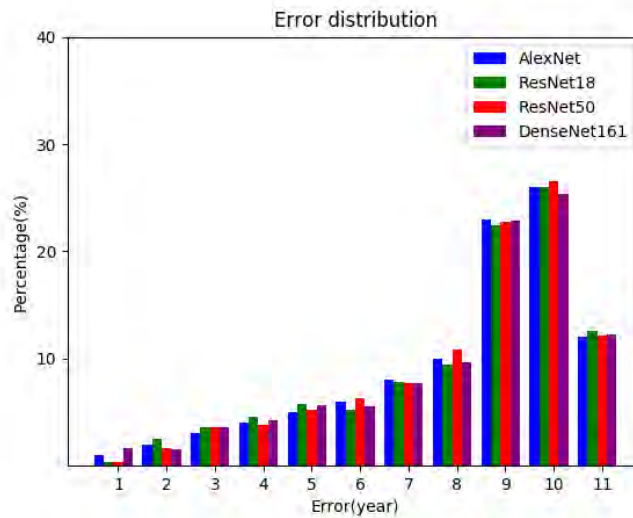
4.2 Accuracy

For regression models, two evaluation metrics are widely used for performance evaluation: mean absolute error (MAE) and root mean squared error (RMSE), both indicate the error of prediction results. MAE is the average over the test sample of the absolute differences between the prediction and the actual observation where all individual differences have equal weights. RMSE is the square root of the average of squared differences between the prediction and the actual observation. Since the errors are squared before they are averaged, RMSE is more useful particularly when large errors are undesirable.

Table 1 summarises the estimation performance of different CNN structures. DenseNet161 shows the best performance among them, and it confirms that deeper CNN structure tends to yield more accurate results[8]. Figure 3 shows the distribution of errors. Around 15% samples have less than 5 years error. Most samples, about 25%, have about 10 years error.

4.3 Findings

The changing of building fashions allows inspectors to roughly determine when buildings are constructed, based on their appearances, materials, components and styles. Inspired by this idea, we explore the feasibility of teaching a machine to estimate the building age by reading housing images and learning the styles. We list house samples in the same age range in Figure 4 and find some interesting patterns. Clearly, it can be observed that more recently constructed buildings tend to have newer facades, and houses are getting more complex both



■ **Figure 3** Error distribution of each CNN structure

in horizontal and vertical space. Duplex houses also prevail in recent decades. Exterior wall materials have also changed over time. Before 2000, newly built houses had wood or brick exteriors; while after that, new houses start to use vinyl siding. These features are hidden in images and could be learned and extracted by the CNN models and then passed to our SVR model for the regression analysis.

5 Conclusions and future work

In this study, a novel approach for direct estimation of building age from Google Street View images is proposed, implemented and tested. The algorithm consists of three major steps: Google Street View images download, image features extraction and building age estimation. Results for the North and West Metropolitan Region of Victoria show that building age estimation can be accurately predicted with deeper convolutional neural networks.

References

- 1 Filip Biljecki and Maximilian Sindram. Estimating building age with 3d gis. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-4/W5, 2017.
- 2 Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, 24(12):5706–5722, 2015.
- 3 J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.
- 4 Harris Drucker, Christopher JC Burges, Linda Kaufman, Alex J Smola, and Vladimir Vapnik. Support vector regression machines. In *Advances in neural information processing systems*, pages 155–161, 1997.
- 5 Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, and César A Hidalgo. Deep learning the city: Quantifying urban perception at a global scale. In *European Conference on Computer Vision*, pages 196–212. Springer, 2016.
- 6 Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. Using deep learning and google street view to estimate the demo-



■ **Figure 4** Samples of buildings with estimated construction year


- graphic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences*, page 201700035, 2017.
- 7 Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
 - 8 Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
 - 9 Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.
 - 10 Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
 - 11 Yan Li, Majid Sarvi, Kourosh Khoshelham, and Milad Haghani. Real-time level-of-service maps generation from cctv videos. In *Transportation Research Board 97th Annual Meeting*, 2018. URL: <https://trid.trb.org/view/1497380>.
 - 12 Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch, 2017. URL: <https://openreview.net/forum?id=BJJsrnfCZ>.
 - 13 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
 - 14 Abbas Rajabifard. *Spatial data infrastructure*. International Federation of Surveyors (FIG), 2012.
 - 15 Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.
 - 16 Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

Center Point of Simple Area Feature Based on Triangulation Skeleton Graph

Wei Lu

Wuhan University, Wuhan, China


whuluwei@whu.edu.cn

 <https://orcid.org/0000-0002-9703-2871>

Tinghua Ai

Wuhan University, Wuhan, China

tinghuaai@whu.edu.cn

 <https://orcid.org/0000-0002-6581-9872>

Abstract

In the area of cartography and geographic information science, the center points of area features are related to many fields. The centroid is a conventional choice of center point of area feature. However, it is not suitable for features with a complex shape for the center point may be outside the area or not fit the visual center so well. This paper proposes a novel method to calculate the center point of area feature based on triangulation skeleton graph. This paper defines two kinds of centrality of vertices in skeleton graph according to the centrality theory in graph and network analysis. Through the measurement of vertices centrality, the center points of polygon area features are defined as the vertices with maximum centrality.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases Shape Center, Triangulation Skeleton Graph, Graph Centrality

Digital Object Identifier 10.4230/LIPICs.GIScience.2018.41

Category Short Paper

Funding This research was supported by the National Key Research and Development Program of China (Grant No. 2017YFB0503500), and the National Natural Science Foundation of China (Grant No. 41531180).

1 Introduction

In geographic information science (GIS), skeleton and center point are two important abstract descriptors of area feature which are extensively used in spatial data compression, cartographic generalization, map annotation configuration, multiscale map matching, spatial relation calculation, etc. Skeleton is a dimension reduction representation of area feature which maintains the geometric and topological characteristics of the area feature. Generally, the skeleton of area feature is a graph structure. The branches reflect the topological relation between different part of an area feature. The extension, length, and width of each part indicate the geometric characteristics of an area feature[7][3]. As for the calculation of center point of an area feature, the most popular used method is the centroid of boundary polygon of an area feature[11]. The pole of inaccessibility evaluation is also used to calculate the center point of area feature[9]. Chen presented a method for calculating the shape center through the triangulation skeleton of area feature[7]. As Chen indicated this method heavily relies on the parameter selection, and it will not guarantee the center point within the area feature. Inspired by Chen's work, this paper provides a new center point extraction method based on skeleton graph of a simple polygon.



© Wei Lu and Tinghua Ai;
licensed under Creative Commons License CC-BY

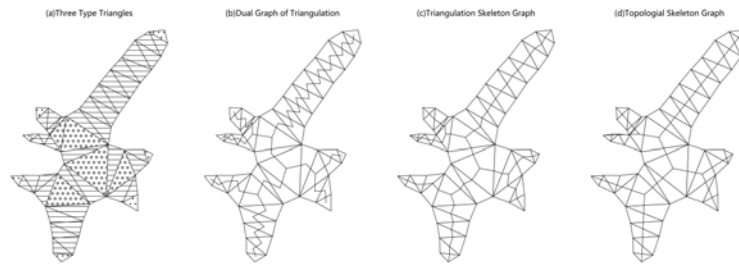
10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 41; pp. 41:1–41:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



■ **Figure 1** Triangulation of Polygon and Related Structures.

This paper presents a method to define the centrality of polygon area feature based on its triangulation skeleton. On the skeleton graph structure of polygon, we define the betweenness and closeness centrality of skeleton graph vertex which is similar to centrality in graph theory[6]. By the centrality definitions, we can extract different kinds of center points of area features. At last, we discuss the algorithm complexity of the methods presented.

2 Triangulation Skeleton Graph of Simple Area Feature

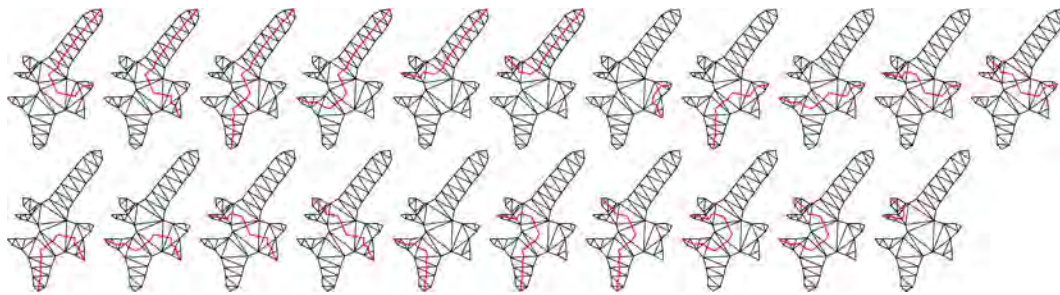
Skeleton or medial axis is a concept firstly used in biology as the descriptor of biological Shape[5]. In computational geometry, this structure has been studied extensively[8][4]. And there are several different definitions for this structure. Skeleton is widely studied and used in areas such as image recognition, medicine analysis, geospatial science, etc. In cartography and geographical information science. A kind of skeleton based on triangulation of polygon is generally used for spatial relation calculation, map annotation, and map generalization[2][1]. This section will give some brief formal definition of this kind skeleton structure and some basic concepts for the definition of centrality of polygon.

2.1 Triangulation of Simple Polygon

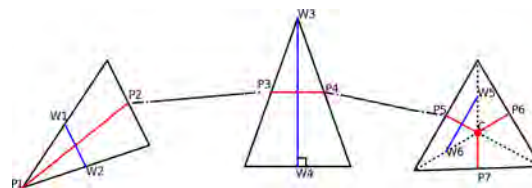
This paper studies the GIS area feature which is formed by simple polygon P in a two-dimensional plane of Euclidean space. A decomposition of P into triangles by a maximal set of non-intersecting diagonals is called a triangulation of P , noted as T_P [4]. This decomposition is not unique for every simple polygon. The number of different triangulation is a Catalan number related to the number of vertices[8]. In paper[2], the authors studied the influence of different triangulation on the form of the skeleton of a polygon. In GIS science area, constrained Delaunay triangulation is used prevalently in engineering projects and scholar researches. According to the definition of T_P , there are three kinds of triangles classified by edge type(Figure 1.a). The one which contains one diagonal is noted as type I triangle, or ear triangle; the one which contains 2 diagonals as type II triangle, or link triangle; the one which contains three diagonals as type III triangle, or branch triangle. The dual graph of triangulation[4](Figure 1.b) represents the topological link relations between sub-areas of a polygon which shows the topological characteristics of different visual feature parts of a polygon.

2.2 Basic Definitions

The triangulation skeleton graph of P , G_P , is defined by a construction process presented in [2]. The vertices of graph G_P can be the vertices of P (**end vertex**), and middle point of diagonals (**link vertex**) of P , and mass centers of triangles in T_P (**branch vertex**). The structure is shown in Figure 1.c.



■ **Figure 2** All Skeleton Paths of a Polygon Triangulation.



■ **Figure 3** Geometric Definition of Cover Width of Three Type of Triangles.

The shortest path between every two vertices in G_P is defined as a **skeleton branch**. If the path between two vertices of G_P doesn't contain any branch vertex, we call the two vertices directly adjacent. The skeleton branch of two end vertices s, t is called **skeleton path**, noted as $P_{s,t}$. As shown in Figure 2, the red skeleton branches are all the skeleton paths of a polygon. If all the link vertices are removed, and the directly adjacent end vertices and branch vertices are connected, we have a topological skeleton graph of P , as shown in Figure 1.d.

We define the cover length, cover width and cover area of the edge of G_P . The cover area of type I and II edge is the area of the corresponding triangle, and cover area of type III edge is $\frac{1}{3}$ of the area of the corresponding triangle. The cover length of each edge is the geometric length of the edge. The geometric definitions of cover width of three type triangles are in the following description shown in Figure 3. The red line segments are the edges, and the blue line segments are the geometric definition of cover width of each edge. For type II triangle, width is the length of the height of triangle on non-diagonal edge, shown as W_3W_4 . For type I triangle, we find a line segment W_1W_2 on triangle parallel to the diagonal edge, and the product of the length of W_1W_2 and the length of edge P_1P_2 will equal to the area of the triangle. For type III triangle, the three sub-area of it can be regarded as type I triangle, the definition of width W_5W_6 for each sub-area is the same as type I triangle.

3 Centrality of Area Feature

Center point of shape is an important attribute of a geographic feature. Generally, the centroid of a polygon will be regarded as the center point. And for special shapes, the center points would not be within the polygon and they will not suitable for some applications, such as annotation of area features. Chen proposed a method based on the main skeleton of a polygon to calculate the center point. There are parameters to be specified when adopting Chen's method which is subjective and sensitive to different data and it will not guarantee the center point always be within the polygon. This study uses the skeleton to define the center point of a polygon from the perspective of graph centrality. In skeleton graph, the

end vertices represent the visual feature points of an area feature. The skeleton paths show the connected characteristics of each pair visual feature points of area feature. The branch vertices are the topological link points of each visual parts. This is the base of our centrality definition of a polygon area.

3.1 Skeleton Graph Vertex Centrality

In graph theory, betweenness centrality is a central measurement of graph vertex based on the shortest path between vertices, which is defined as the number of shortest path through a vertex. In this study, we consider the visual coherence between visual feature parts of a polygon which can be indicated by the shortest path between visual feature vertex of a polygon. We define the betweenness centrality of skeleton graph vertex as the number of skeleton path through a vertex. Through the definition we can conclude that the maximum betweenness centrality vertex is a branch vertex if there are branches in the skeleton of a polygon. Thus, the calculation of betweenness centrality can be applied to the topological skeleton graph which can reduce the calculation complexity for fewer vertices.

► **Definition 1** (Betweenness Centrality of Skeleton Vertex). Betweenness centrality of vertex V is the number of skeleton path through V as:

$$C_b(V) = \sum_s \sum_t P_{s,V,t},$$

for $P_{s,V,t}$ is the skeleton path through V .

The closeness centrality of graph vertex measures the balance of all vertices to the specific vertex by the total length of shortest paths through the vertex. In this paper, we consider the balance between each visual feature vertex and the specific vertex. We define the standard deviation of all the weighted length of the specific vertex between each visual feature vertex. We can have three different kinds of closeness centrality when choosing different weight. The cover length indicates the elongation of the visual part shape of a polygon, and the cover width shows the width of the shape of visual parts, and cover area consider this two factors which are similar to Chen's method.

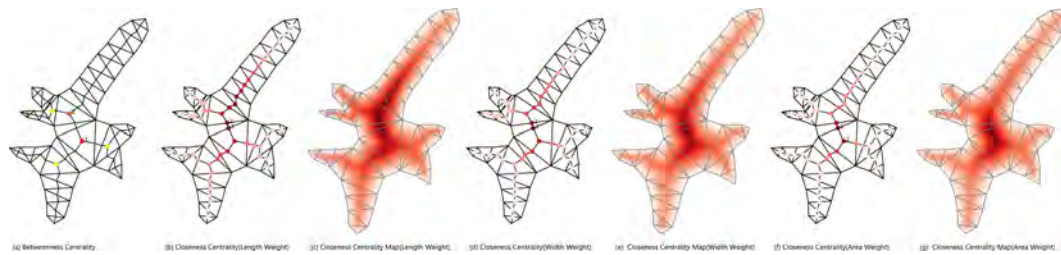
► **Definition 2** (Closeness Centrality of Skeleton Vertex). Closeness centrality of vertex V is the inverse of standard deviation of all the weighted lengths of paths from V to each end vertex as:

$$C_c(V) = \frac{1}{std(d_w(V, s))},$$

for $d_w(V, s)$ is the weighted length between V and end vertex s , the weight can be cover area, cover length, and cover width of graph edge.

3.2 Experiments and Results

To calculate the center point of a polygon, we first calculate the centrality of all vertex of the skeleton graph and find the vertex with the maximum value of centrality, which can be used as the center point of a polygon. In Figure 4, there are different center points by our algorithm. Betweenness centrality indicates the topological connections between visual feature parts of a polygon. The betweenness center point shows the center place have greatest topological importance. Closeness center point reflect the geometric nearness to feature



■ **Figure 4** Two kinds center points examples(a,b,d,f) and interpolate map (c,e,g).

points of a polygon. All the two kinds of center points are within the polygon and indicate the different visual center of a polygon.

In Figure 4(a, b, d, f), two kinds of centrality degree of the vertices are illustrated. We also generate closeness centrality pattern map of three different weight (c, e, g) which is calculated by linear interpolation with the centrality degree all the vertices of skeleton graph and the points of the boundary polygon. Our centrality illustrates the geometric visual center of area feature while the method by border number [10] is about urban structure center by road networks blocks.

4 Complexity Analysis and Discussion

4.1 Complexity Analysis

The calculation of centrality related to polygon triangulation and skeleton construction. According to the triangulation theory, we know that each triangulation has $n - 2$ triangles, and must have at least 2 type I (ear) triangles. If a triangulation contains e type I triangle, then $n \geq 2$, and the number of type III triangles is $e - 2$. By the definition of skeleton graph, each type I triangle and type II triangle form a skeleton edge, and each type III triangle form 3 skeleton edges. Therefore, the number of skeleton edge is $E = n - 2 + 2(e - 2)$. At extreme cases, there are only type I and type III triangles, that is $n - 2 = e + (e - 2)$, thus $2 \leq e \leq n/2$, and we can derive $n - 2 \leq E \leq 2n - 6$. A skeleton graph can also be regarded as a binary tree structure, therefore, the number of vertex and edge maintains $V = E + 1$.

The betweenness centrality needs the calculation of path between two end vertices. And this calculation based on the topological skeleton graph only contains the end vertices and branch vertices. Finding all the paths between end vertices, the complexity is $O(e(2e - 2))$. Under extreme cases in which the triangulation only contains type I and type III triangles, the complexity is $O(n^2)$.

For closeness centrality of vertices in skeleton graph, we need to find the all the path between the end vertices and all other vertices. To find the skeleton branch from each vertex to all end vertex needs a traverse of the skeleton graph, thus the complexity is $O(V)$, and $V = E + 1 = n + 2e - 3$. The number of link vertex and branch vertex is $n - e$. According to the range of e discussed above, we have the complexity of closeness centrality is $O((n - e)(n + 2e - 3)) \sim O(n^2)$.

4.2 Special Cases Discussions

We will consider some special cases. For an "H" shape polygon, the betweenness centrality may have two maximum vertices. In this situation, we can use the closeness center point which can give the difference. For a stripe shape polygon, there is no branch vertex in the

skeleton graph. Under this situation, we only consider the closeness center point. For the polygon with a hole, which is not a simple polygon, we can have a skeleton graph which contains ring structures. This kind of polygon will not be considered for the calculation of center points in this study.

5 Conclusion

This paper presents a definition and calculation of center point of area feature formed by simple polygon. The centrality of a polygon is defined based on the triangulation skeleton graph of a polygon. This method takes into account of the topological and geometric characteristics of visual feature points and parts of a polygon. The center point by this method is within the polygon and shows good visual center characteristics of an area feature. The method proposed has several issues need to be considered. One is the calculation complexity is higher than the mass-based center point in theory. Another is the situations when two candidates will occur. In the future study, we consider extending this paper considering formalizing the definition and comparing with other existing methods by cognitive experiments.


References

- 1 Tinghua Ai and Peter van Oosterom. Gap-tree extensions based on skeletons. In *Advances in Spatial Data Handling*, pages 501–513. Springer, Berlin, Heidelberg, 2002. doi:10.1007/978-3-642-56094-1_37.
- 2 Wolfgang Aigner, Franz Aurenhammer, and Bert Jüttler. On triangulation axes of polygons. *Information Processing Letters*, 115(1):45–51, 2015. doi:10.1016/j.ipl.2014.08.006.
- 3 Xiang Bai and Longin Jan Latecki. Path similarity skeleton graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(7):1282–1292, 2008. doi:10.1109/TPAMI.2007.70769.
- 4 Mark de Berg, Otfried Cheong, Marc van Kreveld, and Mark Overmars. *Computational Geometry: Algorithms and Applications*. Springer, Santa Clara, CA, USA, 3 edition, 2008.
- 5 Harry Blum. Biological shape and visual science (part 1). *Journal of Theoretical Biology*, 38(2):205–287, 1973. doi:10.1016/0022-5193(73)90175-6.
- 6 Stephen P. Borgatti and Martin G. Everett. A graph-theoretic perspective on centrality. *Social Networks*, 28(4):466–484, 2006. doi:10.1016/j.socnet.2005.11.005.
- 7 Tao Chen and Tinghua Ai. Automatic extraction of skeleton and center of area feature. *Geomatics and Information Science of Wuhan University*, 29(5):443, 2004. doi:10.13203/j.whugis2004.05.015.
- 8 Jesús A. De Loera, Jörg Rambau, and Francisco Santos. *Triangulations: Structures for Algorithms and Applications*, volume 25 of *Algorithms and Computation in Mathematics*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. doi:10.1007/978-3-642-12971-1.
- 9 Daniel Garcia-Castellanos and Umberto Lombardo. Poles of inaccessibility: A calculation algorithm for the remotest places on earth. *Scottish Geographical Journal*, 123(3):227–233, 2007. doi:10.1080/14702540801897809.
- 10 Bin Jiang and Xintao Liu. Scaling of geographic space from the perspective of city and field blocks and using volunteered geographic information. *International Journal of Geographical Information Science*, 26(2):215–229, feb 2012. doi:10.1080/13658816.2011.575074.
- 11 Jia-Guu Leu. Computing a shape's moments from its boundary. *Pattern Recognition*, 24(10):949–957, jan 1991. doi:10.1016/0031-3203(91)90092-J.

The Use of Particle Swarm Optimization for a Vector Cellular Automata Model of Land Use Change


Yi Lu

University of New South Wales, Sydney, Australia
yi.lu@unsw.edu.au

 <https://orcid.org/0000-0002-2090-9057>

Shawn Laffan

University of New South Wales, Sydney, Australia
shawn.laffan@unsw.edu.au

 <https://orcid.org/0000-0002-5996-0570>

Abstract

Cellular automata (CA) is an important area of research in GIScience, with recent research developing vector-based models in addition to the traditional raster data formats. One active area of research is the calibration of transition rules, particularly when applied to vector CA. Here we evaluate a particle swarm optimization (PSO) process to calibrate a vector CA model of land use change for a sub-region of Ipswich in Queensland, Australia, for the period 1999-2016. We compare the results with those for a raster CA of the same dataset. The spatial indices of the vector PSO-CA model exceed that of the raster model, with spatial accuracies being 82.45% and 76.47%, respectively. In addition, the vector PSO-CA model achieved a higher kappa coefficient. Vector-based PSO-CA model can be used for the exploration of urbanization process and provide a better understanding of land use change.

2012 ACM Subject Classification Computing methodologies → Modeling methodologies

Keywords and phrases Vector cellular automata (CA), Particle swarm optimization (PSO), Land use simulation, Ipswich

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.42

Category Short Paper

Acknowledgements The authors would like thank the support from China Scholarship Council (CSC) and Queensland Spatial Catalogue.

1 Introduction

Cellular automata (CA) are widely used models of dispersal processes, with example applications including disease spread [6], forest fire spread [8], land use change [2, 15], traffic flow simulation [1], planning support systems [12]. Of these topics, the integration of CA and land use change analysis is of considerable significance given issues of globalization and the expansion of the human population. There are already several case studies applying the method to metropolitan areas, for example in Australia [17], Canada [21], China [23], Europe [5] and the USA [13].

The definition of transition rules, which determine the state conversion of geographical features during simulation, is a core component of CA modelling [16]. A variety of artificial intelligence and evolutionary algorithms have been used to calibrate land use transformation



© Yi Lu and Shawn Laffan;

licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 42; pp. 42:1–42:6

Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

rules in an efficient and objective way, including artificial neural networks (ANN), ant colony optimization (ACO), bee colony optimization (BCO), cuckoo search (CS), decision tree (DT), genetic algorithm (GA), multi-agent system (MAS), particle swarm optimization (PSO), and support vector machines (SVM). It is generally accepted that such methods offer a capacity to minimize the disagreement between the simulations and reference maps, resulting in a set of optimized transition rules and thus improving their accuracy for urban modelling [10]. Nonetheless, most of the above-mentioned methodologies have been validated with raster CA models. There are very few analyses of the integration of evolutionary algorithms and intelligent optimization with vector CA models.

Here we report on the implementation of CA models calibrated using PSO, implemented as both vector and raster formats. The parameters and simulation processes are compared between the two formats using a case study in Queensland, Australia. The analyses were implemented using a prototype system developed using ArcEngine 10.3 and C#.

2 Particle swarm optimization (PSO)

Particle swarm optimization is a widely used intelligent optimization method in artificial intelligence algorithms, an important research area in GIScience. This method explores the optimal solution of problems with regard to a given measure of quality. The method was first proposed by Kennedy and Eberhart [7], and then developed by Shi and Eberhart [19] to enhance the efficient search for a globally optimal solution with inertia weights. The basic unit in the PSO method is the ‘particle’, which refers to one of the potential solutions in the model, and can be described as:

$$particle = (v_n, P_n) \quad (1)$$

where n is the dimensionality of the target problem, v_n and P_n are the velocity and position of a particle at a specific time point. Specifically, v_n can be described as the combination of n velocities (in n dimensions) at time t :

$$v_n = (v_1, v_2, \dots, v_n, t) \quad (2)$$

and similarly, P_n can be represented by n positions in a n -dimensional space at time t :

$$P_n = (P_1, P_2, \dots, P_n, t) \quad (3)$$

Furthermore, the velocity and position of each particle will be updated according to individual and global best positions:

$$\begin{cases} \dot{v}(t+1) = w * v(t) + c1 * rand * (P_{ib} - P(t)) + c2 * rand * (P_{gb} - P(t)) \\ P(t+1) = P(t) + v(t+1) \end{cases} \quad (4)$$

where w is the weight of velocity at time t , $c1$ and $c2$ are two constant weights which are set in advance, and $rand$ is a randomly generated number in the interval $[0, 1]$. P_{ib} is the best individual position of particle i , and P_{gb} is the best global position of particle swarm, namely the best one of all best individual positions. In addition, $v(t+1)$ is the velocity of a particle at time $t+1$, $P(t)$ and $P(t+1)$ are the positions of particle at time t and $t+1$, separately.

3 Case study

3.1 Study area

The study area for this research comprises two districts (Collingwood Park and Redbank Plains) of Ipswich city, with an area of 2,571 ha. Ipswich City is the second oldest local

■ **Table 1** Driving factors of land use change in the study area.

Driving factors	Definition	
	Raster PSO-CA	Vector PSO-CA
disCom	Distance to commercial service	
disPub	Distance to public service	
disHw	Distance to highways	
disSr	Distance to secondary roads	
neigh	5×5 Moore Neighbour	Parcels intersecting a 60 m buffer zone around the 1999 residential cells
popDen	The changed density of population within a parcel over the past decade	
area	not applicable	Area of parcel

government area in the Brisbane-South East Queensland (SEQ), one of the fastest growing metropolitan region in Australia [22, 14]. The current population of approximately 200,000 in Ipswich is projected to double by 2031 [11].

In general, there are 17 land use classes in the study area. During 1999-2016, the main land use transformation was from ‘grazing native vegetation’ and ‘residual native cover’ to ‘residential area’. The area of grazing native vegetation decreased by 233.95 ha over this period, representing 76.03% of all changed land use. Residual native cover is the category with the second largest reduction, at 66.3 ha, or 21.55% of the entire decreased category. There is also a 200-ha increase of residential area during the same period, which is 64.76% of all increased land use types.

3.2 Driving factors

The general objective of this study is to model the set of non-residential cells (in 1999) which were transformed in 2016, thus for the purposes of analysis, ‘grazing native vegetation’ and ‘residual native cover’ were reclassified as ‘non-residential’. Between 1999 and 2016, 188.47 ha of non-residential land was transformed to ‘residential’, while 809.77 ha of non-residential land remained unchanged. These ‘non-residential’ and ‘residential’ land parcels are defined as the vector cells of our CA model. The raster data sets were derived by converting the land use map into 30 m grid cells, which is consistent with the Landsat Thematic Mapper data commonly used to derived such maps.

The position values of a single PSO particle correspond to one possible combination of weights. Therefore, the number of dimensions is equal to the number of driving factors of land use change in the study area (Table 1). There are six common driving factors to both the raster and vector CA models, and vector CA having an additional driving factor to represent parcel area.

3.3 Simulations

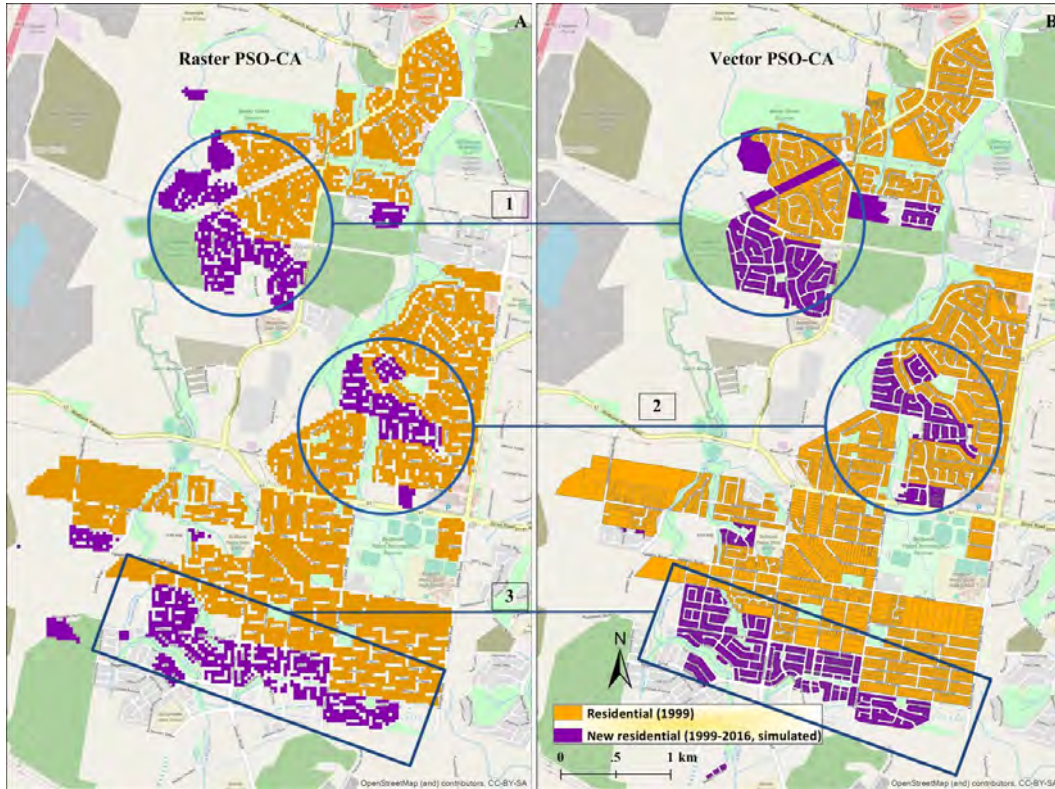
On the basis of previous work by Feng et al. [10] and experimentation, the values of w , c_1 , c_2 were set as 1, 1.2 and 1.2, which represent the contributions of the current velocity and best position of a particle, as well as the best position of particle swarm. 50% of the transformed and non-transformed non-residential cells were randomly selected and normalized as the sample of PSO training. Derived weights for the two models are given in Table 2.

The transfer probability P_{tran} , of non-residential cells is calculated as:

$$P_{tran} = \frac{1}{1 + \exp((-1) * \sum_n^1 (w_i * x_i))} \quad (5)$$

■ **Table 2** Weights of vector and raster PSO-CA models.

Driving factors		disCom	disPub	disHw	disSr	neigh	popDen	area
Weight values	Raster PSO-CA	0.023	-2.181	0.270	-0.970	2.500	-0.320	na
	Vector PSO-CA	0.025	0.312	0.390	0.080	0.650	0.880	0.800



■ **Figure 1** Simulation results of raster (A) and vector (B) PSO-CA models (Base map: OpenStreetMap).

■ **Table 3** Spatial accuracies of PSO-CA models.

Model type	Spatial accuracy (%)	Kappa coefficient
Raster PSO-CA	76.47	0.886
Vector PSO-CA	82.45	0.916

where w_i is the weight of corresponding driving factor, x_i is the normalized value of a non-residential cell. P_{tran} is in the interval $[0, 1]$. A larger number of iterations, which means a shorter iteration interval, are required for completing CA-based simulations [3]. Accordingly, the number of iteration is set as 100 in this study.

The simulation results (Figure 1) show that the general distribution of simulated new residential cells is similar. Specifically, these cells are located in the north (part 1), east (part 2) and south (part 3) of the study area. These are near the 1999 residential areas, consistent with previous studies [9].

Two spatial indices, spatial accuracy and kappa coefficient, are calculated using observed and real land use for years 1999 and 2016. The value of spatial accuracy indicates the

proportion of correctly predicted new residential cells, and the inter-rater agreement for cell categories is demonstrated by kappa coefficient [4]. It is clear from the results that the vector CA performs better than the raster CA (Table 3).

4 Discussion and conclusion

The weights of driving factors, which describe their contribution to the transformation of non-residential cells during the period 1999 to 2016, is the main reason for the differences between simulation results. In the raster PSO-CA, neighbouring cells have the largest positive contribution to land use transformation, which is as much as 2.5. Distances to highways and commercial service are the second and third positive driving factors, but only with values of 0.270 and 0.023, respectively. The remaining three driving factors have negative influences on land use transformation from non-residential to residential in raster PSO-CA. Apart from the inconsistency of weight values in the raster PSO-CA, all the seven driving factors of vector PSO-CA have positive influences on the same type of land use transformation, where population growth (0.880), area of cell (0.800) and neighbouring cells (0.650) ranking in the top three. The vector PSO-CA models are more reasonable considering the fact that land use transformation is usually dependent on a series of spatial variables in terms of accessibilities or proximities [15, 20].

The spatial accuracy of PSO-CA is 5.98% higher for the vector format (Table 3). In addition, the kappa coefficient for the raster PSO-CA is also 0.03 lower than the vector CA. Therefore, the vector-based PSO-CA has the capability to produce a more accurate prediction of land use change, which is consistent with previous research on vector CA [18].

In this paper, the effect of data format on PSO-CA model has been assessed by taking a sub-region of Ipswich, Southeast Queensland, Australia. Considering the weights of driving factors, spatial accuracy and kappa coefficients, vector-based PSO-CA achieves a higher accuracy of simulation, which produces a more realistic model of the expansion of residential area. Future research will have a further exploration of the uncertainties of random disturbance [10], which could lead to a different simulation result (such as another combination of driving weights).

References


- 1 Robert Barlovic, Ludger Santen, Andreas Schadschneider, and Michael Schreckenberg. Metastable states in cellular automata for traffic flow. *The European Physical Journal B-Condensed Matter and Complex Systems*, 5(3):793–800, 1998.
- 2 Michael Batty and Yichun Xie. From cells to cities. *Environment and Planning B: Planning and Design*, 21(7):S31–S48, 1994.
- 3 Min Cao, Guo'an Tang, Quanfei Shen, and Yanxia Wang. A new discovery of transition rules for cellular automata by using cuckoo search algorithm. *International Journal of Geographical Information Science*, 29(5):806–824, 2015.
- 4 Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213, 1968.
- 5 Eric de Noronha Vaz, Peter Nijkamp, Marco Painho, and Mário Caetano. A multi-scenario forecast of urban change: A study on urban growth in the Algarve. *Landscape and Urban Planning*, 104(2):201–211, 2012.
- 6 Ruth J Doran and Shawn W Laffan. Simulating the spatial dynamics of foot and mouth disease outbreaks in feral pigs and livestock in Queensland, Australia, using a susceptible-infected-recovered cellular automata model. *Preventive Veterinary Medicine*, 70(1-2):133–152, 2005.

- 7 Russell Eberhart and James Kennedy. A new optimizer using particle swarm theory. In *Micro Machine and Human Science, 1995. MHS'95., Proceedings of the Sixth International Symposium on*, pages 39–43. IEEE, 1995.
- 8 L Hernández Encinas, S Hoya White, A Martín Del Rey, and G Rodríguez Sánchez. Modelling forest fire spread using hexagonal cellular automata. *Applied Mathematical Modelling*, 31(6):1213–1227, 2007.
- 9 Yongjiu Feng and Yan Liu. A cellular automata model based on nonlinear kernel principal component analysis for urban growth simulation. *Environment and Planning B: Planning and Design*, 40(1):117–134, 2013.
- 10 Yongjiu Feng, Yan Liu, Xiaohua Tong, Miaolong Liu, and Susu Deng. Modeling dynamic urban growth using cellular automata and particle swarm optimization rules. *Landscape and Urban Planning*, 102(3):188–196, 2011.
- 11 Queensland Government. South East Queensland Regional Plan 2009–2031, June 2009.
- 12 Richard E Klosterman. The what if? Collaborative planning support system. *Environment and Planning B: Planning and Design*, 26(3):393–408, 1999.
- 13 Verda Kocabas and Suzana Dragicevic. Assessing cellular automata model behaviour using a sensitivity analysis approach. *Computers, Environment and Urban Systems*, 30(6):921–953, 2006.
- 14 Tiebei Li, Jonathan Corcoran, David Pullar, Alistair Robson, and Robert Stimson. A geographically weighted regression method to spatially disaggregate regional employment forecasts for South East Queensland. *Applied Spatial Analysis and Policy*, 2(2):147–175, 2009.
- 15 Xia Li and Anthony Gar-On Yeh. Neural-network-based cellular automata for simulating multiple land use changes using gis. *International Journal of Geographical Information Science*, 16(4):323–343, 2002.
- 16 Xiaoping Liu, Xia Li, Lin Liu, Jinqiang He, and Bin Ai. A bottom-up approach to discover transition rules of cellular automata using ant intelligence. *International Journal of Geographical Information Science*, 22(11-12):1247–1269, 2008.
- 17 Yan Liu, Yongjiu Feng, and Robert Gilmore Pontius. Spatially-explicit simulation of urban growth through self-adaptive genetic algorithm and cellular automata modelling. *Land*, 3(3):719–738, 2014.
- 18 Niandry Moreno, Fang Wang, and Danielle J Marceau. Implementation of a dynamic neighborhood in a land-use vector-based cellular automata model. *Computers, Environment and Urban Systems*, 33(1):44–54, 2009.
- 19 Yuhui Shi and Russell Eberhart. A modified particle swarm optimizer. In *Evolutionary Computation Proceedings, 1998. IEEE World Congress on Computational Intelligence., The 1998 IEEE International Conference on*, pages 69–73. IEEE, 1998.
- 20 Fang Wang, Jean-Gabriel Hasbani, Xin Wang, and Danielle J Marceau. Identifying dominant factors for the calibration of a land-use cellular automata model using rough set theory. *Computers, Environment and Urban Systems*, 35(2):116–125, 2011.
- 21 Fang Wang and Danielle J Marceau. A patch-based cellular automaton for simulating land-use changes at fine spatial resolution. *Transactions in GIS*, 17(6):828–846, 2013.
- 22 Douglas Ward, Stuart R Phinn, and Alan T Murray. Monitoring growth in rapidly urbanizing areas using remotely sensed data. *The Professional Geographer*, 52(3):371–386, 2000.
- 23 Yongke Yang, Pengfeng Xiao, Xuezhi Feng, and Haixing Li. Accuracy assessment of seven global land cover datasets over China. *ISPRS Journal of Photogrammetry and Remote Sensing*, 125:156–173, 2017.

Towards a Comprehensive Temporal Classification of Footfall Patterns in the Cities of Great Britain


Karlo Lugomer¹

Department of Geography, University College London
Pearson Building, Gower Street, WC1E 6BT, London, United Kingdom
karlo.lugomer.14@ucl.ac.uk

 <https://orcid.org/0000-0002-0820-3772>

Paul Longley²

Department of Geography, University College London
Pearson Building, Gower Street, WC1E 6BT, London, United Kingdom
p.longley@ucl.ac.uk

 <https://orcid.org/0000-0002-4727-6384>

Abstract

The temporal fluctuations of footfall in the urban areas have long been a neglected research problem, and this mainly has to do with the past technological limitations and inability to consistently collect large volumes of data at fine intra-day temporal resolutions. This paper makes use of the extensive set of footfall measurements acquired by the Wi-Fi sensors installed in the retail units across the British town centres, shopping centres and retail parks. We present the methodology for classifying the diurnal temporal signatures of human activity at the urban microsite locations and identify characteristic profiles which make them distinctive regarding when people visit them. We conclude that there exist significant differences regarding the time when different locations are the busiest during the day, and this undoubtedly has a substantial impact on how retailers should plan where and how their businesses operate.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases temporal classification, temporal profiles, time series cluster analysis, Wi-Fi sensors, retail analytics

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.43

Category Short Paper

Acknowledgements This research was sponsored by the UK Economic and Social Research Council (ESRC) Consumer Data Research Centre (ES/L011840/1) and a Ph.D. studentship co-funded by the ESRC and the Local Data Company.

1 Introduction

Spatial classifications have been a subject of a wide range of research papers in geography and GIScience. The popularity of clustering can be justified by the vast amount of readily available spatial data and need for interesting characteristics and patterns extraction [4]. Such classifications aim to describe the extent to which place A is similar to place B and to

¹ Award ES/L011840/1

² Award ES/L011840/1



© Karlo Lugomer and Paul Longley;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 43; pp. 43:1–43:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

use the derived clustering solution to make predictions about the characteristics of locations where data are incomplete and thus inform the industrial or public planning policymakers.

While the geographical classifications have been extensively covered in the past literature, little has been done to characterise bigger samples of places based on the recorded activity patterns on the finer temporal scales. In the past, this could have been done only by manual surveying, which is a costly and laborious process and does not enable the continuous data acquisition. These shortcomings have been addressed after the rapid development and wide-scale adoption of smartphones and Wi-Fi, GPS and Bluetooth technologies, which together made possible the collection of high volumes of data at small time periods, while, regarding spatial resolution, coming even to the granularity of an individual.

Knowing about where people go at which times in the weekly, daily or (sub-)hourly time frames has great practical importance for many fields. A good example of a sector where this is particularly relevant is retailing. Knowing what time of the day a specific retail unit can expect to see the highest number of potential customers passing by is vital to understanding whether that particular location is suitable for a specific category of retail business. For example, pubs and bar operators will be more interested in the places where footfall is significant in the evenings. This is contrary to the coffee shop operators, which will seek to exploit the large flow of morning commuters and midday lunch and coffee consumers.

This paper aims to use the footfall measurements collected by the Wi-Fi sensors to characterise urban microsite locations based on the features of the recorded temporal signatures of footfall. In other words, we are interested in finding out whether urban locations tend to differ in terms of diurnal temporal distribution of footfall and if so, how common each profile is. This classification presents the first step in acquiring a broader understanding of how urban places function and why people tend to find themselves at particular places at particular times of the day or days of the week.

2 National footfall data set

The data for this project were acquired through the network of Wi-Fi sensors installed by the Local Data Company (LDC) in the different UK cities from July 2015 until August 2017. They were placed in the three different categories of retail centres: shopping centres, out-of-town retail parks and, most commonly, urban town centres, i.e. high streets.

The initial set of retail centres for sensor installations was chosen based on the research sample design tailored to incorporate different cities of Great Britain, capturing centres of different sizes and diverse set of geodemographic characteristics of their catchment areas. The criteria for the sample locations outside London were dominant Output Area Classification (OAC) Supergroup, which is based on the cluster analysis of the 2011 Census variables [3]; town centre size expressed by the number of businesses and the town centre type, i.e. position of the centre in the national hierarchy. The primary criterion for the locations in Greater London was, on the other hand, the population size of retail centres' respective catchment areas. The Wi-Fi sensors were placed inside the retail units as close to the storefront window as possible.

2.1 Data acquisition

The Wi-Fi sensors work by receiving the probe requests sent out by the smartphones that are scanning for the available Wi-Fi networks. When a pedestrian carrying a smartphone with Wi-Fi and background scanning turned on passes by the Wi-Fi sensor, the sensor records the data contained in that probe request. The data includes the time stamp, the device

signal strength and the MAC address, which is hashed at the sensor level to preserve the privacy of the device owners. The idea is to derive the accurate measurements of the number of passers-by, monitor their fluctuations over time and use them to characterise locations based on their temporal distribution.

2.2 Data pre-processing

The approach described in the previous subsection comes with limitations, as derived footfall is prone to measurement errors due to factors which cause overcounting or undercounting [7].

Overcounting is caused by the fact that Wi-Fi sensors typically capture probe requests from devices which dwell locally (for example, workers in the retail unit and surrounding offices, devices other than smartphones such as printers, etc.). Undercounting stems from the fact that some passers-by do not have Wi-Fi probing capabilities enabled on their smartphones or they are simply missed due to the presence of some physical obstructions or signal interferences. The overcounting factors can be eliminated automatically by filtering methods and undercounting factors can to a certain extent be accounted for by the calibration in which passers-by are counted manually on site. After that, the ground truth is compared to the filtered sensor measurements, an adjustment factor is calculated by dividing those two figures and then used to adjust the measures. A more detailed treatment of those factors and ways to eliminate them is given in [7] and [9].

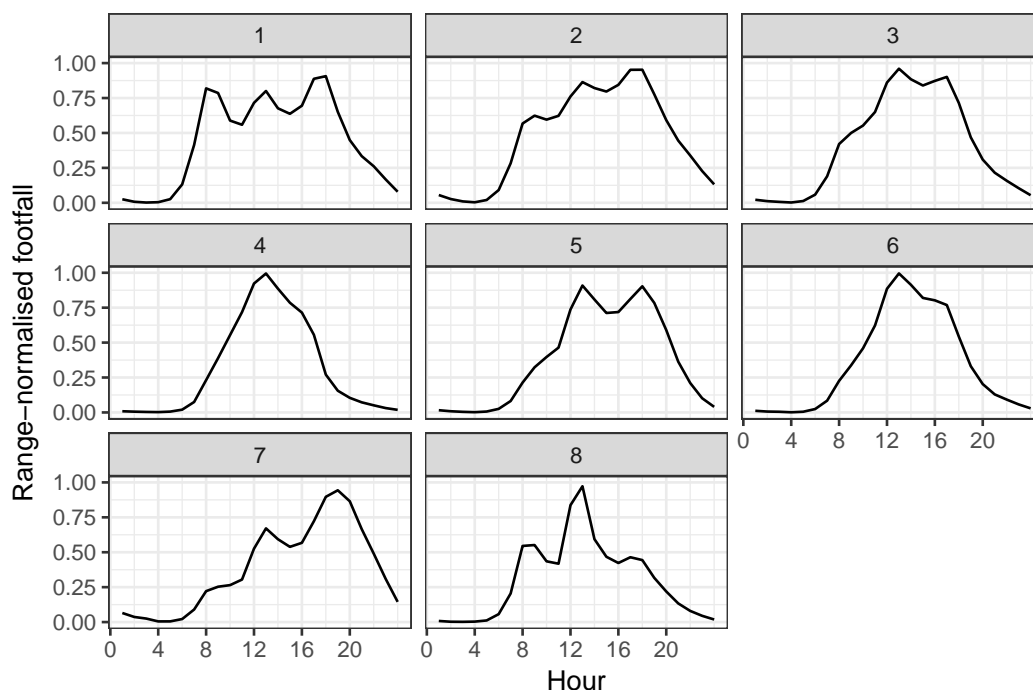
After identifying the devices of interest which serve as the proxy for people, the data were cleaned from outliers, as in this case we are interested in detecting the general functional characteristics of the location, rather than unusual events. The missing data were inputted by linear interpolation or inferred by taking the historical data for the corresponding hours and days of the week in cases where gaps of missing data were too wide for reliable interpolation. One representative weekly profile was then generated for every location by taking the median of every hour separately. The result comprised averaged time series each comprising 168 hours of the week for each of the 605 selected locations.

3 Clustering methodology

Since temporal profiles of different days of the week differ, it is not sensible to create a temporal classification for a "typical, average day" for each location. When the variation of footfall across time is visually inspected at the chosen location, Mondays through Thursdays generally display mutually similar profile shapes, whereas Fridays begin to differ if that location has pronounced nightlife activity. Same is true for Saturdays; however, due to the absence of the majority of workers, the daytime activity profile is usually different. In the first instance, the classification was therefore conducted for the footfall between Mondays and Thursdays for each location. The previously cleaned data were range-normalised.

The next step was to choose from the myriad of distance measures and clustering algorithms suitable for the time series clustering [2][5][6][8].

According to [1], the distance measures are commonly classified as (dis)similarities in either time, shape or change. The similarity in time can be regarded as a special case of similarity in shape, so the two go under the collective term shape-based methods [8][1]. In our case, we are interested in detecting the clusters of similar shapes of footfall profiles, however, at the same time, knowing at what time peaks or troughs occurred is also relevant. That said, we examined the shape-based methods more closely and Dynamic Time Warping (DTW) and Euclidean distances (ED) were found to be the most useful for our particular problem. A further justification for this is found in the recent detailed comparison of the



■ **Figure 1** Temporal profiles of microsite locations (*data source*: Local Data Company (2015–2017)).

different distance measures [2], in which it was concluded that despite some plausible progress made in the time series classification domain, DTW remains hard to beat and it is at the same time computationally less intensive than some of the newly proposed methods such as the Collection of Transformation Ensembles (COTE). In addition, it was found that, on reasonably large data sets comprising thousands (and in some cases only hundreds) of series, the difference between the classification error rate of the DTW and the ED diminishes [11]. In our case, the cleaned data set comprises 605 locations, which means that while warping may be advantageous, the ED could still suffice. Both ED and DTW with a relatively small width of the warping window equal to one hour were tested and coupled with several different partitional and hierarchical methods (k-means, PAM and Ward’s method). The ED fed into Ward’s algorithm provided the best trade-off between the mathematical validity, as measured by clustering validity indices [10] and interpretability.

4 Results and discussion

The optimal clustering solution was found to comprise eight distinct temporal profiles, as shown in the Figure 1.

The number of cases across clusters is unevenly distributed (Table 1), however, since we aimed to detect the interesting functional differences between places, trying to balance the number of cases would produce clusters in which such interesting properties would have been inherently lost.

According to the Table 1, the most common temporal profile in the retail centres of Great Britain (27.93% of the sampled microsite locations) is a two-peaked profile with a maximum around midday and late afternoon - appropriately labelled as *Consistent afternoons*. Unlike with similar profiles, such as One-directional commute, the drop of footfall during the early

■ **Table 1** The breakdown of cluster cases.

Cluster	Proposed name	Cases	Percentage (%)
1	Commute and lunch	84	13.88
2	Gradual rise	80	13.22
3	Consistent afternoons	169	27.93
4	Midday top	119	19.67
5	One-directional commute	29	4.79
6	Lunch time with minor afternoon commuter inflow	90	14.88
7	Quiet mornings, busy evenings	19	3.14
8	Busy lunchtimes with both commuting peaks	15	2.48
	Total	605	100.00

afternoon, i.e. between 2 pm and 5 pm is almost insignificant, which means that such locations benefit from consistently high footfall throughout most of the day. The second most common temporal profile (*Midday top*, comprising 19.67% locations) is a simple one-peaked profile with maximum activity recorded around midday. Such locations likely attract lunch goers. Next cluster is *Lunch time with minor afternoon commuter inflow*, comprising 14.88% of the locations. It is a one-peaked profile with a minor secondary peak in the late afternoon, which is not strictly speaking a peak, but rather a part of the profile where a drop of footfall slows down due to the impact of late afternoon commuters. However, in these locations, commuters are not as numerous as is the case in some other locations, so secondary peaks are not formed.

Similarly numerous, clusters 1 (*Commute and lunch*) and 2 (*Gradual rise*) account for 13.88% and 13.22% of the locations, respectively. Both are three-peaked profiles and are characterised by busier customer traffic during all three characteristic periods during the day - morning rush hour, lunchtime and afternoon rush hour. The difference is that Gradual rise cluster expects more customers towards the end of the day and intra-day differences of footfall volume are not as pronounced. Commute and lunch, on the other hand, has more pronounced peaks and intermediate drop and corresponding locations may expect the similar volume of passing footfall during all three periods, with a peak in the late afternoon recording slightly higher footfall than other two peaks.

The profiles captured by the remaining three minor clusters are not as commonly encountered across the British retail space, however, since they are functionally specific, it is worth further investigating their temporal distribution of footfall.

As was already mentioned, *One-directional commute* cluster is characterised by the two-peaked profiles of microsite locations (4.79%) with a more significant drop in customer traffic after the lunchtime, as compared to the similarly shaped Consistent afternoons cluster. Interestingly, these locations do not record any peak during the morning rush hour but do record one during the afternoon rush hour. Next, *Quiet mornings, busy evenings* cluster (3.14%) is to a certain extent similar to the Gradual rise locations, but morning footfall is much smaller, and differences between the peaks are much more pronounced. Moreover, the maximum footfall is, on average, reached between 7 pm and 8 pm, which seemingly makes these locations more attractive for the dinner and pub goers. And finally, occurring at only 15 of the sampled locations (2.48%), *Busy lunchtimes with both commuting peaks* is characterised by its distinctive dominant lunchtime peak and two smaller peaks during the rush hours.

5 Conclusion and future work

The initial aim of this paper was to test whether different microsite locations in urban areas display different diurnal footfall patterns and if that was the case, to further inspect if the readings from the Wi-Fi sensors could serve to derive the temporal classification of footfall patterns. This cluster analysis proved that there exist significant differences in footfall patterns among urban microsite locations. We identified eight clusters of distinct functional characteristics and described each of them.

As part of the future work, we aim to combine the identified profiles with the ancillary data on local vacancy rates, retail occupancy structure, i.e. local compositions of store types, in addition to the relative distributions of footfall that were presented here. The geodemographic characteristics of the retail centre catchment areas or the underlying Workplace Zones will also be considered as the relevant factors worth further investigation. The ultimate goal is to identify and explain the functional characteristics of the national set of retail centres based on both structural and dynamical properties of space.

References

- 1 Anthony Bagnall, Eamonn Keogh, Stefano Lonardi, Gareth Janacek, et al. A bit level representation for time series data mining with shape based similarity. *Data Mining and Knowledge Discovery*, 13(1):11–40, 2006.
- 2 Anthony Bagnall, Jason Lines, Aaron Bostrom, James Large, and Eamonn Keogh. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 31(3):606–660, 2017.
- 3 Christopher G Gale, A Singleton, Andrew G Bates, and Paul A Longley. Creating the 2011 area classification for output areas (2011 oac). *Journal of Spatial Information Science*, 12:1–27, 2016.
- 4 Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of intelligent information systems*, 17(2-3):107–145, 2001.
- 5 Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.
- 6 T Warren Liao. Clustering of time series data—a survey. *Pattern recognition*, 38(11):1857–1874, 2005.
- 7 Karlo Lugomer, Balamurugan Soundararaj, Roberto Murcio, James Cheshire, and Paul Longley. Understanding sources of measurement error in the wi-fi sensor data in the smart city. In *Proceedings of the 25th GIS Research UK (GISRUK) Conference, Manchester, UK, April 18-21, 2017*, 2017.
- 8 Pablo Montero, José A Vilar, et al. Tsclust: An r package for time series clustering. *Journal of Statistical Software*, 62(1):1–43, 2014.
- 9 Roberto Murcio, Bala Soundararaj, and Karlo Lugomer. Movements in cities: Footfall and its spatio-temporal distribution. In Paul Longley, James Cheshire, and Alex Singleton, editors, *Consumer Data Research*, chapter 6, pages 85–96. UCL Press, London, 2018.
- 10 Mark A Newell, Dianne Cook, Heike Hofmann, and Jean-Luc Jannink. An algorithm for deciding the number of clusters and validation using simulated data with application to exploring crop population structure. *The Annals of Applied Statistics*, pages 1898–1916, 2013.
- 11 Jin Shieh and Eamonn Keogh. i sax: indexing and mining terabyte sized time series. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 623–631. ACM, 2008.

Is This Statement About A Place? Comparing two perspectives^{*†}

Alan M. MacEachren

GeoVISTA Center, Dept. of Geography, Penn State University, University Park, PA, USA
maceachren@psu.edu

Richard Caneba

College of Information Science & Technology, Penn State University, University Park, PA, USA

Hanzhou Chen

Department of Geography, Penn State University, University Park, PA 16802, USA

Harrison Cole

Department of Geography, Penn State University, University Park, PA 16802, USA

Emily Domanico

Department of Geography, Penn State University, University Park, PA 16802, USA

Nicholas Triozzi

Dept. of Anthropology, Penn State University, University Park, PA 16802, USA

Fangcao Xu

Department of Geography, Penn State University, University Park, PA 16802, USA

Liping Yang

Department of Geography, Penn State University, University Park, PA 16802, USA

Abstract

Text often includes references to places by name; in prior work, more than 20% of a sample of event-related tweets were found to include place names. Research has addressed the challenge of leveraging the geographic data reflected in text statements, with well-developed methods to recognize location mentions in text and related work on automated toponym resolution (deciding which place in the world is meant by a place name). A core issue that remains is to distinguish between text that mentions a place or places and text that is about a place or places. This paper presents the first step in research to address this challenge. The research reported here sets the conceptual and practical groundwork for subsequent supervised machine learning research; that research will leverage human-produced training data, for which a judgment is made about whether a statement is or is not about a place (or places), to train computational methods to do this classification for large volumes of text. The research step presented here focuses on three questions: (1) what kinds of entities are typically conceptualized as places, (2) what features of a statement prompt the reader to judge a statement to be about a place (or not about a place) and (3) how do judgments of whether or not a statement is about a place compare between a group of experts who have studied the concept of “place” from a geographic perspective and a cross-section of individuals recruited through a crowdsourcing platform to make these judgments.

2012 ACM Subject Classification Information systems → Information retrieval, Computing methodologies → Natural language processing

* The first author led the research within a graduate seminar on Place & Big Data that the next 6 authors participated in (with equal contribution, listed in alphabetical order). The last author leads the machine learning process to follow. All authors contributed to writing and/or editing parts of the paper.

† The crowdsourced data collection component of the research was reviewed by the Penn State Institutional Review Board and determined to not require formal IRB review, as it meets criteria for exempt research.



Keywords and phrases geographic information retrieval, spatial language, crowdsourcing

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.44

Category Short Paper

1 Introduction

The research reported here has two primary goals. The first extends beyond and motivates the present paper – to develop geographic information retrieval (GIR) methods to retrieve place-focused unstructured information from text. Our longer term project related to this goal is to explore the potential of machine/deep learning methods to categorize statements into those “about place” (or not). Work reported here is a precursor to that objective. Our second goal, the primary focus of the project reported on here, is to explore the concept of place and what it means for a statement to be “about” a place. To address this objective, we: (a) consider examples of places and attributes that lead to an entity being considered to be a place or not, (b) assess the extent to which a set of individuals with scientific understanding of place as a concept agree on whether short statements (in Twitter) are about place or not and (c) evaluate the potential to use Amazon Mechanical Turk (MTurk) crowdsourcing to build large corpora of statements classified into those that are or are not about a place (for subsequent use in training and testing of machine/deep learning).

2 What is a place?

Place has been a core concept of Geography for centuries. Trying to define “place” in a way that appeals across multiple disciplines has been a beguiling problem for geographers [5]. From a humanist perspective, Tuan [11] defined place as “spatial locations that have been given meaning by human experience.” Golledge [3], from a behavioral science perspective, contended that “although place is a dimensionless spatial term, it is conventionally interpreted as a multidimensional phenomenon (emphasis added).” From a social perspective, place can be characterized as an emergent phenomenon, its evolution is non-linear and shaped by many, varying perspectives, constructed and made tangible by social processes and historical narratives, see: [8]. In spite of many efforts to define place, the concept has been difficult to formalize sufficiently to leverage digital data for understanding place as a dynamic construct [4]. Here, we focus on exploring place-related discourse in language. For a broader overview of place in the context of GIScience and Big Data, see: [7].

3 Typical “places”

As a discussion starting point in a Place & Big Data seminar, 6 students (co-authors) completed two tasks in successive weeks. The first focused on listing and categorizing “places,” the second on listing attributes that distinguish places from other entities. Entities proposed as places varied in scale (from the Treaty Oak, through countries, to The Universe). Some entities were uniquely personal (e.g., “the secret fort near my house growing up”). Others, while personally relevant were also prototypical examples of local places (e.g., “Flightpath Coffee”). Some entities, while locations one can be at or in, are also prominent landmarks (e.g., “Golden Gate Bridge”, Taj Mahal).

One parsing of entities listed is to apply Montello’s [9] four Scales of Psychological Space: figural (smaller than the body), vista (potentially apprehended from one place – single rooms,

■ **Table 1**

Vista (43)	Environmental Scale (63)	Geographical Scale (25)
Treaty Oak	Museum of Modern Art	Pennsylvania
This classroom	Lake Michigan	The Great Basin
Hubble telescope	16801	Midwest
secret fort near my house growing up	Yahoo! Inc. Headquarters	Mesopotamia
The bathroom	Grand Central	United States
Craig O's Pastaria walk-in freezer	JFK International Airport	Mordor
Times Square	Boalsburg, PA	I-99
Intersection Allen St and College Ave.	Korean town in LA	Yugoslavia
Golden Gate Bridge	Manhattan	Africa
My hallway closet	Wall Street	The universe

town squares, small valleys), environmental (requiring locomotion to experience – buildings, neighborhoods, cities), and geographical (much larger than the body, understood through symbolic means). Among 140 entities listed collectively, five (arguably) are figural (e.g., the atom in my foot; my shoe). The table below provides 10 examples each for the other three categories (with totals). Those at vista scale include many personal places. Most environmental and geographical scale entities are named places experienced or known by many people. Geographical scale places were least frequent, suggesting that “place” is more easily associated with locations that can be experienced; it also included the only instances of fictional (Mordor) or historical (Mesopotamia) places. Overall, few linear features were named (e.g., 2 streets, 1 wildlife drive, 1 freeway, 1 interstate, and 1 river – the Nile).

4 Statements about places: expert classification

Understanding which entities count as places is a step toward recognizing statements "about" a place. Addressing the about component is closely related to GIR research on document relevance, (e.g., [1], [10]) and on document geographic focus (e.g., [2],[6]), but focuses on statements, not documents. In this section, we present results of a classification task carried out by the 6 graduate student co-authors. The objective was to explore factors leading to statements (in tweets) being conceptualized as “about a place” (or not), and to analyze differences in opinion among individuals who have studied the concept of place formally.

4.1 Procedure

For this task, 104 tweets were sampled from a large repository, with 8 tweets each from 13 subsets related to different event types (earthquake, ebola, fire, flood, flu, malaria, measles, protest, rebels, riot, tornado, violence, womensmarch). Each sample of 8 included 4 tweets containing a formal place name and 4 tweets without a formal place. Tweets with strong offensive language, unintelligible language, or primarily hashtags and/or URLs were omitted. The sampling goal was to select tweets (whether containing formal place names or not) that varied in likelihood of being considered to be about place. Tasks were presented via Google Forms with a form heading of Is this Tweet about a place? followed by, The goal of this task is to distinguish between tweets that are “about” places (thus that are “on the subject of; concerning” places) and those that are not. Tweets appeared to participants in random order, with two choices: “Yes, it is about a place” or “No, it is not about a place.”

4.2 Results and interpretation

Of the 104 tweets, 20 were judged unanimously to be about a place, with 24 more about a place by a majority (≥ 4 of 6). At the other extreme, 28 tweets were judged unanimously to be not about a place, with 25 more by a majority. Seven tweets resulted in a 3-3 tie.

At the extremes, there are clear characteristics that prompt unanimity in judgments about whether a statement is or is not “about a place.” For those judged as about a place, the statement is often about an event, focused on something local in geographical scale, and/or from the perspective of being on the ground. Linguistic cues in the form of locative prepositions also are common. Examples (with RT and @ references removed) include:

- ... about 20,000 people are here in Santa Ana for Orange County #womensmarch2018
- Apparently it’s testing day for the tornado sirens. Skerd me to death. They’re much louder at 101st and Sheridan!??

For statements judged consistently as not about a place, the most common feature is absence of reference to a geographic scale entity (thus without a name or description). This is the case even if an event probably occurring in a place is mentioned; examples include:

- Proud supporter of this & other groups trying to save this democracy.. #dontbackdown . #unitedwewin . #womensmarch2018
- ... and the government want to send arms for the rebels but not a democracy

That said, statements with place names are not always judged to be about a place; e.g., when a government is the intended meaning rather than the territory as well as when it is clear that the geographic entity mentioned is not the focus of the statement; one example is:

- It would cost \$1 billion a year to eradicate malaria which kills \$1 million people per year, the U.S. spends 10 billion ...

Minority views in near-unanimous “is a place” judgments (5-1) can result from too-quick reading (e.g., not noticing a place name due to abbreviation of unfamiliarity). Other factors leading to a minority view that a statement is not about a place are: statements naming more than one location, interpreting “about” strictly, or a geographic entity with indistinct boundaries. At the other extreme, a liberal definition of “about” (e.g., any mention of a proper place name counts) or considering virtual/social “places” to count (e.g., twitisphere) prompts judgments that a statement is about a place when most individuals feel it is not.

Statements with a 4-2 majority for place typically included a formal place name or abbreviation (e.g., “... about the ebola existing in jhb”) and/or use a preposition tied to an event or a proper noun (e.g., “... the 0749 from Radlett cancelled due to no driver...”). Lack of unanimity, however, is prompted by many factors: unfamiliar abbreviations (jhb for Johannesburg), symbolic interpretation of a name (e.g., White House), unclear connection of name to overall statement (e.g., for hashtags), mention of multiple places (thus not *a* place), context points to other focus (e.g., mentions China, but tweet is “about” measles), or too little context to distinguish place from object (e.g., “the fire hydrant outside my building”).

In contrast to the set above, some statements resulted in a 2-4 minority judging them to be about a place. Factors include: use of negation (Not Baghdad), unclear place reference (“Miss”, could be a person’s title or an abbreviation for the U.S. state), place entities mentioned as context for something else (“If TB Joshua want to heal the Ebola Victims Sierra Leone and Liberia isn’t far away let him take his crusade there pls!we”), place names standing for a person (the White House, as above) or a government (Russia protests ...), vague reference (e.g., the world), description of an event, but with no place name to locate it (e.g., “I want them to stop rioting now”), use of a place name without a corresponding event

6 Discussion

The research reported is a part of a larger effort focused on understanding characteristics of language related to place and creating computational methods to recognize statements (and documents) that are about places. While the initial research (focused on entities considered to be places and place attributes) was carried out in a semi-formal way as part of an ongoing course, results provide a starting point to explore the diverse characteristics that define place, including how place is related to geographic scale, personal experience, and function.

The second two parts of the research together provide insight on the challenges and possibilities for building computational methods to enable large volumes of text to be explored for place-related information. It is clear (from analysis of agreement and disagreement among a group of individuals studying place), that judging whether a statement is “about a place” depends on how “about” is interpreted as well as on the individual’s view of what constitutes a “place”. But, the small number of statements that resulted in a stalemate of conflicting judgments suggests that statements can be reliably categorized as being about a place (or not). The subsequent repeat of the experiment using crowdsourcing shows that reliable results are likely using this approach for all statements except those on which even experts disagree (situations with differences in what “about” means, abbreviated names, symbolic places, or imprecise/vague place references). Thus, we expect that it will be possible to build a large corpus of statements classified as being about place or not and to use them to train and test machine/deep learning methods to carry out this task with large volumes of text.

References

- 1 Paul D Clough, Hideo Joho, and Ross Purves. Judging the spatial relevance of documents for gir. In *European Conference on Information Retrieval*, pages 548–552. Springer, 2006.
- 2 Yong Gao, Dan Jiang, Xiang Zhong, and Jingyi Yu. A point-set-based footprint model and spatial ranking method for geographic information retrieval. *ISPRS International Journal of Geo-Information*, 5(7):122, 2016.
- 3 Reginald G Golledge. Place recognition and wayfinding: Making sense of space. *Geoforum*, 23(2):199–214, 1992.
- 4 Michael F Goodchild. Formalizing place in geographic information systems. In *Communities, neighborhoods, and health*, pages 21–33. Springer, 2011.
- 5 Jay T Johnson. Place-based learning and knowing: critical pedagogies grounded in indigeneity. *GeoJournal*, 77(6):829–836, 2012.
- 6 Michael D Lieberman, Hanan Samet, Jagan Sankaranarayanan, and Jon Sperling. Steward: architecture of a spatio-textual search engine. In *Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems*, page 25. ACM, 2007.
- 7 Alan M MacEachren. Leveraging big (geo) data with (geo) visual analytics: Place as the next frontier. In *Spatial Data Handling in Big Data Era*, pages 139–155. Springer, 2017.
- 8 Doreen Massey. Places and their pasts. In *History workshop journal*, volume 39, pages 182–192. JSTOR, 1995.
- 9 Daniel R Montello. Scale and multiple psychologies of space. In *European conference on spatial information theory*, pages 312–321. Springer, 1993.
- 10 Tumasch Reichenbacher, Stefano De Sabbata, Ross S Purves, and Sara I Fabrikant. Assessing geographic relevance for mobile search: A computational model and its validation via crowdsourcing. *Journal of the Association for Information Science and Technology*, 67(11):2620–2634, 2016.
- 11 Yi-Fu Tuan. *Space and place: The perspective of experience*. U of Minnesota Press, 1977.

Geospatial Semantics for Spatial Prediction

Marvin Mc Cutchan

Vienna University of Technology, Austria
marvin.mccutchan@geo.tuwien.ac.at

Ioannis Giannopoulos

Vienna University of Technology, Austria
igiannopoulos@geo.tuwien.ac.at

Abstract

In this paper the potential of geospatial semantics for spatial predictions is explored. Therefore data from the LinkedGeoData platform is used to predict landcover classes described by the CORINE dataset. Geo-objects obtained from LinkedGeoData are described by an OWL ontology, which is utilized for the purpose of spatial prediction within this paper. This prediction is based on an association analysis which computes the collocations between the landcover classes and the semantically described geo-objects. The paper provides an analysis of the learned association rules and finally concludes with a discussion on the promising potential of geospatial semantics for spatial predictions, as well as potentially fruitful future research within this domain.

2012 ACM Subject Classification Information systems → Association rules, Information systems → Geographic information systems, Software and its engineering → Semantics

Keywords and phrases Geospatial semantics, spatial prediction, machine learning, Linked Data

Digital Object Identifier 10.4230/LIPICs.GIScience.2018.45

Category Short Paper

1 Introduction and Related Research

This paper investigates the potential of geospatial semantics for spatial predictions. For this purpose data is obtained from the LinkedGeoData platform [6], which maintains geospatial data with semantic annotations, provided as linked data. This data is then used to predict CORINE landcover classes within a defined region of interest (ROI). Predictions are carried out by computing association rules using the FP-Growth algorithm. Descriptive statistics are calculated for the corresponding association rules and are used for an evaluation of the potential of geospatial semantics for spatial predictions. For the purpose of this research, spatial prediction is defined as the prediction of a CORINE landcover class for a defined region, based on the classes of geo-objects which fall within that region. Examples for such classes of geo-objects are given: tree, restaurant, river and bar. The proposed methodology ultimately enables to predict landcover classes in areas, where no classifications are yet available.

Spatial predictions are identified as one of the use-cases of Digital Earth [5] as well as a key feature for tackling global problems such as urbanization and climate change [3]. Association analysis can spatially predict and has traditionally been utilized as spatial association rule mining [8] or co-location mining [7] within the domain of Geoinformation science. Spatial association rule mining aims at detecting geo-objects with a frequent spatial relationship. Other papers focus on increasing the performance of co-location mining in terms of computational complexity [1, 10]. Further approaches use contextual information as an auxiliary data source in order to achieve better predictions [11, 14]. Nevertheless, no work



© Marvin Mc Cutchan and Ioannis Giannopoulos;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 45; pp. 45:1–45:6

Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

uses extensive semantic information for association analysis or researches on the contribution of geospatial semantic information for spatial predictions. Thus, this work explores the potential, geospatial semantics hold for spatial predictions. The contribution of this paper is as follows: It demonstrates that data with geospatial semantics enable to score meaningful association rules and are therefore a promising data source for this purpose. Additionally, it is shown that geospatial semantics predict association rules with a high conviction in urban areas as well as that a higher number of distinct classes provide better results.

The paper is structured as follows: Firstly the methodology is outlined, followed by a presentation of the results, including an analysis. The paper finalizes with a discussion of the results as well as its fruitfulness to future research and applications.

2 Methodology

Two major steps are performed within the methodology of this work. First, the data is derived and preprocessed. Second, the association analysis is carried out using the FP-Growth algorithm [2]. The FP-Growth algorithm generates association rules describing which set of classes have a relevant association.

2.1 Data acquisition and preparation

In order to perform an association analysis between linked data of the LinkedGeoData dataset and the CORINE landcover dataset (see table 1), a series of steps are performed for deriving and preparing the data: OpenStreetMap data is downloaded for a ROI. SPARQLIFY[6] is then used to load the OpenStreetMap data into a new local triplestore as linked data in the LinkedGeoData structure. Thus, a local copy of the LinkedGeoData endpoint is created for a specific ROI. This enables to access semantic information of geo-objects of the OpenStreetMap dataset, such as its OWL classes. The OWL classes are defined by the LinkedGeoData ontology and are ultimately used to predict the CORINE landcover classes. The ROI is covered by Austria. Geo-objects are then loaded into a PostGIS database. Each object contains three attributes: A unique identifier, the name of its OWL class as well as a geometry encoded as a well-known binary. Thus, every geometry is enhanced with semantics. The dataset contains 3 different types of geometries, namely, point, polygon and linestring. There are 1.080.819 point-objects, 4.024.536 polygon-objects and 1.893.309 line-object which can have one of the 768 OWL classes. Furthermore, the CORINE dataset is transformed to a polygon dataset where each pixel is presented by a square polygon, called a grid-cell. Each grid-cell contains two attributes: An unique identifier, as well as the class number of the corresponding landcover class. There are 44 classes in the CORINE landcover dataset, however, only 28 of them are present in Austria [13]. There are 56667188 grid-cells within the ROI. Finally, a transaction table is generated in the PostGIS database. Each row of this transaction table contains the identifier of a grid-cell and a class of a geo-object which intersects with the corresponding grid-cell. Thus, the table enables to query which classes appear within a certain grid-cell. The distinct set of classes which intersect with a grid-cell is defined as a transaction t_i . All transactions form a set, denoted as \mathbf{T} . Thus, t_i is a subset of \mathbf{T} .

2.2 Association Analysis

After the preparation of the data, the association analysis is performed. Therefore the FP-Growth algorithm is utilized which computes association rules based on the frequencies

of transactions and the number of all available transactions. An example of a computed association rule is given:

$$\{\mathbf{Building}, \mathbf{Tree}, \mathbf{Tramway}\} \rightarrow \{\mathbf{Continuous\ urban\ fabric}\}$$

This association rule suggests, that the class “Continuous urban fabric” is likely to appear, if the classes “Building”, “Tree” and “Tramway” are present. Association rules can have different confidences. The confidence of an association rule can be calculated by equation 1 [12].

$$conf(\mathbf{X} \rightarrow \mathbf{Y}) = \frac{supp(\mathbf{X} \cup \mathbf{Y})}{supp(\mathbf{X})} \quad (1)$$

$$supp(\mathbf{X}) = \frac{|\{t_i | \mathbf{X} \subseteq t_i, t_i \in \mathbf{T}\}|}{|\mathbf{T}|} \quad (2)$$

The support function $supp(\mathbf{X})$ describes the proportion of a transaction t_i , which contains \mathbf{X} , in the set \mathbf{T} . Its numerator denotes the number of times t_i (which contain \mathbf{X}) is observed among all transactions in \mathbf{T} . Whereas the denominator is defined by the number of all transactions within \mathbf{T} . The confidence can range from [0,1] and states how often a rule has been found in the transaction database. A confidence of 0 corresponds to no confidence that a given rule is true. In contrast, a confidence of 1 states the maximum confidence that an association rule is correct. An association rule can be additionally described by the Conviction [12]:

$$conv(\mathbf{X} \rightarrow \mathbf{Y}) = \frac{1 - supp(\mathbf{Y})}{1 - conf(\mathbf{X} \rightarrow \mathbf{Y})} \quad (3)$$

The conviction is as a measurement of the degree of implication of an association rule. An association rule can be confident merely because \mathbf{Y} appears with a high frequency and \mathbf{X} with a low frequency. A high conviction corresponds to a high degree of implication of an association rule, whereas a low conviction corresponds to a low degree of implication of a rule. Confidence and conviction are going to be used to validate the generated association rules. The FP-Growth algorithm computes rules based on a defined minimum support. The lower the support is set, the more association rules are computed. However, defining the value too low will yield a long runtime. The support is set as low as possible within this study to compute as much association rules as possible in order to gain more insights on the impact of geospatial semantics on spatial predictions. For this purpose, an optimal value of 0.1 was found in an interative manner. In addition, association rules having a conviction lower than 1 were pruned, as a conviction below that value suggests no significant implication. For running the FP-growth algorithm, rapidminer [9] was used. There are 56.667.188 grid-cells and consequently 56.667.188 potential transactions. Due to computational limitations not all of these transactions were used within this study. Therefore 5000 randomly selected transactions per landcover class were chosen. This balanced selection was made in order to avoid a bias in the association rules.

3 Results and Analysis

There are 28 predictable CORINE landcover classes within the ROI (see table 1). Each landcover class is a subclass of a more general parentclass. Tables 2, 3 and 4 summarize the number of learned association rules, the descriptive statistics for each class as well as the corresponding parent class, according to the definition of the CORINE dataset [13].

■ **Table 1** All available 28 CORINE classes in Austria and their description.

1	2	3	4	5	6	7	10
Continuous urban fabric	Discontinuous urban fabric	Industrial and commercial units	Road and rail network and associated land	Port areas	Airports	Mineral extraction sites	Green urban areas
11	12	14	15	18	20	21	23
Sport and leisure facilities	Non-irrigated arable land	Rice fields	Vine yards	Pastures	Complex cultivation patterns	Agriculture with natural vegetation	Broad leaved forest
24	25	26	27	29	31	32	34
Coniferous forest	Mixed forest	Natural grassland	Moors and heathland	Transitional woodland shrubs	Bare rock	Sparsely vegetated area	Glaciers and perpetual snow
35	36	40	41				
Inland marshes	Peatbogs	Water courses	Waterbodies				

■ **Table 2** Predictions for CORINE landcover classes 1-11.

CLASS number	1	2	3	4	5	6	7	10	11
Number of association rules	5752	157	345	226	-	1525	45	496	11
Min (Confidence)	0.11	0.11	0.11	0.11	-	0.11	0.11	0.11	0.11
Max (Confidence)	0.87	0.36	0.58	0.72	-	0.98	0.90	0.71	0.86
Mean (Confidence)	0.32	0.17	0.23	0.25	-	0.60	0.51	0.24	0.52
Standarddeviation (Confidence)	0.17	0.05	0.12	0.14	-	0.25	0.28	0.12	0.28
Min (Conviction)	1.08	1.09	1.08	1.08	-	1.10	1.08	1.08	1.08
Max (Conviction)	7.23	1.50	2.27	3.48	-	61.47	9.23	3.40	7.06
Mean (Conviction)	1.59	1.17	1.31	1.37	-	6.17	3.25	1.34	2.87
Standarddeviation (Conviction)	0.64	0.07	0.27	0.41	-	8.58	2.50	0.34	1.89
CORINE Parentclass	Artificial surfaces								

Observing tables 2, 3 and 4, several trends can be observed: Most rules were computed for class 1(Continuous urban fabric), followed by class 6 (Airports). Association rules predicting class 1 exhibit a relatively high confidence, up to 87%, as well as a relatively high conviction, 7.23. An association rule which exhibits both, a high conviction and high confidence can be considered a meaningful rule. Generally, it can be observed that all subclasses of “Artificial surfaces” yield the most promising results. In contrast, confidence and conviction decline for classes which are a subclass of “Agricultural areas”, with one exception, i.e. class 15 (vine yards), which was predicted with exceptional conviction and confidence. However, no predictions could be made for class 14 (rice fields). The lowest confidence as well as conviction can be observed among subclasses of “Forests and semi-natural areas” and “Wetlands” with one exception, class 34 (Glaciers and perceptual snow). No association rules were computed for classes 27 (Sclerophyllous vegetation), class 29 (Transitional woodland shrub), class 32 (Sparsely vegetated areas), as well as class 36 (Peatbogs). The confidence as well as conviction inclines for subclasses of water bodies, as specially for subclass 41 (water bodies).

4 Discussion and Future Research

Considering the findings based on the results presented in tables 2, 3 and 4 it can be said that geospatial semantics can be used for spatial predictions and exhibit different qualities depending on the landcover class to be forecast. Classes closely related to urban areas are predicted better than classes which can be found more often in rural areas, such as forests or wetlands. A potential explanation for this effect is given: LinkedGeoData is based on

■ **Table 3** Predictions for CORINE landcover classes 12-25.

CLASS number	12	14	15	18	20	21	23	24	25
Number of association rules	5	-	64	2	4	1	-	-	-
Min (Confidence)	0.12	-	0.11	0.12	0.12	0.15	-	-	-
Max (Confidence)	0.13	-	0.86	0.13	0.17	0.15	-	-	-
Mean (Confidence)	0.13	-	0.48	0.12	0.14	0.15	-	-	-
Standarddeviation (Confidence)	0.01	-	0.25	0.01	0.02	-	-	-	-
Min (Conviction)	1.09	-	1.10	1.09	1.09	1.13	-	-	-
Max (Conviction)	1.11	-	6.73	1.10	1.16	1.13	-	-	-
Mean (Conviction)	1.10	-	2.59	1.10	1.12	1.13	-	-	-
Standarddeviation (Conviction)	0.01	-	1.68	0.01	0.03	-	-	-	-
CORINE Parentclass	Agricultural areas						Forests and semi-natural areas		

■ **Table 4** Predictions for CORINE landcover classes 26-41.

CLASS number	26	27	29	31	32	34	35	36	40	41
Number of association rules	1	-	-	2	-	11	8	-	47	12
Min (Confidence)	0.15	-	-	0.11	-	0.15	0.11	-	0.11	0.11
Max (Confidence)	0.15	-	-	0.27	-	0.86	0.25	-	0.27	0.52
Mean (Confidence)	0.15	-	-	0.19	-	0.44	0.16	-	0.15	0.22
Standarddeviation (Confidence)	-	-	-	0.12	-	0.24	0.05	-	0.04	0.12
Min (Conviction)	1.13	-	-	1.08	-	1.13	1.08	-	1.08	1.08
Max (Conviction)	1.13	-	-	1.32	-	7.11	1.29	-	1.31	1.99
Mean (Conviction)	1.13	-	-	1.20	-	2.49	1.15	-	1.14	1.28
Standarddeviation (Conviction)	-	-	-	0.17	-	2.13	0.07	-	0.05	0.25
CORINE Parentclass	Forests and semi-natural areas						Wetlands		Waterbodies	

OpenStreetMap and therefore relies on volunteers collecting geospatial data. Thus, there is a greater likelihood that a higher coverage of geospatial data is present in urban areas, increasing the number of classes per grid-cell. A higher number of classes per grid-cell enable to compute association rules with a higher conviction as they exhibit a higher distinction and therefore result in a lower support. The same argument could be made for the number of available classes: A higher amount of available classes would increase the chances to get association rules with a higher conviction as it would increase the distinction. However, this aspect is not covered in this study. Future research will focus deeper on the investigation of the potential of geospatial semantics for predictive purposes. Therefore two major aspects will be investigated: (1) The effect of the class hierarchy on the quality of spatial predictions. For this purpose classes will be exchanged with their parent class. (2) Future studies will investigate the impact of adding other geospatial data with different ontologies. The obtained knowledge can be used as input in spatial human-computer interaction [4], for future geosensor networks in order to create better predictions as well as to measure the impact on integrating different geospatial data sources with semantic annotations. This could help to explain yet undiscovered geospatial phenomena and it is therefore argued that further analysis in this domain is paramount to research progress.

References

- 1 Witold Andrzejewski and Pawel Boinski. GPU-accelerated collocation pattern discovery. In Barbara Catania, Giovanna Guerrini, and Jaroslav Pokorný, editors, *Advances in Databases and Information Systems*, pages 302–315, Berlin, 2013. Springer.
- 2 Christian Borgelt. An implementation of the fp-growth algorithm. In *Proceedings of the 1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations*, OSDM '05, pages 1–5, New York, NY, USA, 2005. ACM. doi:10.1145/1133905.1133907.
- 3 Max Craglia, Kees de Bie, Davina Jackson, Martino Pesaresi, Gábor Remetej-Fülöpp, Changlin Wang, Alessandro Annoni, Ling Bian, Fred Campbell, Manfred Ehlers, John van Genderen, Michael Goodchild, Huadong Guo, Anthony Lewis, Richard Simpson, Andrew Skidmore, and Peter Woodgate. Digital earth 2020: towards the vision for the next decade. *International Journal of Digital Earth*, 5(1):4–21, 2012. doi:10.1080/17538947.2011.638500.
- 4 Ioannis Giannopoulos, Peter Kiefer, and Martin Raubal. Mobile outdoor gaze-based geo-HCI. In *Geographic Human-Computer Interaction, Workshop at CHI 2013*, pages 12–13, 2013.
- 5 M. F. Goodchild. The use cases of digital earth. *International Journal of Digital Earth*, 1(1):31–42, 2008. doi:10.1080/17538940701782528.
- 6 Jon Jay Le Grange, Jens Lehmann, Spiros Athanasiou, Alejandra Garcia Rojas, Giorgos Giannopoulos, Daniel Hladky, Robert Isele, Axel Cyrille Ngonga Ngomo, Mohamed Ahmed Sherif, Claus Stadler, and Matthias Wauer. The geoknow generator: Managing geospatial data in the linked data web. http://jens-lehmann.org/files/2014/lgd_geoknow_generator.pdf. (Accessed on 04/30/2018).
- 7 Y. Huang, S. Shekhar, and H. Xiong. Discovering collocation patterns from spatial data sets: a general approach. *IEEE Transactions on Knowledge and Data Engineering*, 16(12):1472–1485, 2004. doi:10.1109/TKDE.2004.90.
- 8 Krzysztof Koperski and Jiawei Han. Discovery of spatial association rules in geographic information databases. In Max J. Egenhofer and John R. Herring, editors, *Advances in Spatial Databases*, pages 47–66, Berlin, 1995. Springer.
- 9 Rapidminer. Lightning fast data science platform | rapidminer. <https://rapidminer.com/>. (Accessed on 04/30/2018).
- 10 Arpan Man Sainju and Zhe Jiang. Grid-based collocation mining algorithms on gpu for big spatial event data: A summary of results. In Michael Gertz, Matthias Renz, Xiaofang Zhou, Erik Hoel, Wei-Shinn Ku, Agnes Voisard, Chengyang Zhang, Haiquan Chen, Liang Tang, Yan Huang, Chang-Tien Lu, and Siva Ravada, editors, *Advances in Spatial and Temporal Databases*, pages 263–280, Cham, 2017. Springer International Publishing.
- 11 Muhammad Shaheen, Muhammad Shahbaz, and Aziz Guergachi. Context based positive and negative spatio-temporal association rule mining. *Knowledge-Based Systems*, 37:261–273, 2013. doi:10.1016/j.knosys.2012.08.010.
- 12 Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Addison-Wesley Longman Publishing, Boston, MA, USA, 2005.
- 13 UBA. CORINE Landcover Nomenklatur (deutsch). http://www.umweltbundesamt.at/fileadmin/site/umwelthemen/raumplanung/1_flaechennutzung/corine/CORINE_Nomenklatur.pdf. (Accessed on 04/30/2018).
- 14 Cunjin Xue, Wanjiao Song, Lijuan Qin, Qing Dong, and Xiaoyang Wen. A spatiotemporal mining framework for abnormal association patterns in marine environments with a time series of remote sensing images. *International Journal of Applied Earth Observation and Geoinformation*, 38:105–114, 2015. doi:10.1016/j.jag.2014.12.009.

Docked vs. Dockless Bike-sharing: Contrasting Spatiotemporal Patterns

Grant McKenzie

Department of Geography, McGill University, Montréal, Canada

Abstract

U.S. urban centers are currently experiencing explosive growth in commercial dockless bike-sharing services. Tens of thousands of bikes have shown up across the country in recent months providing limited time for municipal governments to set regulations or assess their impact on government-funded dock-based bike-sharing programs. Washington, D.C. offers an unprecedented opportunity to examine the activity patterns of both docked and dockless bike-sharing services given the history of bike-sharing in the city and the recent availability of dockless bike data. This work presents an exploratory step in understanding how dockless bike-sharing services are being used within a city and the ways in which the activity patterns differ from traditional dock station-based programs.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases bike-share, dockless, bicycle, transportation, spatiotemporal patterns

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.46

Category Short Paper

1 Introduction

Cities in the United States are in the midst of a bike-share revolution of sorts [8]. Seemingly overnight, GPS-enabled bicycles have popped up in urban centers from Seattle to Miami, offering access to inexpensive, mobile-payment-based, one-way rentals. Users simply unlock a bike with their mobile device, cycle to their destination, park and lock it on any public land, and walk away. These new dockless bike-share services sell themselves as low cost alternatives to traditional dock-based bike-sharing programs, allowing users the freedom to park a bike virtually anywhere in contrast to the traditional model of designated docking stations.

There is no shortage of companies entering the U.S. dockless bike-sharing space. While dockless programs are quite common in much of Asia and Europe, the U.S. has recently seen substantial investment from companies such as Mobike, Spin, Jump, and LimeBike (Figure 1a). In October of 2017, just as it entered the Washington, D.C. market, *LimeBike*¹ (LB), reported 300,000 unique users and \$225 million in funding [2]. Similar to other dockless services, LimeBike offers 30 minute rentals for \$1 USD and operates on any public space within the metro Washington D.C. area.

Bike-sharing in general is not new to the U.S. and one of the oldest bike-share programs in the country, *Capital Bikeshare* (CB), currently serves the greater metro D.C. area. Originally started under the name SmartBike DC in 2008, it boasts an annual ridership of over 2.1 million² and costs either \$2 USD per 30 min rental or access through membership subscription. CB is a dock-based bike-sharing service where users lock and unlock bicycles from docking stations distributed around the city (Figure 1b). Importantly, and in contrast to the dockless

¹ <http://www.limebike.com/>

² <https://www.capitalbikeshare.com>



© Grant McKenzie;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 46; pp. 46:1–46:7

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



(a) LimeBike (Dockless).

(b) Capital Bikeshare (Docked).

■ **Figure 1** Docked vs. dockless bike-sharing platforms. Photographs: Wikimedia / CC License.

bike-share companies mentioned previously, Capital Bikeshare is owned by the municipal governments it serves (i.e., D.C., Virginia, and Maryland).

There have been numerous studies aimed at the social impact [5] and mobility patterns [10, 9] of bike-sharing programs as well as method for intelligently redistributing bikes throughout urban centers [6]. However, very little research has compared traditional dock-based models to new dockless systems. Given the dramatic influx of dockless bike-sharing companies in the U.S. over the last six months [1], this study is one of the first to compare and contrast the spatial and temporal usage patterns in a city that supports both. One important factor contributing to the novelty of this work is that the public is just now gaining access to much of these data.³ The vast majority of these new dockless bike-sharing companies do not share data related to the locations of their fleet. As far as I am aware, Washington, D.C.'s Department of Transportation (DDOT) is the only U.S. city requiring these companies to provide a publicly accessible application programming interface (API) showing the current location of any dockless bicycles available for rent.⁴

This short paper takes a first step in better understanding the differences in activity patterns between docked and dockless bike-sharing programs. The insight gained through this exploratory research can be used to better inform urban planners, transportation engineers, and the general public on how cyclists and citizens interact with their city.

2 Data

Capital Bikeshare trips for the month of March, 2018 were accessed for this work,⁵ a total 238,936 individual trips. Attribute information for these trips include bike ID, time stamps for the start and end of the trip (to the nearest second), and start and end station IDs. Station IDs were matched with point locations through data available from DC.gov's open data portal. All stations outside of D.C., namely those in Maryland and Virginia, were removed thus restricting trips to only those within the district. This reduced the number of accessible stations from 499 to 269 and number of trips to 209,973. To permit comparison between the two bike-sharing services, the CB data was rounded to the nearest five minute interval.

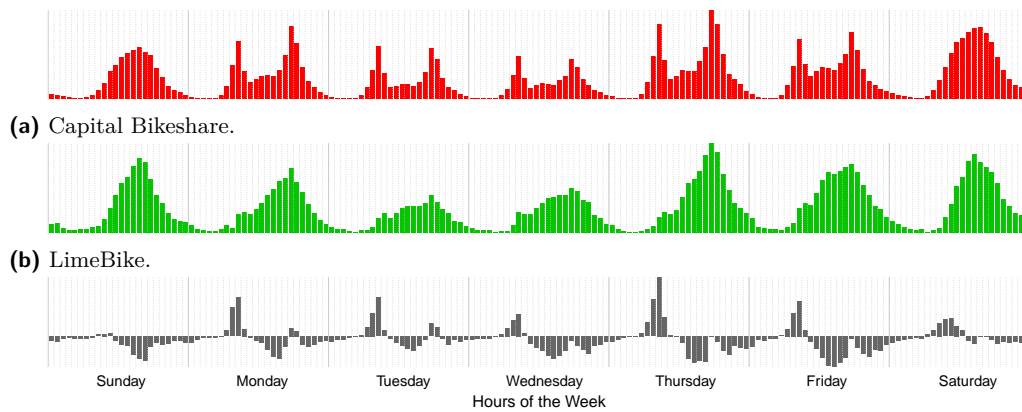
LimeBike data were accessed through their API⁶ every five minutes from March 10th through March 31, 2018. These data include the bike ID, geographic coordinates (to roughly the nearest meter), and time stamp of the available bicycle (at a five minute temporal

³ LimeBike's D.C. API was made public on February 6, 2018.

⁴ <https://github.com/ubahnverleih/WoBike/issues/9#issuecomment-355047664>

⁵ Data are available at <https://www.capitalbikeshare.com/system-data>

⁶ <https://lime.bike/api/partners/v1/bikes>



(c) LimeBike subtracted from Capital Bikeshare temporal patterns. Y-axis ranges from 0.6 to -0.2.

■ **Figure 2** Temporal signatures for Capital Bikeshare and LimeBike in Washington, D.C. aggregated to hours of a standard week.

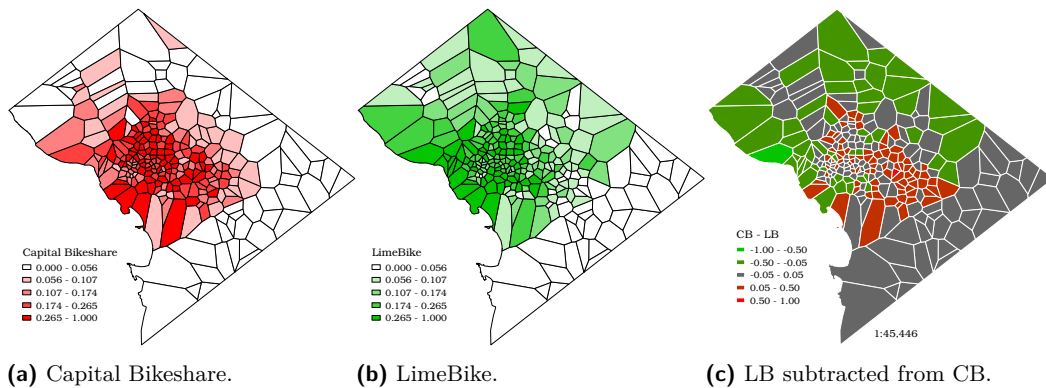
resolution). Further steps were necessary to convert the LB bike *availability data* into trips. The data snapshots captured every five minutes were sorted in order and a trip was recorded as the last time stamp that a bike ID was marked as available, to the next time stamp that the same bike ID reappeared in the data. Assuming GPS accuracy errors within an urban setting, only those bike IDs that moved more than 50 meters were recorded as trips. In total, 154,024 trips were taken by LimeBike users over this time period within the district boundary.

3 Temporal Differences

The mean duration of a trip for both services was approximately 18 minutes though CB showed a median duration of 11 minutes while LB reported 5 minutes (the temporal resolution of data collection). The tendency towards longer trips by CB users is significant and may be partially due to the necessity of finding a docking station instead of leaving the bike in any public space.

The temporal popularities of the two bike-sharing services are shown in Figure 2. This shows bike trip start times aggregated to the nearest hour of a week and independently normalized to account for the larger number of CB trips. We see an expected diurnal pattern with the majority of trips taking place during daylight hours for both services. One difference is the weekday morning peak in Figure 2a, notably missing from Figure 2b. Figure 2c shows the LB temporal pattern subtracted from the CB pattern. The weekday morning peak in CB activity is more apparent here and most pronounced at 8 a.m. This also shows that LB is more popular in the early and late afternoons. Note, however, that there is a negligible difference between the two bike-share services at 5 p.m. on weekdays, peak of the evening commute.

An assessment, based purely on temporal patterns within these data, suggests that the docked CB is used more for commuting to and from work than the dockless LB. Contrarily, CB is used more frequently outside of commuting hours and particularly in the mid-afternoon.



■ **Figure 3** Normalized trip starts assigned to Capital Bikeshare station-based Voronoi polygons.

4 Spatial Patterns

While the temporal activity patterns of bike-sharing services is one dimension on which to assess their similarities and differences, spatial activity patterns offer a different perspective.

By definition, docked or dockless bike-sharing systems consist of fundamentally different architecture. These differences make it difficult to compare them spatially. While dockless bike locations are scattered throughout the city (where ever someone chooses to stop), CB bike trips are restricted to starting and ending at docking station. To compare these two datasets, a Voronoi tessellation was used to partition Washington, D.C. into polygons based on the locations of CB docking stations. In theory, each of these polygons represents the region to which a CB user was traveling based on their chosen docking station. Admittedly there are limitations to this approach (e.g., water body restrictions), but this analysis was deemed suitable for this short paper.

The number of CB trips starting from each station were summed across the dataset and matched to the appropriate Voronoi polygon. LB trip starting points were also intersected with these same polygons and summed. The total trip count for each bike-sharing service in each polygon was then normalized for each service independently. This was done to account for the larger number of CB trips thus allowing for comparison between the two services. Figures 3a and 3b show the spatial distribution of trip starting points for CB and LB respectively. Figure 3c demonstrates the difference between the two services as the LB value for each Voronoi polygon subtracted from the CB value.

These maps demonstrate that CB ridership is more focused on the central business district, of Washington, D.C. than LB. Intersecting these polygons with land use data from D.C.'s Office of Planning, we find the ratio of commercial, industrial, or mixed use buildings to residential housing is nearly double for CB (0.35) compared to LB (0.17). This supports the temporal pattern analysis that suggests that CB is used more frequently for commuting than LB.

These maps indicate that the southeastern portion of the district is less likely to use any bike-sharing service than anywhere else in the district. One possible explanation is that the 2015 American Community Survey reported these predominantly residential regions, namely Wards 7 and 8, as having both the lowest household income in the district and largest number of individuals below the federal poverty line. While relatively inexpensive, both of these bike-sharing services rely on credit cards as the basis for payment, making it less likely that lower income individuals can use these services.

From a combined *spatiotemporal* perspective, the largest trip volume difference between services, across Voronoi polygons is weekdays between 3 p.m. and 5 p.m. whereas the smallest overall difference is weekdays between 2 a.m. and 4 a.m. LB shows the largest temporal variance in trip volume to the west of the downtown core, near the Georgetown neighborhood with peak usage on Fridays at 2 p.m. In contrast, CB peaks at 5 p.m., Monday through Thursday in the downtown commercial region of the district.

4.1 Data-driven Dock Locations

Access to dockless bike-share data offers an opportunity for a docked bike-share company such as CB. Given that dockless bikes can be left virtually anywhere, we can assume that bikes are most often parked at the most convenient locations for their users. This information can be used to assess the optimality of current docking station locations.

K-means [7] was used to cluster the dockless LB locations with a value of 269, the current number of DB stations in D.C., set as the number of clusters. Provided the weighted centers of these new clusters, the average distance between each cluster center and its nearest existing CB station was computed. This resulted in a mean distance of 305.4 m with a median of 181.4 m). This clustering approach ignores buildings and roads, however, and since many DB docking stations are located near intersections, these new cluster centers were snapped to the nearest road intersection and the average distance to existing stations was calculated again. The snapping had a minimal impact reducing the mean distance to 300.1 m and median to 180.2 m. This median distance indicates that the existing DB docking stations, on average, are reasonably well situated throughout Washington, D.C., at least as reflected by LB users. This approach demonstrates that having access to dockless bike-share data can have a substantial impact on infrastructure planning, potentially saving a city considerable effort and financial investment [4].

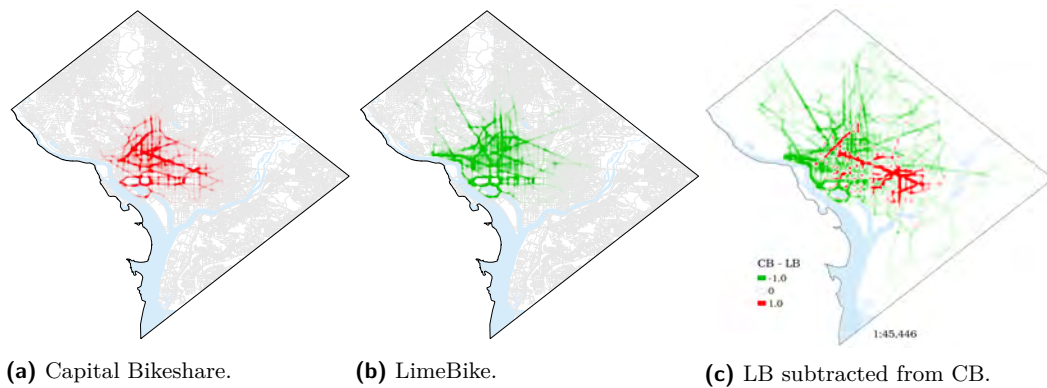
4.2 Network Patterns

The previous section's comparison of trip starting points⁷ is useful for understanding the different spatial distributions of bike-sharing services in D.C. An alternative approach is to examine the spatial distributions of trips on the D.C. road network. The shortest path was calculated between each start and end location along the D.C. road network using Dijkstra's algorithm [3] as implemented in pgRouting.⁸ Once routing analysis was complete, point geometries were generated every 10 m along each of the resulting line geometries. Using these points, kernel density estimates were created for CB and LB independently, producing the two *heat* maps shown in Figures 4a and 4b. Normalizing the kernel density values, LB was subtracted from CB to produce Figure 4c.

These results depict a similar, but more nuanced pattern than the Voronoi polygons shown in Figure 3. While the results are based on a simple shortest-path approach to determining trip routes, the analysis does offer important insight into how urban cyclist interact with their city. LB shows relatively more activity outside the city core, towards historic Georgetown, around the Tidal Basin, and along the Potomac River, areas that cater to leisure activity and are not typically considered areas of business. CB, by comparison, dominates commercial centers, along Massachusetts Avenue, Union train station, and the streets around Capital Hill. These results again support the notion that Capital Bikeshare tailors more to commuters than dockless services such as LimeBike.

⁷ Start points and end points since every end point is a start point for the next trip.

⁸ <http://pgrouting.org/>



■ **Figure 4** Start and end points of bike-share trips mapped to shortest path on the D.C. road network displayed as kernel density maps.

5 Conclusions & Next Steps

This work investigates the spatial and temporal dimensions of docked and dockless bike-share services in Washington, D.C. Though much of this analysis is exploratory, the findings suggest that there are clear difference in how these two services are used. Capital Bikeshare tends to be more commuter focused whereas LimeBike reflects more leisure or non-commute related activities. The results of these analyzes have important implications for urban planners, transportation safety boards, and transportation engineers as these findings may influence infrastructure budgeting, maintenance planning, and new development opportunities.

The results presented in this paper are preliminary since access to this spatial and temporal resolution of commercial bike-share data in the U.S. is still new and the recent influx of bike-share services in cities is disrupting the status quo. Analyzing more data over a longer time period will provide additional insight. Future work will examine the impact of new modes of dockless transportation (e.g., electric scooters), compare these patterns to light-rail ridership, and further examine the behavioral motivations for selecting one service over another.

References

- 1 Eliot Brown. Dockless bike share floods into U.S. cities, with rides and clutter. *The Wall Street Journal*, March 2018. Online; posted 26-03-2018.
- 2 Biz Carson. Limebike now valued at \$225 million after investors go all in on bike-sharing craze. *Forbes*, October 2017. Online; posted 16-11-2017.
- 3 Edsger W Dijkstra. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271, 1959.
- 4 Leif Dormsjo. District of Columbia, Capital Bikeshare Development Plan. Technical report, Government of the District of Columbia, 09 2015.
- 5 Elliot Fishman, Simon Washington, and Narelle Haworth. Bike share: a synthesis of the literature. *Transport reviews*, 33(2):148–165, 2013.
- 6 Zhaoyang Liu, Yanyan Shen, and Yanmin Zhu. Inferring dockless shared bike distribution in new cities. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 378–386. ACM, 2018.
- 7 James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

- 8 Felix Salmon. Bring on the bikocalypse. *Wired*, February 2018. Online; posted 01-02-2018.
- 9 Yu Shen, Xiaohu Zhang, and Jinhua Zhao. Understanding the usage of dockless bike sharing in singapore. *International Journal of Sustainable Transportation*, pages 1–15, 2018.
- 10 Jon Wergin and Ralph Buehler. Where do bikeshare bikes actually go? Analysis of capital bikeshare trips with GPS data. *Transportation Research Record: Journal of the Transportation Research Board*, 2662:12–21, 2017.

OpenPOI: An Open Place of Interest Platform

Grant McKenzie

Department of Geography, McGill University, Montréal, Canada

Krzysztof Janowicz

STKO Lab, University of California, Santa Barbara, USA

Abstract

Places of Interest (POI) are a principal component of how human behavior is captured in today's geographic information. Increasingly, access to POI datasets are being restricted – even silo-ed – for commercial use, with vendors often impeding access to the very users that contribute the data. Open mapping platforms such as OpenStreetMap (OSM) offer access to a plethora of geospatial data though they can be limited in the attribute resolution or range of information associated with the data. Nuanced descriptive information associated with POI, e.g., ambience, are not captured by such platforms. Furthermore, interactions with a POI, such as checking in, or recommending a menu item, are inherently place-based concepts. Many of these interactions occur with high temporal volatility that involves frequent interaction with a platform, arguably inappropriate for the “changeset” model adopted by OSM and related datasets. In this short paper we propose OpenPOI, an open platform for storing, serving, and interacting with places of interests and the activities they afford.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases place, point of interest, open data, gazetteer, check-in

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.47

Category Short Paper

1 Motivation

Gazetteers play an important role in how we understand the world. They facilitate the labeling of geographic space thus forming the foundation of location-based services [3]. Historically, gazetteers have been categorized by scale, resolution, and theme. Some of the more traditional gazetteers are global in scale but at fairly coarse resolutions focusing largely on geographic features on the macro and meso levels such as airports, populated places, and rivers. Local gazetteers have tended to focus on a specific theme at higher resolution within a limited geographic boundary, e.g., Difangzhi local Chinese histories [4]. With advances in technology and commercial investment, digital gazetteers have quietly taken on new roles, forming the foundation on which a lot of the technology we use today, is built. Anyone who has used a mobile device in the past ten years has benefited from digital gazetteers be it through navigation/wayfinding using Google Maps or photograph tagging on Instagram.

In the last several years, context-based technology has continued to drive commercial investment, as many information technology companies realize the value of location information. This has led to substantial investments in digital mapping technology [2] as well as the underlying spatial data that drives these platforms [7]. Digital gazetteers are increasingly storing the locations, names, and categories of *local* businesses and venues, today generally referred to as *points of interest* (POI). For instance, the location, name and hours of operation of the *mom-and-pop* shop at the end of your street is now stored alongside millions of other *place* records in a global gazetteer that forms the basis of Silicon Valley's latest mapping



© Grant McKenzie and Krzysztof Janowicz;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 47; pp. 47:1–47:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

products. Not only are these companies capturing information related to the location and hours of operation, but they are also enlisting citizens to contribute data on everything from menu recommendations to general ambience. Pushing this a step further, the technological scope of many of these commercial entities means that they can determine *popular times* for many of these places based on users' mobile device-reported locations [11]. The amount of auxiliary, or descriptive, data stored about these points of interest arguably contributes to a variation of the POI acronym, namely *Place* of interest. The content contributed about POI are much more than geometric points and really serve to give users of these data an understanding of fine-grained characteristics of a place.

Unfortunately, the high financial cost of developing these place of interest dataset has led to much of these data being siloed within companies, solely being used within (or sold as) their services or products. As is the case with many data silos, the redundancy between gazetteers is high. Companies such as Yelp, Foursquare, and Facebook have all invested heavily in data collection and development of their own proprietary POI datasets, rarely with collaborative interests. This has resulted in multiple academic efforts to match and conflate these datasets [6, 10] and occasional legal and regulatory action [9]. Though many companies offer limited access to their POI datasets for third party application development, recent high-profile events related to data privacy have resulted in tighter restrictions on outside access [1].

One concern related to the construction of these POI datasets is the reliance on volunteered contributions. Most of the data stored in these proprietary data silos were contributed by individuals not employed by the companies. Users are actively choosing to share, or are coerced [5] into sharing, often personal information with these platforms which are in turn monetized and sold back to those same users. While there is an argument to be made for the value added by these companies through their services and platforms, the reality is that most users no longer have legal rights, or even digital access, to the data that they've contributed to these platforms. These silos also hurt the research community as they limit access to attribute data needed for modern recommender systems and work on geographic information retrieval more broadly. Hence, we see this paper and the data and services it introduces as a *research enabler* for the community. Such dataset papers play an important role in geoinformatics research and are gaining importance in many other communities [8].

2 OpenPOI

Considering this, we introduce the *OpenPOI platform*, a dataset and service for storing, sharing, and interacting with a common set of places of interest. Following the open and user-contributed, geo-data model proposed by OpenStreetMap (OSM) and others, OpenPOI aims at supplying highly descriptive content related to local places of interest and enable a high level of interaction and sharing. Both of these approaches sit outside the mission of the OSM community and are likely not suitable for the *changeset* model and validation approach adopted by OSM. Through the OpenPOI platform, users can share recommendations, opinions, and place-specific information as well as *check-in*,¹ post photographs, or access any form of information they would like concerning a place. The purpose is to enable free and open access to *patial* information that is owned and shared by the community. This should be appealing to place-based social media users as the project is completely transparent, allowing everyone open access to all data and code associated with the platform. For researchers, it

¹ Check-in, in this case, refers to the social act of publishing one's presence at a location.

offers a valuable resource on which to study human activity behavior and a place of interest dataset that can be used as the basis for any application, study, or research project.

As this platform is in a prototype phase, we give a very brief overview of the components, the current state of the platform, and some directions for moving forward. Currently, there are three components to the OpenPOI project: The dataset, the web service, and the mobile application.

3 Dataset

OpenStreetMap nodes are the source for all POI in this current version of the platform. As OSM is user-contributed and regularly updated, it provides the most extensive coverage of non-proprietary POI in the world. As the OpenPOI user-base grows, new POI may be added and existing POI updated or removed, branching the dataset from the OSM community while still maintaining links through original OSM node IDs. Future versions of the platform will ingest changes from OSM and publish changes, with basic attributes, back to OSM, following community best-practices and appropriate validation. In this prototype version of the platform, country specific OSM PBF files were downloaded for the United States and Canada. After thoroughly testing the platform using these data, global OSM planet files will be used. The *OsmPoiPbf* POI extraction script² was used to extract point of interest nodes from the raw OSM files resulting in a series of CSV files that were automatically inserted into a PostGIS-enabled PostgreSQL database. Once in the database, a duplication check was made before building a spatial index on the point geometries.

This PostGIS-enabled *PostgreSQL* relational database is used to store all point geometries in the OpenPOI dataset. Data related to users, check-ins, and tags are all stored in *MongoDB*, a document-oriented database system, often classified as *NoSQL*. The primary reason for the two different storage formats is to keep the POI geometry data spatially indexed and separate from the application-level data. The rate at which POI geometries are changed is far less than that of descriptive content, tags, and check-ins. As MongoDB was developed with consistency across database replicas in mind, it serves this purpose well. The current version of the database stores user profile information, time stamps and locations of check-ins as well as collections of tags and attributes assigned to a specified POI. Data extracts are available for each of these data collections or access to the data is available via the OpenPOI application programming interface (API).

4 Web Service

The current version of the OpenPOI API allows for basic interaction with the underlying OpenPOI dataset. Again, the API, including source code, is freely available and accessible via creative commons license. The API forms the basis for the OpenPOI mobile application and forthcoming data extraction tool. *NodeJS*³ in combination with the *Express*⁴ framework supply the foundation for back-end development. A set of public API endpoints are now available and currently permit the following data requests:

- Provided a latitude and longitude, return an array of nearby POI objects.
- Provided a User ID or POI ID, return an array of check-in objects.

² <https://github.com/MorbZ/OsmPoisPbf>

³ <https://nodejs.org/>

⁴ <https://expressjs.com/>

- Provided a POI ID, return an array of tag objects for the specified POI.

In addition to requesting data related to POI, users can also interact with the data through submission of various types of content, namely,

- Add a check-in object to a POI given the POI ID and User ID.
- Add an array of tags (hash-pairs) to a POI given the POI ID and User ID.
- Create a new user object.

These endpoints form the core of the OpenPOI platform functionality with additional endpoints being added as development continues. Documentation including example requests and required parameters is available at <https://github.com/ptal-io/OpenPOI-Server>. The current version of the API does not require authentication, nor does it limit requests, though authentication will be required in future versions of the platform.

5 Mobile Application

The OpenPOI mobile application sits as a front-end interface through which the OpenPOI dataset is accessed and updated. The mobile application communicates through the aforementioned public API endpoints keeping the entire project modular. Anyone can build a mobile, desktop, or web-based front-end that interacts with the data through this API. The mobile application presented here is one possible interactive window into the dataset.

The OpenPOI mobile application⁵ is currently in development using the *React-Native* framework.⁶ React-Native allows developers to use the JavaScript scripting language in combination with the React library to create mobile apps that are compiled into platform-specific applications. The current release of the mobile application has been compiled for use on an Android mobile device and can be downloaded for testing at <http://openpoi.org>. An iOS version of the OpenPOI application is forthcoming. The prototype version of the application is limited in functionality to a few core interactions, but serves the purpose of demonstrating the value of such an application.

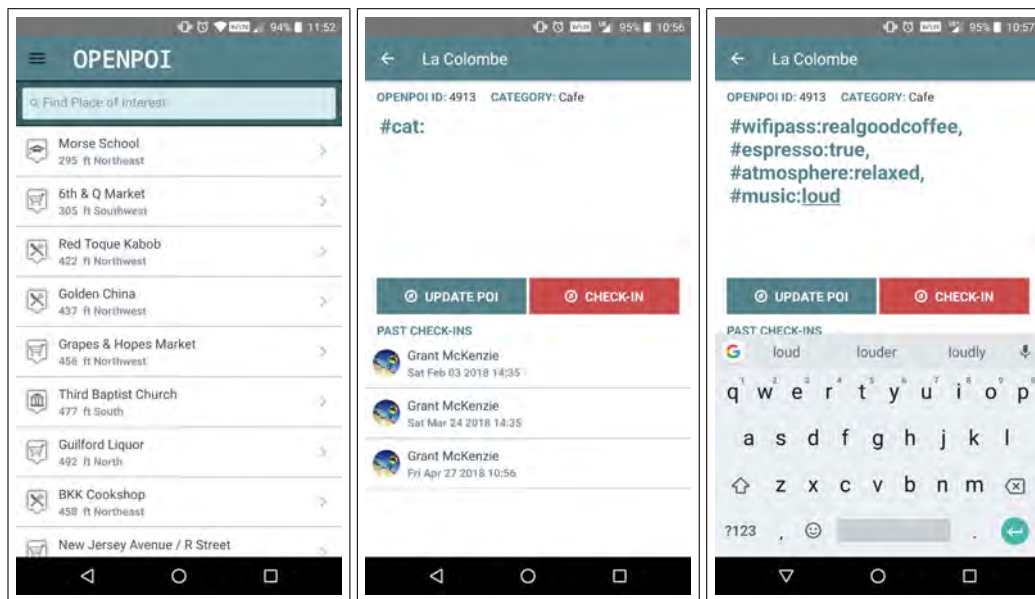
Upon logging in and ensuring the location services are enabled, a list of nearby places of interest are shown to the user ranked by proximity (Figure 1a). The *list view* shows the name, category icon,⁷ distance and direction from the device's current location. After selecting a POI from the list, the user is presented a screen listing descriptive information and permitting two forms of interaction. Users can check in to the POI by selecting the *check-in* button, in which case they are added to the database and list of previous check-ins to the specified POI (Figure 1b), or they can choose to update the POI with attribute information (Figure 1c). Virtually any type of descriptive textual information can be added to the POI on this screen using a *key-value pair*. Through this method, users can specify a *key* term such as *wifipass* by prepending a hashtag symbol. This term is then followed by a colon and the value associated with the key term. Attribute information is separated by these hashtags allowing for free text entry of any information. Currently, the application prompts user to update the category of the POI, but future versions will suggest potential key terms that may be most useful for the specified POI.

While the unrestricted ability to add any type of character-based content to a POI will undoubtedly lead to noisy data, the purpose of this application is not to restrict what people

⁵ Source Code: <https://github.com/ptal-io/OpenPOI-App>

⁶ <http://www.reactnative.com/>

⁷ Currently based on OSM's amenity category taxonomy



(a) Nearby places of interest. (b) POI details and check-in screen. (c) Adding *key-value hashtag pairs*.

■ **Figure 1** Three screens of the mobile OpenPOI application.

can or cannot enter, it is to get as much content as possible contributed to the application so that users can decide for themselves which information they care about and researchers have access to a wide variety of data. In today's area of *big data* and machine learning, it is much easier to clean, organize and extract meaning from large, noisy data than to work with a very limited supply of well structured content. Along these same lines, the underlying motivation for this application is not financial, meaning that clean, curated, and validated data is a secondary thought after free and open access to a large, heterogeneous, POI-specific dataset.

6 Summary & Next Steps

As digital gazetteers and POI datasets becomes increasingly silo-ed behind commercial firewalls, additional efforts must be made to ensure continued access to these types of geospatial information. In this short paper we introduced the OpenPOI platform and provide a brief overview of the components, functionality, and motivation for its development. We believe that such a dataset and services will be of value for the research community and act as a *research enabler* for many researchers in a wide range of disciplines.

Next steps for this platform will focus on three primary areas. First, a robust automated work flow for the extraction and merging of global places of interest from OpenStreetMap is in development. This process will merge the latest updates and changes from the OSM community with the rich attribute information and check-ins added through the OpenPOI platform. Additional effort will focus on inclusion and conflation of other data sources. Second, further functionality for interacting with the OpenPOI dataset, e.g., adding new places and updating geometry, are in the works along with associated API documentation. Last, further development on the mobile application will focus on rigorous testing of the core features, addition of a mapping screen, and overall interface development.

References

- 1 Josh Constine. Facebook restricts apis, axes old instagram platform amidst scandals. *TechCrunch*, April 2018. Online; 04-04-2018.
- 2 The Economist. The battle for territory in digital cartography. *The Economist*, June 2017. Online; 08-06-2017.
- 3 Michael F Goodchild and Linda L Hill. Introduction to digital gazetteer research. *International Journal of Geographical Information Science*, 22(10):1039–1044, 2008.
- 4 James M Hargett. Song dynasty local gazetteers and their place in the history of difangzhi writing. *Harvard Journal of Asiatic Studies*, 56(2):405–442, 1996.
- 5 Grant McKenzie and Krzysztof Janowicz. Coerced geographic information: The not-so-voluntary side of user-generated geo-content. In *Eighth international conference on geographic information science*, 2014.
- 6 Grant McKenzie, Krzysztof Janowicz, and Benjamin Adams. A weighted multi-attribute method for matching user-generated points of interest. *Cartography and Geographic Information Science*, 41(2):125–137, 2014.
- 7 Greg Miller. The huge, unseen operation behind the accuracy of google maps. *The Economist*, December 2014. Online; 08-12-2014.
- 8 Jennifer C Molloy. The open knowledge foundation: open data means better science. *PLoS biology*, 9(12):e1001195, 2011.
- 9 Jack Nicas. Google rival yelp claims search giant broke promise made to regulators. *The Wall Street Journal*, September 2017. Online; 11-09-2017.
- 10 Tessio Novack, Robin Peters, and Alexander Zipf. Graph-based matching of points-of-interest from collaborative geo-datasets. *ISPRS International Journal of Geo-Information*, 7(3):117, 2018.
- 11 Google Support. Popular times, wait times, and visit duration. Technical report, Alphabet Inc., 2016. Accessed: 28-04-2018.

Exploring Shifting Densities through a Movement-based Cartographic Interface

Aline Menin

Univ. Grenoble Alpes, CNRS, Grenoble INP¹, LIG, 38000 Grenoble, France
aline.menin@univ-grenoble-alpes.fr

Sonia Chardonnel

Univ. Grenoble Alpes, CNRS, Science Po Grenoble², PACTE, 38000 Grenoble, France
sonia.chardonnel@univ-grenoble-alpes.fr

Paule-Annick Davoine

Univ. Grenoble Alpes, CNRS, Grenoble INP¹, LIG, 38000 Grenoble, France and Univ. Grenoble Alpes, CNRS, Science Po Grenoble², PACTE, 38000 Grenoble, France
paule-annick.davoine@univ-grenoble-alpes.fr

Luciana Nedel

Federal University of Rio Grande do Sul, Institute of Informatics, Porto Alegre, Brazil
nedel@inf.ufrgs.br

Abstract

Animated maps are widely used for representing shifting densities. Though there is evidence that animations can provide better memory recall than static charts, it could be a consequence of using better techniques for animation than for static representations. However, the lack of control makes them frustrating for users, while animated choropleth maps can cause change blindness. In this paper, we propose an interactive animation technique which employs the lenticular printing metaphor and benefits from the user's proprioceptive sense to explore density changes over time. We hypothesized that using a tangible interface based on the body movement would improve memory recall and, consequently, animated map reading.

2012 ACM Subject Classification Human-centered computing → User interface design, Human-centered computing → Visualization, Human-centered computing → Geographic visualization, Human-centered computing → Gestural input, Human-centered computing → Mobile computing

Keywords and phrases proprioceptive interaction, lenticular technique, shifting densities, tangible interfaces, mobility analysis

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.48

Category Short Paper

Acknowledgements We would like to acknowledge the Region Auvergne-Rhône-Alpes for funding this research.

1 Introduction

Shifting densities describe the density changes in different space areas over time. Studying them support the measurement of urban mobility while exploring indicators such as attractiveness changes and the space use frequency according to the activities performed [19].

¹Institute of Engineering Univ. Grenoble Alpes

²School of Political Studies Univ. Grenoble Alpes



This indicator is often represented through animated choropleth maps. Even if there is evidence that animation can improve performance on memory-recall and map-reading if compared to static graphs [14], this representation has shown drawbacks like change blindness. Consequently, users regularly fail to detect basic changes within animated choropleth maps [13]. They can be frustrating while dealing with complex changing maps that are difficult to control [14].

Tversky et al. [21] found out that most of the so-called successful applications of animation turn out to be a consequence of a better visualization or study procedures such as interactivity or prediction that are known to improve learning independent of graphics. The same authors say that the drawback of animation may be perceptual and cognitive limitations in the processing of a changing visual situation. Animations are fleeting, they disappear, and when they can be reinspected, this is done in motion, where it may be difficult to perceive all the minute changes simultaneously.

Interactivity could be the key to overcome the drawbacks of animation while improving learning and giving the user the power of controlling speed, stop and start, zoom in and out, and so on. Therefore, we propose a motion-based interaction for map animation, which explores the lenticular printing metaphor combined with the user's proprioceptive sense to explore the use of the space in urban areas represented by animated choropleth maps. Our approach brings together the benefits of controlling the animation by using the lenticular foil technique, which allows the user to see spatial information separately, and to see relations and dependencies of phenomena by changing the view [11], and making use of their proprioceptive sense by tilting a mobile device.

2 Related Work

Previous studies have proposed visualization and interaction techniques to improve animation effectiveness for shifting densities exploration. André-Poyaud et al. [1] present the concept of territorial heartbeat, through which one could observe the density variation by sensing the city pulse: in the morning, the periphery people move towards the main agglomerations, which receive even more density, and then they leave gradually from the end of the afternoon to beginning of the evening. This technique has been explored in combination with the 3D view, by varying the z -axis height and color according to the density change on different map regions [15, 5]. Le Roux et al. [17] use animated timelines to represent the social segregation over time. Users can control time periods by pausing/playing or directly clicking on the timeline to choose periods.

In thematic cartography, researchers have explored the lenticular foil technique for improving map-reading and, especially, for displaying 3D effects without any specific device, e.g. glasses, which is called *true-3D*. Lenticular printing uses lenticular lenses to change or move the image as it is viewed from different angles. Cartography could benefit from the possibility of giving the user different information from the respective content layers since they can be visualized separately by auto-stereoscopic accentuation or sequential insertion. Moreover, it could reduce the drawbacks of graphic density and its associated bad legibility, as well as improving information communication [11].

The lenticular foil technique can display both 2D and 3D effects. Up to this date, it has been mostly applied to display information by using *true-3D*. Buchroithner et al. [7] employed this technique to create an interactive map of the Granatspitz Massif in the Eastern Alps aiming to exhibit the touristic places in the region. Similarly, Wagman [23] uses the same technique to support tourism in Manhattan city, in which the viewer has the impression of

seeing several layers of information, almost like an hologram. The map can be seen from three different angles, showing the New York's subway system, the neighborhood, and the streets grid.

By using visualization environments composed of immersive and stereoscopic augmented reality combined with tangible input, Bach et al. [3] showed that direct manipulation with 3D holographic visualizations improves time and accuracy for tasks that require coordination between perception and interaction. In 3D visualizations employing spatial information with mobile devices, Buschel et al. [8] found that users perceived spatial interaction as more supportive, comfortable and preferable to touch input.

Besançon et al. [6] explore the possibility of using both tactile and tangible input for fluid dynamics data visualization using a portable, position-aware device. Their approach was better appreciated by the users than a traditional mouse-and-keyboard setup. Moreover, Arvola and Holm [2] showed that device-orientation based panning on hand-held devices is useful when engagement is considered important, and their results strengthen the idea that more intensive bodily interaction can be more engaging. In fact, the user's proprioceptive sense could help to retain information by using their body's position as a recall reference. This sense corresponds to user's feeling regarding the pose of their own body and the strength or effort being employed in the movement. It could assist interaction through tangible user interfaces, in which a person interacts with the digital information through the physical environment.

3 Space Use over Time

Studies on shifting density are of interest for areas as diverse as crisis and health management, social segregation, or mobility issues. Bañgate et al. [4] propose a multi-model of human behavior during seismic crisis based on the social attachment theory. During the simulation of the model, they consider where people are performing their activities at each hour of the day. Likewise, Davoine et al. [10] proposed a visualization tool to improve the study of social vulnerability considering spatio-temporal variations regarding visited places at personal and professional aspects.

The investigation of social segregation helps to dynamically consider place effects on individual behavior and to target areas to implement interventions more connected with the real rhythm of the city [17]. Moreover, the day course perspective may help to isolate some critical and sensitive periods in which changes in place attributes occur, as well as days and nights constitute an important timescale for humans as they impose a biological rhythm, which helps to understand the mobility impacts upon health [22].

In France, Household Travel Surveys (HTS) constitute a valued database on urban mobility. We explored the data from a large HTS regularly carried out in the Rhône-Alpes region since 1976, from which we recovered the 2010 edition [9]. This survey provides a large amount of information on the daily mobility of inhabitants aged five and older, as well as on the household and individual aspects. Displacements are described through origin-destination information, which contains information about departure/arrival sector and time.

In this study, we calculated presence density and migration rate varying along a 24-hour period. The first one refers to all people staying in the referred sector, while the second refers to the difference between present people and the sector population.

4 Lenticular Printing Metaphor

4.1 Design Rationale

Animations are appropriate when we need to use multiple maps to represent information changing over time. However, it should be interactive to properly replace static maps, which also facilitates analysis when allowing the user to choose viewpoints [12]. Dorling also points out the brain's poor visual memory as being a problem when animating time. Therefore, we believe the proprioceptive sense could assist the improvement of memory recall while using animation for the analysis of spatio-temporal information.

In virtual reality, proprioception aids users to orientate themselves spatially inside virtual environments [18], and improves object manipulation, which is also better performed when using a handheld object to guide the user from the physical space [20]. Additionally, the use of tangible user interfaces reduces the cognitive workload, while physical mobility may increase user creativity, which indicates that less constrained interaction styles are likely to help users to think and communicate. Tangible interfaces that engage the body can leverage body-centric experiential cognition [16].

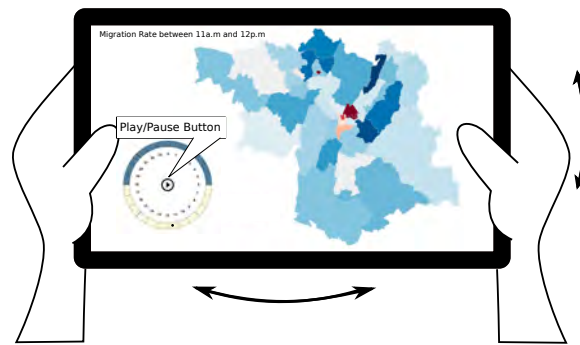
Based on these assumptions, our animation technique uses a mobile device as visualization and interaction interface and grants interactivity through lenticular effects (see Figure 1). We implemented the morphing effect, which changes one image into another through a seamless transition and, thus, it is suitable to represent a series of spatial events gradually [11] along a 24-hour period. Since these transitions are activated by tilting the device, the user could use their wrists orientation as a physical reference to recall the information seen on the screen.

4.2 Implementation and preliminary results

In order to calculate presence density, individuals who reported staying at home all day were assigned to their residence sector during all observation time. Individuals that moved during the day can be considered either *visitors* (people that do not reside in the current sector they are staying) or *residents* (people whose displacements were performed inside their residence sector). Time periods have one-hour duration. Then, for each hour we calculated the number of people staying in each sector. Displacements were recorded from 4:00 AM to 3:59 AM, then we considered people were at their first origin sector from 4:00 AM to the departure time of their first displacement, and at their last destination sector from their last displacement arrival time to 3:59 AM. Following the approach of Le Roux et al. [17], we did not take people that were moving into account.

The density was calculated by dividing the number of persons present in the sector at each time period by the sector surface in square kilometers. The migration rate was determined by the difference between the current population and the number of inhabitants in the sector.

The map can display either the presence density or migration rate at the time. For the first one, we vary density from light to dark red, while for the second one, we vary migration rate from blue nuances (when the sector loses population) to red nuances (when the sector gains population). Time periods can be selected by cyclically tilting the mobile device, from which we recover accelerometer information. This data is mapped to a time period by computing the tilting angle and then matching it with the corresponding period. The indicators are dynamically updated according to the selected time. Finally, the user also disposes of a play/pause button (in the middle of the clock) and the velocity animation is determined by the user's movement speed while tilting the device. We used D3 Data-Driven Documents javascript library for developing our application, which holds a choropleth map and a 24-hour clock to select the time.



■ **Figure 1** The desired time period is selected by tilting the tablet. During the movement, the clock (left) shows the current time period and the map is animated accordingly. Animation can be stopped by pressing the play/pause button and the movement speed is set by the tilting velocity.

5 Final Comments and On-going Work

In this paper, we introduced a technique for exploring animated maps based on a natural and tangible interface. We use the lenticular printing approach to visualize changes in mobility indicators along a 24-hour period by tilting a mobile device. By using a tangible interface we benefit from the user's proprioceptive sense, which aims to improve memory-recall and, consequently, map-reading.

This work is part of a greater project, in which we intend to develop a geovisualization tool for exploring individual mobility data by combining non-conventional interactions and successful visualization techniques for mobility analysis. Therefore, the next step consists of evaluating the proposed technique to test our hypothesis that it really improves animated map-reading.

References


- 1 Isabelle I. André-Poyaud, Sonia Chardonnel, Laure L. Charleux, and Kamila Tabaka. La mobilité au cœur des emplois du temps des citadins. In Florence Paulhiac Yves Chalas, editor, *La mobilité qui fait la ville*, Débats, pages 67–95. CERTU, 2008.
- 2 Mattias Arvola and Anna Holm. Device-orientation is more engaging than drag (at least in mobile computing). In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, pages 939–942. ACM, 2014.
- 3 Benjamin Bach, Ronell Sicat, Johanna Beyer, Maxime Cordeil, and Hanspeter Pfister. The hologram in my hand: How effective is interactive exploration of 3d visualizations in immersive tangible augmented reality? *IEEE transactions on visualization and computer graphics*, 24(1):457–467, 2018.
- 4 Julius Bañgate, Julie Dugdale, Elise Beck, and Carole Adam. Solace a multi-agent model of human behaviour driven by social attachment during seismic crisis. In *4th International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*, pages 1–9. IEEE, 2017.
- 5 Arnaud Banos and Thomas Thévenin. La carte animée pour révéler les rythmes urbains. *Revue Internationale de Géomatique*, 15(1):pp–11, 2005.
- 6 Lonni Besançon, Paul Issartel, Mehdi Ammi, and Tobias Isenberg. Hybrid tactile/tangible interaction for 3d data exploration. *IEEE transactions on visualization and computer graphics*, 23(1):881–890, 2017.

- 7 MF Buchroithner, Klaus Habermann, and Thomas Gruendemann. True 3d visualization of mountainous terrain by means of lenticular foil technology. In *4th ICA Mountain Cartography Workshop*, pages 125–135, 2010.
- 8 Wolfgang Büschel, Patrick Reipschläger, Ricardo Langner, and Raimund Dachselt. Investigating the use of spatial interaction for 3d data visualization on mobile devices. In *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces*, pages 62–71. ACM, 2017.
- 9 CEREMA. *Enquête Ménages Déplacements (EMD), Grenoble, Grande région grenobloise*. Archives de Données Issues de la Statistique Publique, 2010.
- 10 Paule-Annick Davoine, Elise Beck, Isabelle André-Poyaud, Sonia Chardonnel, Céline Lutoff, and Anton Telechev. Géovisualisation pour la réduction de la vulnérabilité socio-spatiale en milieu urbain. *Comité Français de Cartographie*, 211:69–84, 2012.
- 11 Frank Dickmann. The potential of the lenticular foil technique for thematic cartography. *The Cartographic Journal*, 47(3):250–256, 2010.
- 12 Daniel Dorling. Stretching space and splicing time: from cartographic animation to interactive visualization. *Cartography and Geographic Information Systems*, 19(4):215–227, 1992.
- 13 Carolyn Fish, Kirk P Goldsberry, and Sarah Battersby. Change blindness in animated choropleth maps: an empirical study. *Cartography and Geographic Information Science*, 38(4):350–362, 2011.
- 14 Mark Harrower. Visualizing change: Using cartographic animation to explore remotely-sensed data. *Cartographic Perspectives*, 39:30–42, 2001.
- 15 Christophe Hurez. Localisation spatiale et temporelle des personnes et des voitures à partir des enquêtes ménages déplacements. <https://mappemonde-archivage.mgm.fr/num27/fig10/fig10302.html>, 2010. Accessed: 2018-04-23.
- 16 Scott R Klemmer, Björn Hartmann, and Leila Takayama. How bodies matter: five themes for interaction design. In *Proceedings of the 6th conference on Designing Interactive systems*, pages 140–149. ACM, 2006.
- 17 Guillaume Le Roux, Julie Vallée, and Hadrien Commenges. Social segregation around the clock in the paris region (france). *Journal of Transport Geography*, 59:134–145, 2017.
- 18 Anderson Maciel, Luciana P Nedel, Vitor AM Jorge, Juan MT Ibiapina, and Luis FMS Silva. Reality cues-based interaction using whole-body awareness. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, pages 1224–1228. ACM, 2010.
- 19 Matthieu Mille. Des densités habitantes aux densités mouvantes l'exemple de la métropole lilloise. *Cybergeo: European Journal of Geography*, 2000.
- 20 Mark R Mine, Frederick P Brooks Jr, and Carlo H Sequin. Moving objects in space: exploiting proprioception in virtual-environment interaction. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 19–26. ACM Press/Addison-Wesley Publishing Co., 1997.
- 21 Barbara Tversky, Julie Bauer Morrison, and Mireille Betrancourt. Animation: can it facilitate? *International journal of human-computer studies*, 57(4):247–262, 2002.
- 22 Julie Vallée. The daycourse of place. *Social science & medicine*, 194:177–181, 2017.
- 23 David Wagman. Urban maps gives a vintage printing technique a high-tech revival. <https://www.directionsmag.com/article/3272>, 2009. Accessed: 2018-04-20.

Geotagging Location Information Extracted from Unstructured Data

Kyunghyun Min

Department of Civil and Environmental Engineering, Seoul National University 35-209,
Gwanak-gu, Seoul, Republic of Korea
minkh@snu.ac.kr

 <https://orcid.org/0000-0002-4835-7215>

Jungseok Lee

Department of Civil and Environmental Engineering, Seoul National University 35-209,
Gwanak-gu, Seoul, Republic of Korea
rightstone@snu.ac.kr

Kiyun Yu

Department of Civil and Environmental Engineering, Seoul National University 35-209,
Gwanak-gu, Seoul, Republic of Korea
kiyun@snu.ac.kr

Jiyoung Kim

Institute of Construction and Environmental Engineering, Seoul National University 35-215,
Gwanak-gu, Seoul, Republic of Korea
soodaq@snu.ac.kr

Abstract

Location information is an essential element of location-based services and is used in various ways. Unstructured data contain different types of location information, but coordinate values are required to determine the exact location. In Twitter, a typical social network service (SNS) platform of unstructured data, the number of geotagged tweets is low. If we can estimate the location of text by geotagging a large number of unstructured data, we can estimate the location of the event in real-time. This study is a base study on extracting the location information by using the named entity recognizer provided by the Exobrain API and applying geotagging to unstructured data in Hangul (Korean). We used Chosun news articles, which are grammatically correct and well organized, instead of tweets to extract three location-related categories, namely “location,” “organization,” and “artifact”. We used the named entity recognizer and geotagged each sentence in combination of the fields in each category. The results of the study showed that 61% of the 800 test sentences did not have the location-related information, thus hindering geotagging. In 11.75% of the test sentences, geotagging was possible with only the given location information extracted using the named entity recognizer. The remaining 27.25% of the sentences contained information on more than two locations from the same subcategories and hence required location estimation from candidate locations. In future research, we plan to apply the results of this study to develop location estimation algorithm that makes use of the extracted location-related entities from purely unstructured data such as that on SNSs.

2012 ACM Subject Classification Information systems → Content analysis and feature selection

Keywords and phrases Location Estimation, Information Extraction, Geo-Tagging, Location Information, Unstructured Data

Digital Object Identifier 10.4230/LIPICs.GIScience.2018.49

Category Short Paper



© Kyunghyun Min, Jungseok Lee, Kiyun Yu, and Jiyoung Kim;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 49; pp. 49:1–49:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

■ **Table 1** Percentages of Tweets with Location Information.

	Max	Min	Average
% of Geotagged Tweets Per Day	0.22%	0%	0.11%

Funding This study was supported by the research funding of the project on the development of big data management, analysis, and service platform technology for the national land spatial information research project of the Ministry of Land, Infrastructure, and Transport (18NSIP-B081023-05).

1 Introduction

Recently, location-based services are growing rapidly owing to the large amount of data generated in people's lives. A person's behavior or the occurrence of an event is often accompanied by location information. Recently, the use of social network services (SNSs) has increased as a method for human expression. However, less than 0.42% of tweets were geotagged even though Twitter is providing a function to determine the location information [9]. In fact, we collected 611,687 tweets for the entire month of March 2018 and confirmed that they are geotagged only on an average of 0.11% tweets a day, as shown in Table 1. If the tweet is geotagged, a location where a specific article was written or a location that it describes is known. Hence, an incident or an accident mentioned in the SNS or the news article can be checked in real time. Therefore, by extracting the location information from these unstructured data and adding the location information, the occurrence of a specific event and its location can be monitored.

As mentioned earlier, the number of geotagged SNSs is small. As a result, many studies have been carried out only on geotagged posts [5, 7]. Therefore, other factors such as user profiles, text content, and location labeling are used to aid in an ongoing location estimation [4]. One study detected earthquake in real time and inferred the location from the registered location and GPS data created when users sign up the unstructured data platform, Twitter [8]. Further, a study on the extraction of location-related entities from each tweet on twitter using named entity recognition and concept-vocabulary-based extraction has been performed [1]. Recently, research has been performed to detect the location information in text using the conditional random fields (CRF) model [3]. However, a case in which the location information is extracted by using the named entity recognizer for Hangul (Korean) does not exist. In this study, we aim to geotag the unstructured Hangul data with the location information extracted with the entity recognizer.

2 Detection of Location Information by Named Entity Recognition

2.1 Named Entity Recognition

Named entities are the names of persons, organizations, locations, dates, and times. Named entity recognition refers to recognizing and tagging the corresponding entity name among proper names or noun phrases. Named entity recognition is one of the language analysis techniques that is essential in natural language processing tasks used in information retrieval or information extraction. In the English language, high-level recognition and classification performance were shown by using language characteristics such as capital letters [6]. However, in the Korean language, it is difficult to recognize an entity name in the absence of certain features such as capital letters in English. As an alternative,

■ **Table 2** Lists of items in each category (in Parts).

LC	OG	AF
LCP_COUNTRY	OGG_EDUCATION	AF_BUILDING
LCP_CAPITALCITY	OGG_SPORTS	AF_ROAD
LCP_COUNTY	OGG_FOOD	AF_TRANSPORT
LCP_CITY	OGG_HOTEL	AF_CULTURAL_ASSET
LC_TOUR	OGG_POLITICS	:
LCG_MOUNTAIN	OGG_RELIGION	:
LCP_PROVINCE	OGG_ECONOMY	:
:	:	:

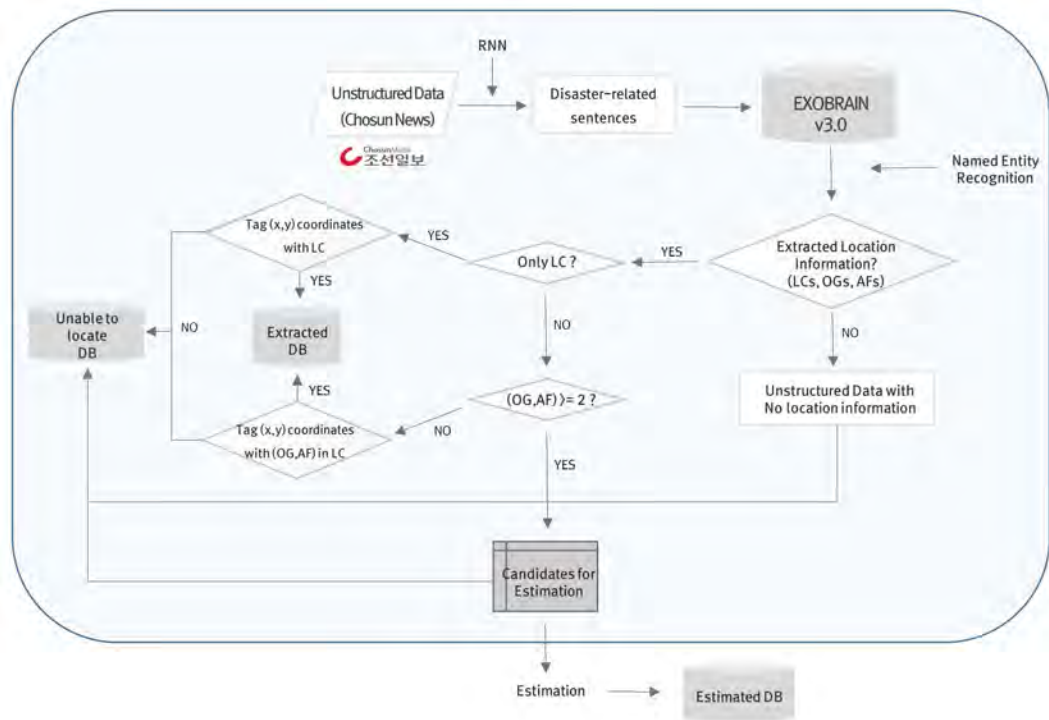
there is a study using word embedding features in recognition and classification of Korean entity names [2]. The entity name recognizer used in this study is the Exobrain language analysis open API provided by Korea Electronics and Telecommunications Research Institute (ETRI). The entity recognition corpus for Exobrain comprises of 10,000 sentences from news articles. It uses the Telecommunications Technology Association's (TTA) standard object name tag set consists of 15 main categories and 146 subcategories for object types in various fields. Location (LC), organization (OG), and artifact (AF) were selected as the necessary main categories for this study. Subcategories that can be used to extract location-related information are partly introduced in Table 2. There are fourteen subcategories for LC, fifteen for OG, and thirteen for AF. LC contains the geographical name, the administrative district name, and the like. OG contains the names of educational institutions, medical institutions, accommodations, and the like. AF indicates the names of cultural properties, buildings, and roads.

2.2 Extracting Location Information

The workflow of this study is presented in Figure 1. We extracted the LC, OG, and AF information from sentences related to fire accidents by using the entity recognizer. If no location-related information that belongs to the three major categories is obtained in the sentence, such a sentence is stored in a database (DB) that cannot be geographically located by geotagging. If extracted location information are geographically hierarchical, the coordinates corresponding to the area are tagged and stored in the extracted DB. If only one OG or one AF information exists in addition to the LC, only one coordinate value can be assigned. However, if more than two OG or AF information are to be assigned, the allocation of the location cannot be determined. In other words, if the text, in this case a sentence, is mentioning more than two locations that are not geographically hierarchical, then location estimation is needed. In our future study, several OGs and AFs will be temporarily stored as estimated candidates so that location estimation can proceed.

3 Test and Results

The sentences used in this study are 800 fire accident-related sentences from the Chosun news articles published in 2017. Since tweets are written by the users in colloquial style that is hard for computers to understand, we chose news articles as an alternative as they are grammatically correct and well structured. To geotag the sentence, not estimate, at least one LC is required. For example, if "Starbucks" is the only retrieved location information



■ **Figure 1** Work Flow Chart.

for OG, the specific Starbucks branch cannot be determined because there are more than thousand Starbucks stores in Korea. As many as 488 sentences through the named entity recognizer did not contain location information, comprising 61% of the total number of sentences. In contrast, sentences with location information including LC, OG, or AF, were 312 in number. Among them, only 94 sentences, i.e., only 11.75% out of the total, could be geotagged; for the remaining 27.25% of sentences, location estimation is required. The results are summarized in Table 3. Figure 2 shows the example visualization of named entity recognition and morphological analysis performed using the Exobrain API. The sentence at the top is written in Hangul, and the one below is the corresponding translated sentence.

4 Conclusion

Recently, the use of SNS has increased, but the location information extracted from unstructured data is lacking. We confirmed the lack of geotagging through the twitter data collected for a month and aimed to solve it through the location estimation from the named entity recognition. In this study, geotagging was performed by extracting the location-related information on LC, OG, and AF from fire accident-related sentences using the Exobrain named entity recognizer as a base study for location estimation. Our experimental results showed that 61% of 800 sentences had no extracted location information, 11.75% of sentences were geotagged, and 27.25% of sentences required location estimation. As the number of sentences has a large number of candidates that can be used for estimation, future studies will focus on improving the accuracy using named entity recognition and CRF model, and the location information can be provided to more unstructured data by developing a location estimation algorithm that uses the extracted location information.

Table 3 Application Result.

	LC, OG, AF		No Location Information	Total
	Extracted Location	Location Estimation needed	-	
Named Entity Recognition	312		488	800
Percentage [%]	11.75%	27.25%	61%	100%

밀양 세종병원 화재는 2018년 1월 26일 경상남도 밀양시 중앙로 114(가곡동)에 있는 세종병원에서 발생한 화재 사고이다.
 LCP_CITY OGG_MEDICINE LCP_PROVINCE LCP_CITY AF_ROAD LCP_COUNTY OGG_MEDICINE

The fire in the Miryang Sejong Hospital is a fire accident at Sejong Hospital on
 LCP_CITY OGG_MEDICINE OGG_MEDICINE
 January 26, 2018 in Gangok-dong, 114, Middle Road, Miryang-si, Gyeongsangnam-do.
 LCP_COUNTY AF_ROAD LCP_CITY LCP_PROVINCE



Figure 2 Example of named entity recognition result for fire-related sentence.

References

- 1 Puneet Agarwal, Rajgopal Vaithiyathan, Saurabh Sharma, and Gautam Shroff. Catching the long-tail: Extracting local news events from twitter. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, pages 379–382, 2012.
- 2 Yunsu Choi and Jeongwon Cha. Korean named entity recognition and classification using word embedding features. In *Journal of Korean Institute of Information Scientists and Engineers*, pages 678–685, 2016.
- 3 Diana Inkpen, Ji Liu, Atefeh Farzindar, Farzaneh Kazemi, and Diman Ghazi. Location detection and disambiguation from twitter messages. *Journal of Intelligent Information Systems*, 49(2):237–253, 2017.
- 4 Farhad Laylavi, Abbas Rajabifard, and Mohsen Kalantari. A multi-element approach to location inference of twitter: A case for emergency response. *ISPRS International Journal of Geo-Information*, 5(5):56, 2016.
- 5 Ryong Lee, Shoko Wakamiya, and Kazutoshi Sumiya. Discovery of unusual regional social activities using geo-tagged microblogs. *World Wide Web*, 14(4):321–349, 2011.
- 6 Xiaohua Liu, Ming Zhou, Furu Wei, Zhongyang Fu, and Xiangyang Zhou. Joint inference of named entity recognition and normalization for tweets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 526–535. Association for Computational Linguistics, 2012.


49:6 Geotagging Location Information Extracted from Unstructured Data

- 7 Kenta Oku, Koki Ueno, and Fumio Hattori. Mapping geotagged tweets to tourist spots for recommender systems. In *Advanced Applied Informatics (IIAIAAI), 2014 IIAI 3rd International Conference on*, pages 789–794. IEEE, 2014.
- 8 Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.
- 9 Cheng Zhiyuan, Caverlee James, and Lee Kyumin. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 759–768, 2010.

Linked Open Data Vocabularies for Semantically Annotated Repositories of Data Quality Measures

Franz-Benjamin Mocnik

Heidelberg University, Institute of Geography
Im Neuenheimer Feld 348, 69120 Heidelberg, Germany
mocnik@uni-heidelberg.de

 <https://orcid.org/0000-0002-1759-6336>

Abstract

The fitness for purpose concerns many different aspects of data quality. These aspects are usually assessed independently by different data quality measures. However, for the assessment of the fitness for purpose, a holistic understanding of these aspects is needed. In this paper we discuss two Linked Open Data vocabularies for formally describing measures and their relations. These vocabularies can be used to semantically annotate repositories of data quality measures, which accordingly adhere to common standards even if being distributed on multiple servers. This allows for a better understanding of how data quality measures relate and mutually constrain. As a result, it becomes possible to improve intrinsic data quality measures by evaluating their effectivity and by combining them.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases data quality, measure, semantics, Linked Open Data (LOD), vocabulary, repository, reproducibility, OpenStreetMap (OSM)

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.50

Category Short Paper

Supplement Material <http://purl.org/data-quality>, <http://purl.org/osm-data-quality>

Funding This work was supported by the DFG project *A framework for measuring the fitness for purpose of OpenStreetMap data based on intrinsic quality indicators* (FA 1189/3-1).

1 Introduction

Data quality and fitness for purpose are major issues for many applications. Are the data of use for a certain application because they are capable of delivering the desired result? Applications each have their own requirements: certain aspects of the data might be more important than other ones for a specific application. Data quality measures quantify how usable the data are in respect to a certain aspect of the data. Among such aspects are the completeness of the data, logical consistency, positional and thematic accuracy, temporal quality, etc. [6] As in many cases no reference data are available – the reference data would then be used instead of the considered data – one aims for *intrinsic measures*, which evaluate aspects of data quality by, for the most part, only referring to the data themselves.

While often examining different aspects of data quality independently, a holistic view is needed in many practical examples. In case of vehicle routing, the completeness of the representation of the road network and the topological quality play a major role, but the geometric quality and the thematic accuracy have an impact as well. The same is true for many other applications: whether a dataset is fit for a certain purpose can only be evaluated



© Franz-Benjamin Mocnik;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 50; pp. 50:1–50:7

Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

when assessing all concerned aspects of the data. Therefore, a repository of data quality measures should ideally address the following needs:

- (N1) **Formal harmonization of measures.** Measures can often not be related because they are implemented independently. Their results are semantically incompatible and their descriptions in publications stay often unrelated. Common standards, including semantic descriptions, allow for harmonizing and combining measures.
- (N2) **Situational interpretation of measures.** When assessing data quality, the results need interpretation. Measures often presume a certain context and work only in a certain setting – they mutually constrain. A repository allows for relating measures to gain a situational interpretation of their results if the relations and dependencies are formally described.
- (N3) **Traceability of complex results.** Data quality measures are described and evaluated in scientific publications but their algorithms are often not properly documented. The publication in a repository under an open license and the semantic annotation allow for tracing how individual measures lead to a complex assessment of data quality.

In this article, we discuss how data quality measures and their dependencies can be described as Linked Open Data (LOD). First, we shortly summarize related work (Section 2). Subsequently, we discuss properties of data quality measures, including relations between different measures (Section 3). These properties are formalized in two vocabularies, which can be used to annotate data quality measures as LOD (Section 4). Such annotations allow for a harmonization of data quality measures and, accordingly, for examining them as a whole. The structure of a repository of data quality measures is discussed by referring to the role that the LOD vocabularies may take in this context (Section 5).

2 Related Work

Numerous data quality measures have been discussed in literature. Senaratne et. al. [13] list measures for Volunteered Geographic Information in general, and Mocnik et. al. [10] for OSM in particular. Such measures can be classified by their grounding, i. e., by the source of information used to assess data quality. A corresponding ontology has been introduced by Mocnik et. al. [10]. Data quality aspects have been discussed by Wand and Wang [14] and been published as a norm [6]. Descriptions of data quality by the properties of the data have been complemented by descriptions of how the data can be used, the fitness for purpose [2, 5]. The concepts of fitness for purpose and data quality have been related by Devillers et. al. [4]. Couclelis has discussed differences between information and knowledge in respect to imperfection [3], which emphasizes the need to relate several data quality measures. Mocnik et. al. have discussed the comparison of intrinsic and extrinsic measures [11]. The importance of traceability has, among others, been discussed by Popper [12].

3 Properties of and Relations Between Measures

In this section, we discuss the semantic foundations of a repository of data quality measures. Both intrinsic and extrinsic measures are often constrained by a context or other measures, creating the need to formally capture such constraints and relations. In the following, we discuss how to describe measures and their interrelations formally. OpenStreetMap will serve as an example while the definitions apply to data quality measures in general.

Measures assign meta information to a dataset. As an example, the *saturation principle* can measure the completeness of a road network represented in some dataset [1]. Thereby,

it is measured whether the length of the road network still increases or already stagnates – stagnation occurs when the road network is (more or less) completely represented in the data. The measure assigns to the dataset meta information about its completeness, e. g., by the increase of the road network's length. In general, measures can be conceptualized as follows:

► **Definition 1.** A *measure* $\mu: D \rightarrow R$ is a function or algorithm that assigns to each dataset $d \in D$ a result $\mu(d) \in R$. A measure is called a *data quality measure* if the result refers to the quality of the dataset.

In geographical applications, measures are of particular interest if they describe a dataset spatially. Many datasets explicitly expose a spatial dimension while others include them implicitly [9, 7]. We call a measure spatial if its result explicitly exposes a spatial dimension, aggregated by a discrete grid. The saturation principle can, e. g., be applied independently to a collection of grid cells for assessing the completeness of the road network for each of them.

► **Definition 2.** A measure $\mu: D \times G \rightarrow R$ is called *spatial* in case of G being a discrete grid that tessellates some region in \mathbb{R}^n or S^n .

The saturation principle works in case of a road network for OSM [1] but it remains unclear whether it also works in other contexts, e. g., for the electrical grid. In addition, the principle only works in case that the increase of road length is in a meaningful interval. This fact can be expressed as a condition ξ to the information resulting from the measure: if the increase is outside a certain range, the measure cannot be expected to deliver meaningful information¹. Similar concepts even apply to other measures. We accordingly define:

► **Definition 3.**

- (a) A measure $\mu: D \rightarrow R$ is called to be *valid in a context* c if the result $\mu(d)$ has a meaningful interpretation in respect to c .
- (b) A spatial measure $\mu: D \times G \rightarrow R$ is called to be *valid in an area* $G' \subset G$ if $\mu(d, g)$ has a meaningful interpretation for all $(d, g) \in D \times G'$. The measure is called to *meet condition* ξ if μ is valid in the area $G' := \{g \mid \xi(g)\} \subset G$.

Many conditions cannot be provided in general but depend on the examined place. The saturation principle, e., g., only works if volunteers contribute data about the examined area. Otherwise, the length of the road network does not increase, independent of its completeness. A second measure can be used to examine the presence of mapping activity in a particular area and, in turn, to determine in which areas the saturation principle provides meaningful information. Such relations between measures can, more formally, be described as follows:

► **Definition 4.** A spatial measure $\mu: D \times G \rightarrow R$ is said to *presume another spatial measure* $\nu: \tilde{D} \times G \rightarrow \tilde{R}$ under a condition ξ if μ is valid in the area $G' \subset G$ where ν meets ξ .

Even in before evaluating a spatial measure by computing its result for some region, one might want to know what to expect from the measure. The saturation principle might, e. g., not be able to properly distinguish between a completeness of 95 and 100 per cent. If the repository contains information about such limits of the expected results, one can decide in before whether to evaluate the saturation principle. We define:

► **Definition 5.** Assume R to be a totally ordered set. Then, the *minimum/maximum* of a spatial measure $\mu: D \times G \rightarrow R$ is defined as the minimum/maximum for both components:

$$\min \mu := \min_{d,g} \mu(d, g) \quad \text{and} \quad \max \mu := \max_{d,g} \mu(d, g).$$

¹ It needs to be discussed in detail and in respect to each measure what *meaningful information* refers to.

■ **Table 1** Linked Open Data vocabulary for describing data quality measures.

Classes (selection)	Definition
<code>dq:measure, :dataQualityMeasure, :result</code>	Definition 1
<code>dq:spatialMeasure</code>	Definition 2
<code>dq:context</code>	Definition 3(a)
<code>dq:grounding</code>	grounding of a data quality measure [10]
Individuals (selection)	Definition
<code>dq:extrinsicPerceptionBasedGrounding</code>	perception-based grounding [10]
<code>dq:intrinsicDataBasedGrounding, :extrinsic...</code>	data-based grounding [10]
<code>dq:intrinsicGroundingInProcessedData, :ext...</code>	grounding in processed data [10]
<code>dq:intrinsicGroundingInRulesPatternsKnowledge, :extrinsicGroundingInRulesPatternsKnowledge...</code>	grounding in rules/patterns/knowledge [10]
Predicates (selection)	Definition
<code>dq:implementedBy</code>	who implemented the measure
<code>dq:documentedBy</code>	who documented the measure
<code>dq:api</code>	URL of the REST API
<code>dq:typeOfResult</code>	Definition 1
<code>dq:assesses</code>	assessed data quality aspect [6]
<code>dq:validInContext, :validInArea</code>	Definition 3
<code>dq:usesGrounding</code>	refers to the grounding-based ontology of data quality measures [10]
<code>dq:presumes</code>	Definition 4
<code>dq:maximumResult, :minimumResult</code>	Definition 5

These formal definitions describe how measures relate and which properties they have. In the next section, we discuss how these formal definitions can semantically be expressed by the use of Linked Open Data (LOD) vocabularies.

4 Semantic Annotation Using Linked Open Data Vocabularies

The semantic annotation of a measure allows for a better interpretation of the measure's results and for an understanding of the context of the measure. When being able to relate measures by their semantics, one can make sense of them as a whole. Here, we discuss two new LOD vocabularies for semantically annotating measures, with the aim of expressing the definitions of the preceding section and of further properties.

The first of the two vocabularies describes data quality measures and their relations (`dq`; <http://purl.org/data-quality>; Table 1). The class `measure` represents measures in general; its subclass `dataQualityMeasure`, data quality measures; and its subclass `spatialMeasure`, spatial measures. If a measure is only valid in a certain context or area, this can be described by `validInContext` and `validInArea`, respectively. The predicate `presumes` expresses that a spatial measure presumes another one.

The vocabulary can also be used to represent the source of the data quality information when evaluating a data quality measure. Data refers to the environment by relating symbols to objects and processes, i. e., the data are grounded in the environment. When data is assessed, the original grounding is compared to an additional one, which is described by the

■ **Table 2** Linked Open Data vocabulary for describing data quality measures for OpenStreetMap.

Classes (selection)	Definition
<code>osmdq:spatialMeasure</code>	spatial measure (Definition 2) related to OSM
<code>osmdq:spatialDataQualityMeasure</code>	spatial data quality measure (Definition 2) related to OSM
<code>osmdq:elementType</code>	type of the OSM element (node, way, area, relation)
<code>osmdq:node, :way, :area, :relation</code>	OSM node, OSM way, OSM area, OSM relation
<code>osmdq:tag, :key, :value</code>	OSM tag, and corresponding key and value
Predicates (selection)	Definition
<code>osmdq:assessesElementType</code>	type of element that is assessed in particular
<code>osmdq:assessesTag</code>	tags of the elements assessed

grounding-based ontology of data quality measures [10]. The vocabulary allows for a formal representation of this ontology, by which data quality measures can be classified.

OSM-related data quality measures can be characterized by which elements they assess in the OSM dataset. This characterization is captured by a second LOD vocabulary (`osmdq`; <http://purl.org/osm-data-quality>; Table 2). In particular, `assessesElementType` describes whether a particular type of element is assessed (node, way, area, or relation). The predicate `assessesTag` refers to the tags of the elements that are assessed by the measure.

The two vocabularies described in this section can be used to annotate data quality measures and OSM-related data quality measures in particular. This allows for making sense of such measures as a whole, in particular when combining them. In the next section, we discuss the structure of a repository that contains semantically annotated measures.

5 A Repository of Quality Measures

A repository needs to expose executable algorithms as well as semantic information if it shall address the needs (N1)–(N3) of the introduction. Accordingly, different techniques have to be combined. Here, we exemplarily discuss which techniques can practically be used to build a repository² of data quality measures for addressing the needs (N1)–(N3).

The algorithm related to a measure is in many cases simple to understand, but its evaluation is often more complex than the central parts of the algorithm would suggest. For instance, the dataset needs to be distributed among a number of machines for efficient processing, the data need to be indexed, the history of the data might be made accessible, etc. The use of a common query language ensures the traceability of the results when the algorithms are made publicly available.

The measures in the repository should be semantically annotated by the vocabularies that have been discussed in the preceding section. Without semantics, it is hard to combine different measures and make sense of them as a whole. The use of the vocabularies, however, allows for a formal representation of the information necessary to combine different results and for taking account of mutual constraints between measures. When several measures and their results are combined, there is a need to trace how these results have been concluded. The use of formal vocabularies in combination with executable algorithms makes the evaluation of single measures and their interrelations between measures more transparent and traceable.

² see <https://osm-measure.geog.uni-heidelberg.de> for an exemplary implementation of these ideas

Both the algorithms and the semantic annotation can be stored in a repository using a version control system. In addition, they need to be offered on a website, where the semantic information is available as LOD. The algorithms can be run on a REST server³ that executes the code, aggregates by the ISEA3H Discrete Global Grid System⁴ [8], and caches the result. This setup ensures the effective use of the LOD vocabulary in the context of a repository.

6 Outlook

We have discussed how measures can relate and mutually constrain. In addition, we have introduced vocabularies for representing these relations and further properties. The vocabularies integrate well into a repository of data quality measures.

Intrinsic data quality measures only consume the data themselves. Despite this advantage, they can be unreliable because they cannot rely on any additional source of information. When comparing intrinsic and extrinsic measures by the use of a repository, one is able to trace the mutual dependencies of these measures. This allows for a better understanding of their relations and, as a consequence, improves the applicability of intrinsic measures.

Reasoners can take advantage of semantic annotations when relating measures. The formal representation of mutual dependencies allows thus for computationally combining data quality measures by their potentially similar (or dissimilar) results as well as by their mutual constraints, which renders synergy effects. As a result, more stable measures can be derived and data quality and fitness for purpose can be assessed more situationally.

References

- 1 Christopher Barron, Pascal Neis, and Alexander Zipf. A comprehensive framework for intrinsic OpenStreetMap quality analysis. *Transactions in GIS*, 18(6):877–895, 2014.
- 2 Nicholas R. Chrisman. The role of quality information in the long-term functioning of a geographic information system. *Cartographica*, 21(2):79–87, 1984.
- 3 Helen Couclelis. The certainty of uncertainty: GIS and the limits of geographic knowledge. *Transactions in GIS*, 7(2):165–175, 2003.
- 4 Rudolphe Devillers, Yvan Bédard, and Roberg Jeansoulin. Multidimensional management of geospatial data quality information for its dynamic use within GIS. *Photogrammetric Engineering and Remote Sensing*, 71(2):205–215, 2005.
- 5 Andrew U. Frank. Metamodels for data quality description. In Robert Jeansoulin and Michael F. Goodchild, editors, *Data quality in geographic information. From error to uncertainty*, page 15–29. Hermès, Paris, 1998.
- 6 International Organization for Standardization. ISO 19157:2013. Geographic information. Data quality, 2013.
- 7 Franz-Benjamin Mocnik. *A scale-invariant spatial graph model*. PhD thesis, Vienna University of Technology, 2015.
- 8 Franz-Benjamin Mocnik. A novel identifier scheme for the ISEA Aperture 3 Hexagon Discrete Global Grid System. *Cartography and Geographic Information Science*, 2018.
- 9 Franz-Benjamin Mocnik and Andrew U. Frank. Modelling spatial structures. *Proceedings of the 12th Conference on Spatial Information Theory (COSIT)*, page 44–64, 2015.

³ e. g., using <http://github.com/giscience/measures-rest>

⁴ e. g., using <http://github.com/giscience/geogrid> and <http://github.com/giscience/geogrid.js>

- 10 Franz-Benjamin Mocnik, Amin Mobasher, Luisa Griesbaum, Melanie Ecker, Clemens Jacobs, and Carolin Klöner. A grounding-based ontology of data quality measures. *Journal of Spatial Information Science*, 16, 2018.
- 11 Franz-Benjamin Mocnik, Alexander Zipf, and Hongchao Fan. The inevitability of calibration in VGI quality assessment. *Proceedings of the 4th Workshop on Volunteered Geographic Information: Integration, Analysis, and Applications (VGI-Analytics)*, 2017.
- 12 Karl Popper. *The logic of scientific discovery*. Routledge, London, 1992.
- 13 Hansi Senaratne, Amin Mobasher, Ahmed Loai Ali, Cristina Capineri, and Mordechai Haklay. A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science*, 31(1):139–167, 2017.
- 14 Yair Wand and Richard Y. Wang. Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11):86–95, 1996.

Need A Boost? A Comparison of Traditional Commuting Models with the XGBoost Model for Predicting Commuting Flows

April Morton

Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37830, USA
mortonam@ornl.gov

Jesse Piburn

Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37830, USA
piburnjo@ornl.gov

Nicholas Nagle

Department of Geography, University of Tennessee, Knoxville, 1000 Phillip Fulmer Way,
Knoxville, TN 37916, USA
nnagle@utk.edu

Abstract

Commuting models estimate the number of commuting trips from home to work locations in a given area. Since their infancy, they have been increasingly used in a variety of fields to reduce traffic and pollution, drive infrastructure choices, and solve a variety of other problems. Traditional commuting models, such as gravity and radiation models, typically have a strict structural form and limited number of input variables, which may limit their ability to predict commuting flows as well as machine learning models that might better capture the complex dynamics of the commuting process. To determine whether machine learning models might add value to the field of commuter flow prediction, we compare and discuss the performance of two standard traditional models with the XGBoost machine learning algorithm for predicting home to work commuter flows from a well-known United States commuting dataset. We find that the XGBoost model outperforms the traditional models on three commonly used metrics, indicating that machine learning models may add value to the field of commuter flow prediction.

2012 ACM Subject Classification Applied computing → Law, social and behavioral sciences

Keywords and phrases Machine learning, commuting modeling

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.51

Category Short Paper

1 Introduction

Knowing how many people commute from various home to work locations is important for solving problems in a wide variety of domains. Commonly referred to as commuting flows, these movements form a complex socio-economic network that can be used to better understand the transport of people, goods, money, information, and diseases at different spatial scales [7]. Having a better grasp of these processes is important for policy- and other decision-makers who aim to tackle a variety of issues such as reducing traffic and pollution, planning the development of new infrastructure, and preventing the spread of disease.

In response to the need for better understanding the movement of commuters, researchers have developed a suite of commuting models used for estimating population flows, planning



© April Morton, Jesse Piburn, and Nicholas Nagle;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 51; pp. 51:1–51:7

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

transportation systems, analyzing urban traffic, and many other applications [7, 8, 10, 5]. This collection of techniques has traditionally consisted of different versions of what are commonly known as gravity and radiation models [7]. In general, these models are based on simple equations with a small number of input variables that have been chosen based on the assumption that the number of trips between two locations is related to their residential and work populations, the distance between the locations and/or the number of opportunities (e.g. other jobs) between them [7].

Though useful, both gravity and radiation models are analytical models with crafted functional forms and a small number of input variables [9]. This potentially limits their ability to capture the more complex dynamics that more flexible models, such as machine learning algorithms, may be able to. To determine whether machine learning models might add value to the field of commuter flow prediction, we compare and discuss the performance of a standard gravity and radiation model with the XGBoost machine learning algorithm for predicting home to work commuter flows from a well-known United States (U.S.) commuting dataset. We find that the XGBoost model outperforms the traditional models on three commonly used metrics, showing promise for machine learning models in the field of commuter flow prediction.

2 Related Work

The goal of commuting modeling is to predict the matrix of commuters $T = (T_{ij})_{1 \leq i, j \leq n}$ that move from every zone i to every other zone j within a set of n distinct zones. Assuming there are a total of N commuters, the estimated matrix $\hat{T} = \hat{T}_{ij}$ is derived by first estimating the set of probabilities $(p_{ij})_{1 \leq i, j \leq n}$ that a randomly drawn commuter from the set of N commuters moves between all zones i and j , and then drawing at random N trips from the set of estimated probabilities $(\hat{p}_{ij})_{1 \leq i, j \leq n}$. Oftentimes, additional constraints are added to ensure that the total number of commuters m_i leaving each zone i , the total number of commuters n_j working in each zone j , or both, is preserved.

In order to estimate the probabilities $(p_{ij})_{1 \leq i, j \leq n}$, researchers have traditionally used variants of the well-known gravity and radiation laws [9]. Gravity laws are based on the assumption that the number of trips T_{ij} between two locations i and j is related to the total number of commuters m_i leaving zone i , the total number of commuters n_j working in zone j , and decays directly as a function of the distance d_{ij} between the zones [6]. The importance of the distance in predicting the probabilities is typically controlled by parameters α , β , and/or γ .

Radiation laws, on the other hand, are based on the assumption that the number of trips T_{ij} between two locations i and j depends on the total number of commuters m_i leaving zone i , the total number of commuters n_j working in zone j , and the number of intervening opportunities s_{ij} between the two zones [7]. In the commuting literature, s_{ij} is typically defined as the total number of commuters working in all zones whose centroid falls in the circle centered at i with radius d_{ij} (not including zones i or j) [7]. In some forms of this law, a parameter β is introduced to control the effect of the number of intervening opportunities between the home and work zones. Table 1 provides equations for the traditional gravity and radiation laws chosen in this study.

The XGBoost model is a subset of a broader class of models, called machine learning models, that use a set of known input and output data to "learn" a model that can then be given new input data to estimate unknown output data [1]. In the case of commuter flow modeling, one might use a set of known input variables m_i , n_j , d_{ij} , s_{ij} , and known output

■ **Table 1** Traditional commuting laws.

Law	Equation
Gravity with exponential law	$\tilde{p}_{ij} \propto m_i n_j e^{-\beta d_{ij}}$
Extended radiation law	$\tilde{p}_{ij} \propto \frac{[(m_i + n_j + s_{ij})^\beta - (m_i + s_{ij})^\beta](m_i^\beta + 1)}{[(m_i + s_{ij})^\beta + 1][(m_i + n_j + s_{ij})^\beta + 1]}$

variables T_{ij} , to learn the structure of a machine learning model that can then take in new values of m_i , n_j , d_{ij} and s_{ij} , to estimate unknown values of T_{ij} . The XGBoost model is well known for winning several machine learning competitions and depends on three primary parameters commonly referred to as the maximum tree depth (r), number of estimators (e), and learning rate (k) [4].

3 Methodology

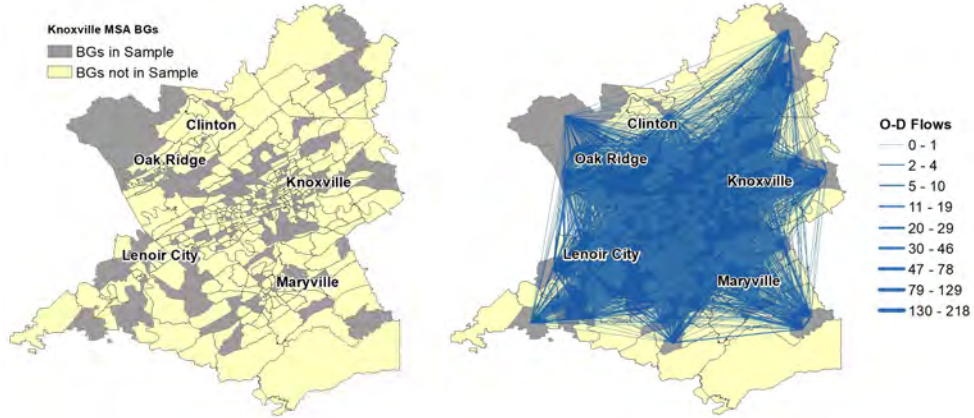
To determine whether the XGBoost model might add value to the field of commuter flow prediction, we compare and discuss the performance of a standard gravity and radiation model with the XGBoost machine learning algorithm for predicting a subset of home to work commuter flows within the Knoxville Metropolitan Statistical Area (MSA). From this point forward, we refer to the home location as the origin location and the work location as the destination location. The following subsections discuss the specific gravity, radiation, and XGBoost models chosen, as well as the data, study area, evaluation metrics, and experimental setup, in greater detail.

3.1 Models

In this study, we compare the performances of the gravity model based on an exponential distance decay function, the radiation model based on the extended radiation law, and a standard implementation of the XGBoost model. Table 1 provides the equations for both the gravity and radiation laws underlying the gravity and radiation models selected. Additionally, for both the gravity and radiation models, we ensure that the number of workers in each destination zone j is preserved by simulating all $(\tilde{T}_{ij})_{1 \leq i, j \leq n}$ from the multinomial distribution $\mathcal{M}(n_j, (\frac{\tilde{p}_{ij}}{\sum_{k=1}^n \tilde{p}_{kj}})_{1 \leq i, j \leq n})$. From this point forward, whenever we use the terms gravity or radiation model, we are referring specifically to the gravity and radiation models chosen in this study. The exponential distance decay function and extended radiation model were chosen because of their decent performance in a recent study conducted by [7]. The standard XGBoost model was chosen because of its flexibility and proven track record as the winner of several machine learning competitions [4].

3.2 Data and Study Area

We use each of the three models to predict commuting flows reported in a Census dataset called the 2010 Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics (LODES) [3]. The 2010 LODES dataset is a partially synthetic dataset that provides residential, workplace, and origin to destination commuter flow totals for a variety of U.S. Census-defined regions. We focus our study on estimating commuting flows between origin and destination Census block groups. Additionally, we focus our study on a subset of origin and destination block groups in the Knoxville MSA. More specifically, we consider all origin and destination block group pairs within a random sample of 120 block groups in



■ **Figure 1** The spatial boundaries for all 2010 Knoxville block groups (bgs), the subset of sampled bgs in the study area, and all origin-destination (o-d) commuting flows between the sampled block groups.

■ **Table 2** Evaluation metrics.

Metric	Equation
Common Part of Commuters (<i>CPC</i>)	$CPC(T, \tilde{T}) = \frac{2 \sum_{i,j=1}^n \min(T_{ij}, \tilde{T}_{ij})}{\sum_{i,j=1}^n T_{ij} + \sum_{i,j=1}^n \tilde{T}_{ij}}$
Common Part of Links (<i>CPL</i>)	$CPL(T, \tilde{T}) = \frac{2 \sum_{i,j=1}^n (\mathbb{1}_{T_{ij} > 0} \cdot \mathbb{1}_{\tilde{T}_{ij} > 0})}{\sum_{i,j=1}^n \mathbb{1}_{T_{ij} > 0} + \sum_{i,j=1}^n \mathbb{1}_{\tilde{T}_{ij} > 0}}$
Root Mean Squared Error (<i>RMSE</i>)	$RMSE(T, \tilde{T}) = \sqrt{\frac{1}{n} \sum_{i,j=1}^n (T_{ij} - \tilde{T}_{ij})^2}$

the Knoxville MSA. In total, there are $n = 14,280$ block group pairs within this subset, and $N = 15,288$ commuters who travel these routes. Figure 1 provides a visual map of the study area and data.

We use the LODES dataset to determine m_i , n_j , and T_{ij} , and another dataset, called the 2010 U.S. Census Block Group Shapefiles [2], to obtain the distances d_{ij} and intervening opportunities metrics s_{ij} for all origin block groups i and destination block groups j in the study area. Whenever we calculate a distance for a set of locations, we use the haversine formula to determine the great-circle distance between them.

3.3 Evaluation Metrics

To evaluate how well each of the models perform, we use three metrics commonly used in the commuting modeling literature. The first two, known as the Common Part of Commuters (*CPC*) and Common Part of Links (*CPL*) metrics, measure the similarity between the true commuting flow network and a predicted network. The third metric, known as the Root Mean Squared Error (*RMSE*), measures the prediction accuracy (how similar the true flow counts are to the predicted flow counts). Table 2 provides the equations for each metric.

3.4 Experimental Setup

To select the optimal hyperparameters and then compare the winning models, we split our data into training, validation, and testing sets via nested cross validation. More specifically,

we first split our data into 10 unique training and testing set pairs via 10-fold cross validation. We refer to each of these training/testing set pairs as outer folds. We then further split the training sets of each outer fold into 10 more unique training and validation sets via a second round of 10-fold cross validation.

For our gravity and radiation models, we choose the optimal $\beta \in [0, 0.1, \dots, 1]$ for each outer fold by first simulating one possible \tilde{T}_{ij} from the models corresponding to each β on the training sets of each inner fold. We then select the β that corresponds to the model with the highest average *CPC* score over all inner folds. Once the optimal β s are selected for each outer fold, we use the winning models to compute one possible \tilde{T}_{ij} on the testing sets of each outer fold.

For the XGBoost model, we choose the optimal maximum tree depth r , number of estimators e , and learning rate k , by first using a randomized grid search to simulate 100 random samples (r, e, k) from the Cartesian product of $r \in [2, 3, \dots, 7]$, $e \in [25, 26, \dots, 275]$, and $k \in [0.1, 0.2, \dots, 0.5]$. We then find the optimal combination (r, e, k) for each outer fold by first simulating \tilde{T}_{ij} from the models corresponding to each of the 100 parameter combinations (r, e, k) on the training sets of each inner fold, and then selecting the (r, e, k) set that corresponds to the model with the highest average *CPC* score over all inner folds. Once the optimal parameter combinations are selected for each outer fold, we next use the optimal models to compute \tilde{T}_{ij} on the testing sets of each outer fold. During each simulation, we round the output data \tilde{T}_{ij} to the nearest non-negative integer.

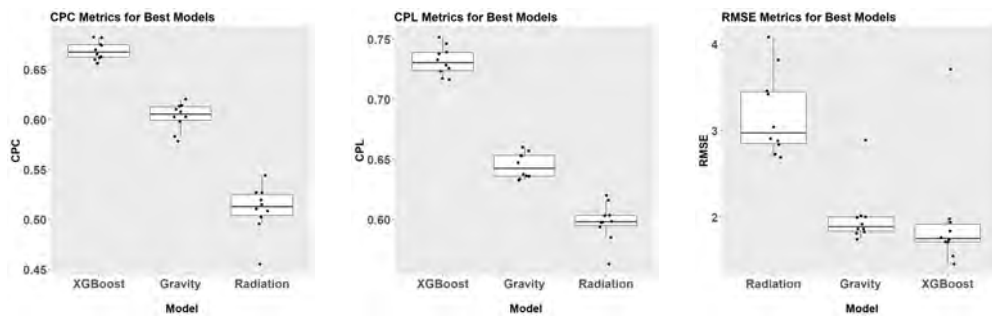
4 Results

Figure 2 shows box plots of the *CPC*, *CPL*, and *RMSE* scores produced by each model type for all outer folds. In addition, each of these figures shows the actual values for each metric over each outer fold, randomly adjusted, or "jittered", on the y -axis to prevent overlap. More specifically, we see in Figure 2 that all *CPC* and *CPL* scores produced by the XGBoost model are higher than all *CPC* and *CPL* scores produced by the gravity model, which are in turn higher than all of the *CPC* and *CPL* scores produced by the radiation model. Since all scores, rather than just all median scores, are higher in the XGBoost model than both other models, we are confident that the XGBoost model outperforms the gravity and radiation models on the *CPC* and *CPL* metrics. On the other hand, though we see that the median *RMSE* of the XGBoost model is also the best, or lowest, median *RMSE* among all three models, not all of the XGBoost model's *RMSE* scores are lower than scores coming from the other models. For example, the XGBoost *RMSE* score from one of the 10 testing sets is worse than all of the gravity model's *RMSE* scores and worse than eight of the radiation model's *RMSE* scores. This suggests that, though the XGBoost model produces a network with more similar structure to the ground truth network, it may also produce flow counts that are very far apart from one another.

5 Conclusion and Future Work

In this paper, we compared and discussed the performance of a standard gravity and radiation model with the XGBoost machine learning algorithm for predicting origin/destination commuter flows for a subset of block groups in the Knoxville MSA. We parameterized each model using two well known Census datasets and then evaluated and compared each model using the *CPC*, *CPL* and *RMSE* metrics.

Overall, we found that the XGBoost model far outperformed the gravity and radiation models on both the *CPC* and *CPL* metrics, indicating that it was able to re-create the



■ **Figure 2** Box plots and horizontally jittered *CPC*, *CPL* and *RMSE* scores for the best performing models on each testing set.

original network better than the traditional models. However, we also discovered that the XGBoost model sometimes led to higher *RMSE* scores than both the gravity and radiation models, despite having the lowest median *RMSE* value. This may indicate that, given certain training/testing set combinations, the XGBoost model has the potential to produce estimates that are very far off from the ground truth flows. Thus, despite the fact that the XGBoost model re-creates the overall flows better than the gravity and radiation models, certain (though likely rare) links may have larger errors.

Though this study does indicate that the XGBoost model likely adds value to the field of commuter flow prediction, there are a few limitations and opportunities worth noting. For example, in a follow-up study it may be worth comparing more complex commuting models with the XGBoost model to determine if it still performs better. Additionally, one might want to add other machine learning models to the framework to determine if they add additional value on top of the XGBoost model. Furthermore, there may be other non-conventional input variables worth considering in the machine learning models that may further improve their performances.

6 Copyright

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (<http://energy.gov/downloads/doe-public-access-plan>).

References

- 1 Ethem Alpaydin. *Introduction to machine learning*. MIT Press, 2014.
- 2 United States Census Bureau. Block group shapefiles for Tennessee [data file], 2010. URL: <https://www.census.gov/geo/maps-data/>.
- 3 United States Census Bureau. LEHD Origin-destination employment statistics [dataset], 2010. URL: <https://lehd.ces.census.gov/data/>.

- 4 Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.
- 5 Andrea De Montis, Marc Barthélemy, Alessandro Chessa, and Alessandro Vespignani. The structure of interurban traffic: a weighted network analysis. *Environment and Planning B: Planning and Design*, 34(5):905–924, 2007.
- 6 Sven Erlander and Neil F Stewart. *The gravity model in transportation analysis: theory and extensions*, volume 3. VSP, 1990.
- 7 Maxime Lenormand, Aleix Bassolas, and José J Ramasco. Systematic comparison of trip distribution laws and models. *Journal of Transport Geography*, 51:158–169, 2016.
- 8 Celik H Murat. Sample size needed for calibrating trip distribution and behavior of the gravity model. *Journal of Transport Geography*, 18(1):183–190, 2010.
- 9 Caleb Robinson and Bistra Dilkina. A machine learning approach to modeling human migration. *arXiv preprint arXiv:1711.05462*, 2017.
- 10 Jan Rouwendal and Peter Nijkamp. Living in two worlds: a review of home-to-work decisions. *Growth and Change*, 35(3):287–303, 2004.

Modeling Road Traffic Takes Time

Kamaldeep Singh Oberoi

Normandie Univ, UNIROUEN, UNIHAVRE, INSA Rouen, LITIS, 76000 Rouen, France
kamaldeep-singh.oberoi1@etu.univ-rouen.fr

Géraldine Del Mondo

Normandie Univ, INSA Rouen, UNIROUEN, UNIHAVRE, LITIS, 76000 Rouen, France
geraldine.del_mondo@insa-rouen.fr

Yohan Dupuis

CEREMA, 76121 Le Grand-Quevilly, France
yohan.dupuis@cerema.fr

Pascal Vasseur

Normandie Univ, UNIROUEN, UNIHAVRE, INSA Rouen, LITIS, 76000 Rouen, France
pascal.vasseur@univ-rouen.fr

Abstract

To model dynamic road traffic environment, it is imperative to integrate spatial and temporal knowledge about its evolution into a single model. This paper introduces temporal dimension which provides a method to reason about time-varying spatial information in a spatio-temporal graph-based model. Two types of evolution, topological and attributed, of time-varying graph (TVG) are considered which require the time domain to be discrete and/or continuous, and the TVG proposed includes time-varying node/edge presence and labeling functions. Theoretical concepts presented in this paper will guide us through the process of application development in future.

2012 ACM Subject Classification Information systems → Spatial-temporal systems, Computing methodologies → Modeling methodologies, Mathematics of computing → Graph theory

Keywords and phrases Qualitative Spatio-temporal Model, Time Varying Graph, Road Traffic, Intelligent Transportation Systems

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.52

Category Short Paper

Acknowledgements This work takes part in the DAISI project. This project has been funded with the support from the European Union with the European Regional Development Fund (ERDF) and from the Regional Council of Normandy.

1 Introduction

Road traffic evolves in space-time. This evolution is made explicit by time-varying spatial relations between different objects like vehicles, pedestrians, buildings etc. which directly affect the flow of traffic in an urban environment. To extract useful information from the movement of traffic, we need to model it in a reasoning system. To this effect, we proposed a spatial model in [7]. It includes different physical objects present in an urban area and, using quantitative information, aims to extract qualitative spatial knowledge which can enhance the robustness of Advanced Driver Assistance Systems (ADAS) currently in use. The model uses



© Kamaldeep S. Oberoi, Géraldine Del Mondo, Yohan Dupuis, and Pascal Vasseur;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 52; pp. 52:1–52:7

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

graphs as data structure which abstract the real world information. Furthermore, dynamic phenomenon can be described using time-varying graphs.

In this paper, we propose a time formalization which provides a temporal dimension to the spatial model described in [7]. We define a time-varying graph (TVG) which models the change in its structure as well as in its node/edge attributes. The objective of this paper is to propose a theoretical formalization of TVG and link it with spatial graph proposed in our previous work.

The paper is organized as follows. Section 2 presents the related work. In Section 3, we describe, in brief, different spatial graphs. Then, in Section 4, we propose the formalization of time-varying graph and Section 5 concludes the paper along with future work.

2 Related Work

In this section, we will first mention some research related to modeling of road traffic and urban environment in Intelligent Transport Systems (ITS) domain. Then we will mention some techniques for including time and modeling time-varying graphs present in the literature.

Although there is a lot of research which exists for modeling road traffic environment for ITS applications, the one we would like to focus on and compare with our work is Local Dynamic Map (LDM) [4], a multi-level database, which has been standardized in Europe. It has been developed for cooperative systems and uses a four layered model to store data. Although we plan to use similar database architecture as in LDM, the main contribution of our model is data abstraction using graphs, which can initiate the use of graph algorithms to comprehend the evolution of road traffic.

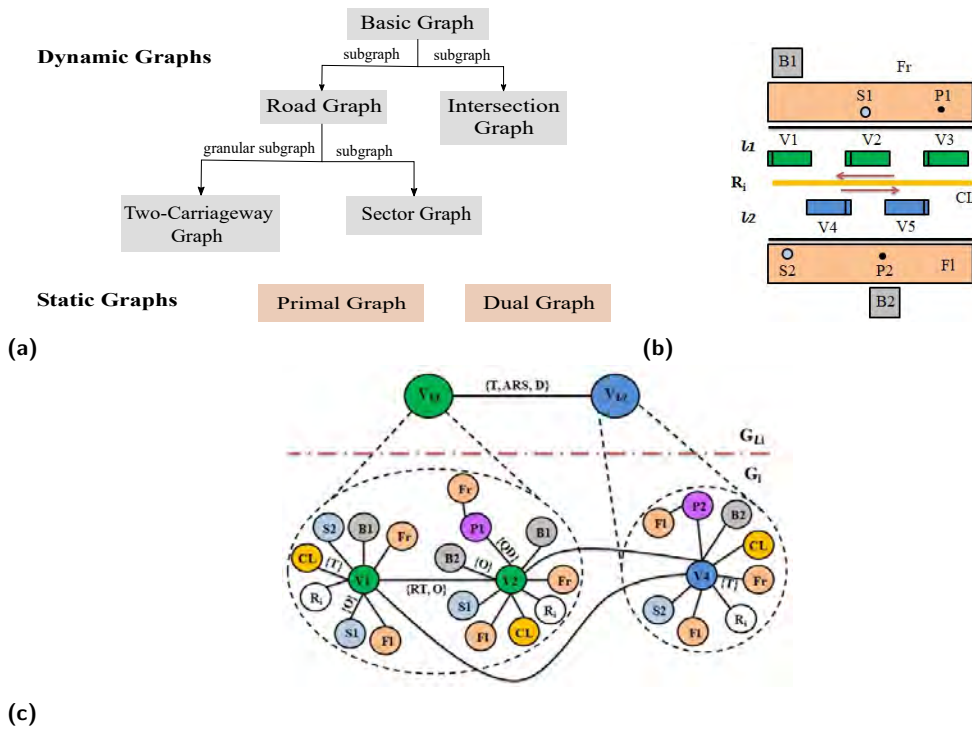
Time modeling has applications in various domains. It can be modeled using intervals or instants [5]. In our model, we consider time to consist of both. A lot of researchers have proposed different methods for time modeling and a survey of such methods is given in [8].

A time-varying phenomenon (like road traffic) can be modeled using time-varying graphs (TVG). Some models for TVGs are described in [3] and [9]. However, the model proposed in this paper is motivated from the one described in [2], as it is suitable for highly dynamic networks and it uses continuous time domain.

3 Spatial Graph Model

The qualitative spatial model we proposed in [7] is based on graph theory in which different objects are represented as nodes and qualitative spatial relations between those objects are included as edges. The objective is to have a spatio-temporal model which can help to understand the dynamics of road traffic in an urban environment from the perspective of evolving spatial relations between static and/or non-static objects.

Figure 1a shows the hierarchy of spatial graphs which are derived from Basic Graph $G = (V, E)$ containing information about the urban environment at finest level of detail. V is the set of nodes which represent real-world objects and E is the set of edges which represent spatial relations between different nodes. From G , a Road Graph for each road segment in that environment is derived, which contains the spatial relations between objects present on that road segment. G also contains the relations between objects present at each intersection and hence an Intersection Graph (for each intersection) is derived from it. If a road segment is divided into bi-directional carriageways, Two-Carriageway Graph for that road segment is computed which gives information at a coarser level than Road Graph. If, on the other hand, it is divided into sectors, Sector Graph for each sector present on that road segment is



■ **Figure 1** (a) Hierarchy of graphs (b) An i th road segment divided into two carriageways (c) (Finer) Road graph (G_i) and (coarser) Two-carriageway graph (G_{Li}).

defined. Graphs which provide static information about the road network, primal and dual graphs, are also included in the model.

Road graph in Figure 1c includes different objects (vehicles, buildings, pedestrians, road markings, vertical structures, roadsides), present on the road segment in Figure 1b, as nodes with its edges representing spatial relations [7]. The nodes of two-carriageway graph represent groups of vehicles moving in opposite directions (for bi-directional road segment).

4 Time and Temporal Graph Model

Let us now add a temporal dimension to the proposed spatial model to theorize about dynamic aspects of the road traffic. In this section, we will first describe the structure of the temporal domain along with the temporal primitives considered. Then we will move onto the formalization of the temporal graph model, which is the main contribution of this paper.

4.1 Structure of Time

Before diving into the formalization of temporal graph, some characteristics of time [8] need to be clarified. We assume time to be linear, dense and positively unbounded $[0, \infty[$ (we use $[]$ notation to represent a closed interval and $] [$ for open). Consider a time domain (\mathbb{T}, \leq) which is a set of totally ordered time points with order relation \leq . Since it is dense, the domain of \mathbb{T} is $\mathbb{R}_{>0}$ (set of positive real numbers) and $\exists t_j \mid t_i < t_j < t_k, \forall t_i, t_j, t_k \in \mathbb{T}, i, j, k \in \mathbb{R}_{>0}$. That means that there is always a time point between two adjacent time points (digitization of such time points could give different results depending on temporal granularity considered). This time domain has two time primitives: instants and intervals, and we assume that intervals are

bounded by instants [1]. For clarification, term "time instants" is used to represent individual points on a discrete time line whereas term "time points" is used to represent individual points on a continuous time line. We define $\mathcal{T} \subset \mathbb{T}$ as the time during which our model is functional and we call it the lifetime of the model, which is bounded. We consider that \mathcal{T} can represent discrete as well as continuous time. In the former case, the domain of \mathcal{T} is $\mathbb{Z}_{>0}$ (set of positive integers) and it consists of discrete time instants and intervals. For continuous time, \mathcal{T} belongs to $\mathbb{R}_{>0}$ and consists of continuous time intervals. A closed non-zero duration time interval is given as the pair $[t_{start}, t_{end}] \mid t_{start} < t_{end}, t_{start}, t_{end} \in \mathcal{T}$, where t_{start} and t_{end} are zero duration instants which bind the interval.

4.2 Temporal Graph Model

Our model of time varying graph (or TVG) is motivated from [2]. We include node/edge presence functions and define labeling functions. We consider the evolution of the graph in terms of change in its structure (or topology), called "topological evolution", and change in the value of node/edge attributes given as labels, called "attributed evolution". We define a TVG as $\mathcal{G} = (V, E, \mathcal{T}, \rho_V(\mathcal{T}), \rho_E(\mathcal{T}), A_V(\mathcal{T}), A_E(\mathcal{T}))$ where V and E are the sets of nodes and edges, respectively, included in the spatial graph described in Section 3, $\mathcal{T} \subset \mathbb{T}$ is the lifetime of the model, $\rho_V : V \times \mathcal{T} \rightarrow \{0, 1\}$ is the time-varying node presence function, $\rho_E : E \times \mathcal{T} \rightarrow \{0, 1\}$ is the time-varying edge presence function, A_V is time-varying node labeling function and A_E is time-varying edge labeling function. For topological evolution, \mathcal{T} is discrete and for attributed evolution, it is continuous. In both types of evolution, the time at which the change occurs, is called characteristic date [2]. In topological evolution, the characteristic date is when a node or edge is added/removed in the graph. Similarly, in attributed evolution, attribute value changes at a characteristic date. Such dates defined within discrete/continuous lifetime \mathcal{T} provide an explicit way to model time when different kinds of changes occur.

4.2.1 Time-varying Node Labels

In our previous work [6], we described nine classes into which real world objects, present in an urban environment, can be classified. These objects, along with groups of vehicles belonging to each carriageway on a bidirectional road segment, are included as nodes in our model. For simplification and homogeneity, we consider that the nodes representing groups of vehicles belong to a separate class called "Group". Each (of now ten) class of nodes, given by the set of classes $C_V = \{c_1, c_2, \dots, c_{10}\}$, is assumed to have a unique set of attributes during \mathcal{T} , given by $K_{c_i} = \{\kappa_1, \kappa_2, \dots, \kappa_m\}$, where the number of attributes m varies for every $c_i \in C_V$, $1 \leq i \leq 10$. As in [10], we assign an attribute vector $[\kappa_1(v_i), \kappa_2(v_i), \dots, \kappa_m(v_i)]$ to i th node $v_i \in V_{c_j}$, $1 \leq j \leq 10$ with V_{c_j} being the set of nodes which belong to a class $c_j \in C_V$. An element of the attribute vector $\kappa_x(v_i)$, $1 \leq x \leq m$ is the value of the attribute κ_x for a node v_i . This attribute vector of a node v_i is considered to be the label for that node. Assume that classes for all nodes in \mathcal{G} and set of attributes K_{c_j} for all classes $c_j \in C_V$, $1 \leq j \leq 10$ are given as *a priori*. For a node $v_i \in V_{c_j}$, a node labeling function can be given as $A_{v_i}(v_i, K_{c_j}) = [\kappa_1(v_i) \ \kappa_2(v_i) \ \dots \ \kappa_b(v_i)]_{1 \times b}$ where $b = |K_{c_j}|$. Considering

all nodes in a class $c_j \in C_V$, a node labeling function for this class is written as

$$A_{V_{c_j}}(V_{c_j}, K_{c_j}) = \begin{bmatrix} \kappa_1(v_1) & \kappa_2(v_1) & \dots & \kappa_b(v_1) \\ \kappa_1(v_2) & \kappa_2(v_2) & \dots & \kappa_b(v_2) \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \kappa_1(v_a) & \kappa_2(v_a) & \dots & \kappa_b(v_a) \end{bmatrix}_{a \times b}$$

with $a = |V_{c_j}|$. When the time is included with $A_{V_{c_j}}$, we get a time-varying node labeling function $A_{V_{c_j}}(V_{c_j}, K_{c_j}, \mathcal{T})$ for a class $c_j \in C_V$. This function will give as output a 3-D vector with dimensions $|V_{c_j}| \times |K_{c_j}| \times |\mathcal{T}|$. Since in case of attributed evolution \mathcal{T} is continuous, the value of $|\mathcal{T}|$ will change according to the temporal granularity considered. Finally, we define a time-varying node labeling function $A_V(\mathcal{T})$ which gives as output the value of all attributes for the nodes present in \mathcal{G} during \mathcal{T} in the form of 4-D vector, where fourth dimension has ten entries, one for each class.

4.2.2 Time-varying Edge Labels

To define the time varying edge labeling function, we classify an edge in TVG on the basis of the classes of nodes which are its end points. For example, an edge between a vehicle and a building is classified as Vehicle-Building edge. In [6], sets of relations for thirteen different classes of edges are proposed. Since we have defined class "Group" of nodes in the previous section, we classify edge between two group nodes into Group-Group class, and hence fourteen classes of edges are possible. The value of relations on these edges, given by a corresponding attribute vector, acts as the edge label. Given the set of node classes $C_V = \{c_1, c_2, \dots, c_{10}\}$, set of edge classes can be written as $C_E = \{c_x c_y | c_x, c_y \in C_V, 1 \leq x \leq 10, 1 \leq y \leq 10\}$ and has fourteen elements. It is possible to have a class of type $c_x c_x \in C_E$ provided $c_x = Vehicle \vee c_x = Group \vee c_x = Intersection \vee c_x = Roadsegment$ since relations between two vehicles, groups, intersections and road segments are allowed. Each class $c_p c_q \in C_E, 1 \leq p \leq 10, 1 \leq q \leq 10$ is made up of two classes c_p and c_q of nodes, between which an edge exists belonging to $c_p c_q$. For every $c_p c_q \in C_E, R_{pq} = \{r_1, r_2, \dots, r_n\}$ is the set of relations corresponding to that class, where n depends on the class [6]. Given an edge $e_{vv'} \in E, v, v' \in V$, it belongs to class $c_p c_q \in C_E \iff v \in V_{c_p} \wedge v' \in V_{c_q}, V_{c_p}, V_{c_q} \subset V$. An edge labeling function for $e_{vv'}$ given R_{pq} is written as $A_{e_{vv'}}(e_{vv'}, R_{pq}) = [r_1(e_{vv'}) \ r_2(e_{vv'}) \ \dots \ r_d(e_{vv'})]_{1 \times d}$ where $d = |R_{pq}|$. The attribute vector, given as output, contains the values of different relations which exist on edge $e_{vv'}$. Let $E_{c_p}^{c_q}$ represent the set of edges which belong to class $c_p c_q \in C_E$. Then edge labeling function for $E_{c_p}^{c_q}$ gives a 2D vector as output

$$A_{E_{c_p}^{c_q}}(E_{c_p}^{c_q}, R_{pq}) = \begin{bmatrix} r_1(e_1) & r_2(e_1) & \dots & r_d(e_1) \\ r_1(e_2) & r_2(e_2) & \dots & r_d(e_2) \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ r_1(e_c) & r_2(e_c) & \dots & r_d(e_c) \end{bmatrix}_{c \times d}$$

where $c = |E_{c_p}^{c_q}|$. Over \mathcal{T} , time varying edge labeling function $A_{E_{c_p}^{c_q}}(E_{c_p}^{c_q}, R_{pq}, \mathcal{T})$ for a class $c_p c_q \in C_E$ gives as output the 3-D vector with dimensions $|E_{c_p}^{c_q}| \times |R_{pq}| \times |\mathcal{T}|$. Considering all classes of edges in E , time varying labeling function $A_E(\mathcal{T})$ has 4-D vector as output with fourth dimension related to the total number of edge classes (fourteen).

4.2.3 Graph Evolution

As mentioned before, the evolution of a TVG can be topological or attributed. The change in the graph topology is considered to be discrete in our model. That means, the addition/removal of a node/edge is instantaneous. In addition, we ignore the change in attribute values of nodes/edges and only focus on their presence/absence. Hence, the definition of TVG can be modified to $\mathcal{G}_T = (V, E, \mathcal{T}, \rho_V(\mathcal{T}), \rho_E(\mathcal{T}))$. However, in case of attributed evolution, the TVG is formalized as $\mathcal{G}_A = (V, E, \mathcal{T}, A_V(\mathcal{T}), A_E(\mathcal{T}))$ with \mathcal{T} representing continuous time, which can be discretized if variation in an attribute value is instantaneous. In this case, change in graph topology is ignored. It is noteworthy, that both types of evolution happen simultaneously and hence, the lifetime of the system is the same in both cases.

5 Conclusion and Future Work

In this paper, we proposed the formalization of a time-varying graph (TVG) which provides the temporal dimension to the spatio-temporal graph-based model we are developing, to understand the dynamics of road traffic in a given urban environment. Two types of graph evolution are considered and node/edge presence and labeling functions are defined. Due to limited number of pages, we skip the description of related concept of underlying graph, a static graph which relates spatial and time-varying graphs, and the notion of defining different point of views for visualizing the evolution of TVG. The next step for our work is to define the conceptual framework of the system and implement the ideas proposed. To do this, we first need to compare the existing spatio-temporal data models (conceptual and physical) and adapt one to our needs. The required real-world traffic data is collected by CEREMA, Rouen (France). Our long-term goal is to develop graph algorithms to compute and reason about patterns in evolving road traffic.

References

- 1 James F Allen and Patrick J Hayes. A common-sense theory of time. In *Proceedings of the 9th international joint conference on Artificial intelligence-Volume 1*, pages 528–531. Morgan Kaufmann Publishers Inc., 1985.
- 2 Arnaud Casteigts, Paola Flocchini, Walter Quattrociocchi, and Nicola Santoro. Time-varying graphs and dynamic networks. *International Journal of Parallel, Emergent and Distributed Systems*, 27(5):387–408, 2012.
- 3 Benoit Costes, Julien Perret, Bénédicte Bucher, and Maurizio Gribaudo. An aggregated graph to qualify historical spatial networks using temporal patterns detection. In *18th AGILE International Conference on Geographic Information Science*, 2015.
- 4 ETSI. Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Local Dynamic Map (LDM); Rationale for and Guidance on Standardization. Technical Report ETSI TR 102 863 V1.1.1, ETSI, 2011.
- 5 Antony Galton. A critical examination of allen’s theory of action and time. *Artificial intelligence*, 42(2-3):159–188, 1990.
- 6 Kamaldeep S Oberoi, Géraldine Del Mondo, Yohan Dupuis, and Pascal Vasseur. Spatial Modeling of Urban Road Traffic Using Graph Theory. In *Spatial Analysis and GEOmatics (SAGEO)*, pages 264–277, 2017. URL: <https://hal.archives-ouvertes.fr/hal-01643369>.
- 7 Kamaldeep S Oberoi, Géraldine Del Mondo, Yohan Dupuis, and Pascal Vasseur. Towards a qualitative spatial model for road traffic in urban environment. In *2017 IEEE 20th*

- International Conference on Intelligent Transportation Systems (ITSC)*, pages 1724–1729, Oct 2017. doi:10.1109/ITSC.2017.8317644.
- 8 Lluís Vila. A survey on temporal reasoning in artificial intelligence. *Ai Communications*, 7(1):4–28, 1994.
 - 9 Klaus Wehmuth, Artur Ziviani, and Eric Fleury. A unifying model for representing time-varying graphs. In *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*, pages 1–10. IEEE, 2015.
 - 10 Yang Zhou, Hong Cheng, and Jeffrey Xu Yu. Graph clustering based on structural/attribute similarities. *Proceedings of the VLDB Endowment*, 2(1):718–729, 2009.

Diversity in Spatial Language Within Communities: The Interplay of Culture, Language and Landscape in Representations of Space

Bill Palmer

University of Newcastle, Australia
bill.palmer@newcastle.edu.au

Alice Gaby

Monash University, Australia
alice.gaby@monash.edu

Jonathon Lum

Monash University, Australia
lum.jonathon@gmail.com

Jonathan Schlossberg

University of Newcastle, Australia
schlossberg.jonathan@gmail.com

Abstract

Significant diversity exists in the way languages structure spatial reference, and this has been shown to correlate with diversity in non-linguistic spatial behaviour. However, most research in spatial language has focused on diversity between languages: on which spatial referential strategies are represented in the grammar, and to a lesser extent which of these strategies are preferred overall in a given language. However, comparing languages as a whole and treating each language as a single data point provides a very partial picture of linguistic spatial behaviour, failing to recognise the very significant diversity that exists *within* languages, a largely under-investigated but now emerging field of research. This paper focuses on language-internal diversity, and on the central role of a range of sociocultural and demographic factors that intervene in the relationship between humans, languages, and the physical environments in which communities live.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases spatial language, Frame of Reference, landscape, sociotopography

Digital Object Identifier 10.4230/LIPICs.GIScience.2018.53

Category Short Paper

Funding This research was funded by Australian Research Council Discovery Project DP120102701. We gratefully acknowledge this support.

Acknowledgements We are grateful for the comments of three anonymous reviewers. Any errors remain ours. We thank our language consultants in the Marshall Islands and the Maldives.



© Bill Palmer, Alice Gaby, Jonathon Lum, and Jonathan Schlossberg;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 53; pp. 53:1–53:8

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Diversity in the way languages structure spatial reference has been amply demonstrated, and has been shown to correlate with diversity in spatial behaviour in other, non-linguistic, cognitive modalities (navigation and wayfinding; memory recall and memory recognition; inferential reasoning; gesture; etc.). Some theories have argued for a primary role of language in shaping conceptual representations of space [9, 12, 20]. Others have focused on the role of the environment in which communities live and languages are spoken in motivating spatial representations that are manifest across modalities, including language [15].

Here we present findings on diversity in preferred Frame of Reference in linguistic expressions of spatial relations. In this context, a Frame of Reference (FoR) is a conceptual strategy for locating an object (“figure”) or path in relation to another object (“ground”). This is done by assigning an asymmetry to a scene so that a path or a search domain in which the figure can be found can be projected off the ground object on the basis of a coordinate system fixed to a particular “anchor”. Different FoRs are different strategies for assigning this asymmetry, involving different anchors, and therefore represent different types of coordinate systems. Three FoRs are established: intrinsic, relative, and absolute [8, 9, 12, 15] (see Figure 1). In the intrinsic FoR the coordinate system is anchored in the ground object on the basis of a perceived intrinsic asymmetry in the facets of that object itself (e.g., *in front of the chair* – the search domain/path is projected off a perceived intrinsic ‘front’ of the ground chair, itself the anchor). In the relative and absolute FoRs the anchor is external to the Figure_Ground array. In the relative FoR the coordinate system is anchored in the location of a viewpoint (e.g., *in front of* [i.e., on the viewer’s side of] *the post* – the search domain/path is projected off the facet of the ground post facing the viewpoint anchor). Absolute FoR invokes a set of external coordinates imposed on the scene (e.g., *west of the house* – the search domain/path is projected off the facet of the ground house facing west in an external cardinal coordinate system, with the anchor in those external coordinates).

The two externally-anchored FoRs and a number of other referential strategies for expressing spatial relations can also be divided into those which are egocentric, such as those invoking participants in the speech event as landmarks (e.g., *on my side of the post*) or through the relative FoR (e.g., *in front of the post*); and those which are geocentric, invoking features of the external world, either through the absolute FoR (e.g., *seaward from the village*), or through reference to landmarks (e.g., *towards the sea from the village*) (e.g. [4, 14, 17]).

2 Diversity across languages

Most research in spatial language to date has focused on diversity between languages. This has primarily focused on which referential strategies are represented in the grammars of individual languages [9, 10, 12, 20]. For example, in terms of FoR, some languages provide specialised grammatical means of expressing spatial relations in the relative FoR, and others do not. To a lesser extent research has focused on which of these strategies are preferred overall out of the referential strategies available in individual languages [12]. For example, Mopan (Mayan, Belize) has been characterized as employing intrinsic and absolute (geocentric) FoR, but not relative (egocentric) FoR, with intrinsic preferred and absolute only available in restricted contexts [12]. Tamil (Dravidian, India), on the other hand, has been characterized as allowing intrinsic, relative and absolute, but dispreferring intrinsic [12]. However, considering each language as a whole fails to recognise the very significant diversity that exists *within* languages, a largely under-investigated but now emerging field of research.

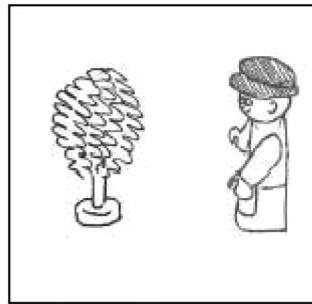
3 Diversity within language communities

A handful of recent studies have now shown that diversity among speakers within a language can be considerable, and that individual language communities are far from homogeneous [2, 17]. Language-internal diversity based on environment has previously been observed in a preference for relative FoR among urban communities and absolute FoR among rural communities [12], for example between urban and rural Tamils [18, 19], and on the basis of scale (table-top space versus navigational scale). However, recent studies have found significant variation on the basis of individual demographic factors such as age, gender and occupation, and community-wide cultural practices such as dominant subsistence mode.

Some language-internal diversity may correlate with different patterns of sociocultural interaction with the environment of the language locus. For example, in one Ancash (Quechuan, Peru) community in the Andes, individuals who work in the highlands as herders show significantly higher rates of geocentric reference than those who do not: “both highland pastoralism and the use of the Absolute FoR draw on a similar cognitive ability to keep track of one’s position among various landmarks in a fixed coordinate system” [22]. Gender is another factor that may correlate with variation in spatial reference. Mopan is cited above as preferring intrinsic FoR with absolute used in restricted contexts [12], but this language-level generalisation oversimplifies the situation and masks patterns of behaviour. For example, cardinal directions are used across the board more often by Mopan men, who work in the fields, than by Mopan women, who work in the home or in the village [5]. Similarly, among Yucatec Mayans (Mexico), men but not women use cardinal direction terms, reflecting occupational biases and cultural practices specific to men, particular in garden work [1, 3, 7]. Other factors such as age or education also play a role. In Dhivehi (Indo-Aryan, Maldives) older speakers, men, and less well educated individuals, who were more likely to have worked outdoors or on the sea, were more likely to use geocentric references than younger speakers, women, and better educated speakers, who were more likely to have always worked indoors [11]. Sometimes community-wide cultural practices play a role. On one Maldivian atoll, speakers living on islands where fishing was the dominant subsistence mode used geocentric expressions at significantly higher rates, independent of the occupation of individual community members, than speakers on other islands on the same atoll where indoor work dominated, who favoured egocentric strategies (see below) [11]. Other studies show inter-generational change. In Australia’s Indigenous Gurindji community, older speakers use absolute FoR more frequently than younger speakers, apparently correlating with a shift to Gurindji Kriol and Aboriginal English, perhaps also related to schooling and other changes to way of life [13].

4 Diversity within Marshallese and Dhivehi

Quantitative analysis of a corpus of data gathered in a recent study of language-internal and language-external variation in spatial reference in two atoll-based languages presents a picture of systematic and partially parallel variation within each language community [11, 16, 17, 21]. One of the first systematic large-scale investigations of language-internal variation in spatial behaviour, this study was conducted among speakers of Marshallese (Austronesian, Marshall Islands) and Dhivehi (Indo-Aryan, Maldives), in order to test the Topographic Correspondence Hypothesis (TCH) [15]. TCH hypothesises a correlation between the features of linguistic spatial referential systems and features of the topography of the environment in which a language is spoken. The results of the atoll study partially support TCH, but demonstrate that language-internal variation exists correlating with a



■ **Figure 1** Sample “Man and Tree” card [6].

range of sociocultural factors beyond the scope of TCH, revealing the limitations of the hypothesis’s focus on environment alone.

Data in this study was elicited using an identical set of formal experimental task-based methodologies, some established, some developed for the project, in each of a range of diverse communities in a range of environments in both languages. A total of 96 participants for Marshallese and 118 for Dhivehi were involved, making this the largest such study by a considerable margin. Data presented below are from the results of a “Man and Tree” elicitation task [23]. In this task, one participant, a ‘director’, selects a card from a set of cards bearing images of a toy man and a toy tree in various configurations, and describes the configuration so a second participant, a ‘matcher’, who selects the corresponding card from their own set, yielding data heavy in spatial reference.

The tree is *in front of* the man (intrinsic FoR).

The tree is *to the left of* the man (relative FoR).

The tree is *west of* the man (absolute FoR).

Quantitative analyses of task results revealed not coarse-grained FoR choice (absolute versus relative, etc.), but preferences among a wide range of referential strategies offered by each language, some involving specialised grammatical constructions, some not. In other words, each language provides its speakers with a range of spatial referential strategies, and speakers vary on which strategies they prefer, and how strong those preferences are. Patterns of strategy preference emerged based on a range of factors. Some representative findings are presented here. Some patterns of strategy preference correlated simply with overall language community regardless of location or individual demography. For example intrinsic FoR accounted for 31% of spatial descriptions offered by Dhivehi participants in a Man & Tree task, but only 10% of descriptions offered by Marshallese participants. However, environment also played a role. For example, among externally-anchored Dhivehi Man & Tree location descriptions, preference for egocentric strategies correlated with degree of urbanisation: egocentric strategies account for 88% of descriptions in the densely urban Maldivian capital Malé, 77% in less urban Addu atoll, and an average of 43% in rural Laamu atoll.

Community-wide practices were also a factor. On Laamu atoll the dominant subsistence mode on some islands is fishing, but on others it is indoor work and small scale farming. Quantitative analyses found 79% of all externally anchored Man & Tree descriptions were geocentric on islands where the dominant subsistence mode is fishing, but only 39% on islands where indoor work and small scale farming dominate, independent of the individual occupation of each participant (see Figure 2). Moreover, individual demographic factors were also important. Laamu participants aged 17-34 produced 44% geocentric descriptions, while the figure for ages 35-49 was 67%, and ages 50-70 was 77%. Cross-cutting that, among

fishermen and sailors, 93% of Man & Tree descriptions were geocentric, but among indoor workers only 55% were. Variation was also observed on the basis of education, literacy, and bilingualism [11, 21]. Finally, linguistic resources and language use were factors: topographic features and cardinals were invoked in equal numbers in Marshallese, but references invoking topographic features were almost entirely absent in Dhivehi, correlating with the encoding of key topographic features in specialised terms in high frequency constructions in Marshallese but not Dhivehi.

5 Sociotopography

Findings such as those outlined in sections 3 and 4 provide strong support for the Sociotopographic Model (STM) [17], an attempt to model the interaction of environmental, sociocultural and linguistic factors in spatial referential systems. Major environmental features are salient to humans and play a role in conceptual representations of space that then interact with linguistic spatial expressions, consistent with the Topographic Correspondence Hypothesis. However, sociocultural factors, as well as affordances of the environment, mediate in the relationship between humans and landscape, a fact that cannot be accounted for within TCH but is captured by STM. In addition, the linguistic resources of the language itself contribute to nonlinguistic representations of space, mediated by language use. Each of these interactions is bidirectional. For example, topographic features and affordances of the environment shape human sociocultural interaction with that environment, while that interaction itself in turn plays a role in modifying and developing the environment through the built environment [17]. Sociotopography is defined in terms of: the natural environment (broadly construed, including topography, path of the sun, prevailing winds etc.); the built environment; and affordances of and sociocultural interaction and associations with the natural and built environment. It is culturally ‘constructed’: humans modify their environment; and conceptualise existing topography in terms of uses, associations and meanings attached to it. Consequently, elements of the landscape that are not attended to by some individuals and by some communities may be prominent to others. A sample implementation of the model is presented in Figure 2.

6 Conclusion

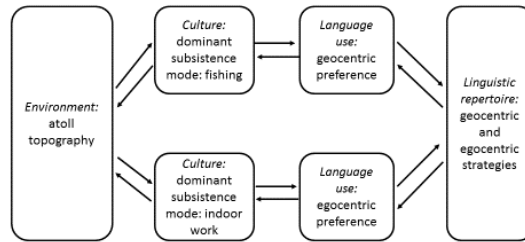
A tendency of much previous research to focus on a language’s overall spatial system rather than individual choices among available spatial referential strategies within a language has led to failed attempts to attribute a determining role to a single factor: to language, or to landscape, or to culture. Our findings demonstrate that all these factors and more play a role. Attending only to strategy choice in languages as a whole obscures patterns that reveal the complex interplay of factors at work in shaping conceptual representations of space: patterns reflecting the nature of the environment, the degree and nature of engagement with the environment, cultural associations placed on the environment, individual and community-wide cultural practices, the linguistic resources of the language itself, and patterns of language use. The Sociotopographic Model attempts to model the interplay of these diverse factors.

References

- 1 Jürgen Bohnemeyer. Spatial frames of reference in Yucatec: Referential promiscuity and task-specificity. *Language Sciences*, 33:892–914, 2011.

- 2 Jürgen Bohnemeyer, Katharine T. Donelson, Randi E. Tucker, Elena Benedicto, Alejandra Capistrán Garza, Alyson Eggleston, Néstor Hernández Green, María de Jesús Selene, Hernández Gómez, Samuel Herrera Castro, Carolyn O'Meara, Enrique Palancar, Gabriela Pérez Báez, Gilles Polian, and Rodrigo Romero Méndez. The cultural transmission of spatial cognition: Evidence from a large-scale study. In *Cogsci 2014 Proceedings*, pages 212–217, 2014. URL: <https://mindmodeling.org/cogsci2014/papers/047/paper047.pdf>.
- 3 Jürgen Bohnemeyer and Christel Stolz. Spatial reference in Yukatek Maya: A survey. In Stephen C. Levinson and David Wilkins, editors, *Grammars of space: Explorations in cognitive diversity*, pages 273–310. Cambridge University Press, Cambridge, 2006.
- 4 Jürgen Bohnemeyer and Randi Tucker. Space in semantic typology: Object-centered geometries. In Peter Auer, Martin Hilpert, Anja Stukenbrock, and Benedikt Szmrecsanyi, editors, *Space in language and linguistics: Geographical, interactional, and cognitive perspectives*, pages 637–666. Mouton de Gruyter, Berlin, 2013.
- 5 Eve Danziger. Language, space and sociolect: Cognitive correlates of gendered speech in Mopan Maya. In Catherine Fuchs and Stéphane Robert, editors, *Language diversity and cognitive representations*, volume 3, pages 85–107. John Benjamins Publishing Company, Amsterdam, 1999. doi:10.1075/hcp.3.09dan.
- 6 Cris Edmonds-Wathen. *Frame of Reference in Iwaidja: Towards a culturally responsive early years mathematics program*. PhD thesis, RMIT, Melbourne, 2012.
- 7 Olivier Le Guen. Modes of pointing to existing spaces and the use of frames of reference. *Gesture*, 11(3):271–307, 2011. doi:10.1075/gest.11.3.02leg.
- 8 Stephen C. Levinson. Frames of reference and Molyneux's question: Crosslinguistic evidence. In Paul Bloom, Mary A Peterson, Lynn Nadel, and Merrill F Garrett, editors, *Language and Space*, pages 109–169. MIT Press, Cambridge, 1996.
- 9 Stephen C. Levinson. *Space in language and cognition: Explorations in cognitive diversity*. Cambridge University Press, Cambridge, 2003.
- 10 Stephen C Levinson and David Wilkins, editors. *Grammars of space: Explorations in cognitive diversity*. Cambridge University Press, Cambridge, 2006.
- 11 Jonathon Lum. *Frames of spatial reference in Dhivehi language and cognition*. PhD thesis, Monash University, Melbourne, 2018.
- 12 Asifa Majid, Melissa Bowerman, Sotaro Kita, Daniel B.M. Haun, and Stephen C. Levinson. Can language restructure cognition? The case for space. *Trends in Cognitive Sciences*, 8(3):108–114, 2004.
- 13 Felicity Meakins, Caroline Jones, and Cassandra Algy. Bilingualism, language shift and the corresponding expansion of spatial cognitive systems. *Language Sciences*, 54:1–13, 2016. doi:10.1016/j.langsci.2015.06.002.
- 14 Carolyn O'Meara and Gabriela Pérez Báez. Spatial frames of reference in Mesoamerican languages. *Language Sciences*, 33(6):837–852, 2011. doi:10.1016/j.langsci.2011.06.013.
- 15 Bill Palmer. Topography in language: Absolute Frame of Reference and the Topographic Correspondence Hypothesis. In Rik De Busser and Randy J. LaPolla, editors, *Language structure and environment. Social, cultural and natural factors*, pages 179–226. John Benjamins Publishing Company, Amsterdam, 2015.
- 16 Bill Palmer, Alice Gaby, Jonathon Lum, and Jonathan Schlossberg. Socioculturally mediated responses to environment shaping universals and diversity in spatial language. In Paolo Fogliaroni, Andrea Ballatore, and Eliseo Clementini, editors, *Proceedings of workshops and posters at the 13th International Conference on Spatial Information Theory (COSIT 2017)*, pages 195–205. Springer International Publishing, Cham, 2018. doi:10.1007/978-3-319-63946-8_35.

- 17 Bill Palmer, Jonathon Lum, Jonathan Schlossberg, and Alice Gaby. How does the environment shape spatial language? Evidence for sociotopography. *Linguistic Typology*, 21(3), 2017. doi:10.1515/lingty-2017-0011.
- 18 Eric Pederson. Geographic and manipulable space in two Tamil linguistic systems. In Andrew U. Frank and Irene Campari, editors, *Spatial Information Theory: A theoretical basis for GIS*, pages 294–311. Springer-Verlag, Berlin, 1993.
- 19 Eric Pederson. Spatial language in Tamil. In Stephen C. Levinson and David Wilkins, editors, *Grammars of space: Explorations in cognitive diversity*, pages 400–436. Cambridge University Press, Cambridge, 2006.
- 20 Eric Pederson, Eve Danziger, David Wilkins, Stephen C. Levinson, Sotaro Kita, and Gunter Senft. Semantic typology and spatial conceptualization. *Language*, 74(3):557–589, 1998.
- 21 Jonathan Schlossberg. *Atolls, islands and endless suburbia: space and landscape in Marshallese*. PhD thesis, University of Newcastle, Newcastle, Australia, 2018.
- 22 Joshua A. Shapero. Does environmental experience shape spatial cognition? Frames of Reference among Ancash Quechua speakers (Peru). *Cognitive Science*, 41:1274–1298, 2017. doi:10.1111/cogs.12458.
- 23 Angela Terrill and Niclas Burenhult. Orientation as a strategy of spatial reference. *Studies in Language*, 32(1):93–136, 2008. doi:10.1075/sl.32.1.05ter.




■ **Figure 2** Strategy tendencies and subsistence mode in Laamu fishing versus non-fishing communities [17].

Flexible Patterns of Place for Function-based Search of Space


Emmanuel Papadakis

Dept. of Geoinformatics - Z_GIS, University of Salzburg, Schillerstr. 30, 5020 Salzburg, Austria
emmanouil.papadakis@sbg.ac.at

 <https://orcid.org/0000-0001-8669-2420>


Andreas Petutschnig

Dept. of Geoinformatics - Z_GIS, University of Salzburg, Schillerstr. 30, 5020 Salzburg, Austria
andreas.petutschnig@sbg.ac.at

 <https://orcid.org/0000-0001-5029-2425>

Thomas Blaschke

Dept. of Geoinformatics - Z_GIS, University of Salzburg, Schillerstr. 30, 5020 Salzburg, Austria
thomas.blaschke@sbg.ac.at

 <https://orcid.org/0000-0002-1860-8458>

Abstract

Place is a human interpretation of space; it augments the latter with information related to human activities, services, emotions and so forth. Searching for places rather than traditional space-based search represents significant challenges. The most prevalent method of addressing place-related queries is based on placenames but has limited potential due to the vagueness of natural language and its tendency to lead to ambiguous interpretations. In previous work we proposed a system-oriented formalization of place that goes beyond placenames by introducing composition patterns of place. In this study, we introduce flexibility into these patterns in terms of what is necessarily or possibly included when describing the spatial composition of a place and propose a novel automated process of extracting these patterns relying on both theoretical and empirical knowledge. The proposed methodology is exemplified through the use case of locating all the shopping areas within London, UK.

2012 ACM Subject Classification Information systems → Geographic information systems, Theory of computation → Modal and temporal logics

Keywords and phrases Functions, Place, Patterns, Function-based search, Place-based GIS

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.54

Category Short Paper

Funding The presented work is framed within the Doctoral College GIScience (DK W 1237N23), funded by the Austrian Science Fund (FWF).

1 Introduction and Related Work

People live and act on space but deal and interact with places. Place is a human invention to describe space [2] and spatial experience. Several disciplines formalize the notion of place to enable a human friendly way of searching space, henceforth place *search*. Indicative approaches treat place as a reference (placename) to the geographical space either in its primitive form, or augmented with semantics, or even as a full-fledged model. An emerging question is to determine the extent to which the existing place search techniques can encapsulate the



© Emmanuel Papadakis, Andreas Petutschnig, and Thomas Blaschke;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 54; pp. 54:1–54:7

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

nature of place, reflecting the meaning infused within. For instance, “is it possible to locate a shopping area if it is not denoted as a shopping mall?” This work contributes to the formalization of place and development of place search methodology that is not limited to placenames or semantic infusion of spatial entities; instead, it treats places as entities that conform to specific spatial patterns. The objective of this methodology is twofold: (1) to provide an enhanced representation of place that relies on both statistical and narrative information and (2) to identify locations and extents of places that possess a set of desired features, in order to yield results that cannot be captured simply through search based on placenames or place properties.

The most prevalent method of place search relies on digital gazetteers [4], which are spatially-referenced catalogs of placenames. The major limitation of this approach is the lack of information, which is narrowed down to placenames, spatial footprints and simple properties such as place types. The use of ontologies [5] overcome these limitations. CIDOC CRM [3] is an upper level ontology that provides a detailed knowledge representation about places facilitating sophisticated search; however, most of the ontologies provide relative, limited or devoid absolute spatial representation. Ontological gazetteers [6], on the other hand, combine the aforementioned methods by enriching the traditional structure of placenames and spatial footprints with additional semantics. Nevertheless, according to [11], the meaning of place is something more than a collection of semantics.

Following a meta-modeling approach, [9] establishes places from narratives by taking into account the relations between semantics instead of lists of properties, combined with spatial information. However, the high dependency on natural language makes this approach context-dependent, as well as, it raises many technical difficulties. On the other side of the spectrum, [10] follows a bottom-up approach. Particularly, it gives emphasis on the extraction of semantic signatures of places, in the form of co-occurrence patterns of points of interest, using LDA topic modeling and statistical analysis. These patterns are then used to discover similar regions that comply to the aforementioned signatures. The unsupervised nature of this method imply certain limitations with respect to describing the plausible meaning of places.

In previous work, we proposed the model of functional space [7]. This emphasizes on a fraction of the meaning of place that is functionality. Particularly, according to this model a place is a system that satisfies people’s purposes by offering certain functions. These functions are regarded as “services” enabled or disabled by the spatial organization that governs a place, known as composition. Under this model, places are formalized based on composition patterns[8], which are defined as sets of components and rules, denoted as *Comps* and *Rules*, respectively. The former refers to the physical entities that constitute a place, whereas the latter is a set of implications between functions and first-order logic formulas that form the composition rules. These rules stand for relations between physical entities and external variables. The overall formalization is visualized in Figure 1 and the set of all the available composition rules, denoted as C_R is shown in Figure 2. Composition patterns are created through text analysis. Specifically, narratives, such as dictionaries, Wikipedia pages, design guidelines and so on, are analyzed to extract information about the functions and the composition of a place. This knowledge is then organized as a composition pattern, essentially offering a commonly accepted blueprint for the place under consideration.

The composition patterns enable function-based search of space [7], that is, locating places that support certain functions. However, the rigid rules that describe the composition patterns can be more restrictive than necessary in some use cases. In particular, since the composition rules are expressed in first-order logic, every associated function is either

Element	Description
Functions	Functions the place offers
Comp	Components that form the place
C_R	Composition rules
F_R	$F \rightarrow f(C), F, \in \text{Functions}, C, \in \text{Comp}, f \text{ logical formula}$

Composition rule	Description
Occurrence (A, N)	Component A appears N times. $N \in \mathbb{Z}^+$
Property (A, <name, value>)	Component C has property <name, value>
PartOf(A, B)	Component A is a part of component B
Correlation (A, B, N)	Proportion of occurrence of components A and B is N. $N \in \mathbb{R}^+$
Topology (A, B, T)	Component A and B have topology T (DE-9IM)
Distribution (A, MI)	The Moran's Index of component A is MI $\in [-1, 1] \subset \mathbb{R}$
Proximity (A, B, Dist)	Average distance of A and B is Dist $\in \mathbb{R}^+$
Organization (SO, A)	Component A has SO distribution. SO $\in [\text{linear, centralized, radial}]$

■ **Figure 1** Composition pattern. ■ **Figure 2** Composition rules.

permitted or forbidden. This hinders the effectiveness of place search, especially when dealing with inconsistent data or in cases of increased vagueness that marks a function to be optional. Furthermore, the extraction of composition patterns highly depends on narratives, which often reflect the ideal or the most general definition of a place, abstracting away the diversity that characterizes the real world.

This study proposes an improved version of the composition patterns of place, described above, that addresses the aforementioned limitations. Specifically, the composition patterns are extended to support flexibility in terms of what is necessarily or possibly included in the composition of a place. In addition, the extraction process of composition patterns is enhanced with empirical knowledge, which revise and complements the knowledge extracted by narratives.

The remainder of this paper is organized as follows. Section 2 proposes the improved composition pattern of place and introduces an empirical methodology for extracting patterns of place. Then, Section 3 demonstrates the applicability of the proposed model for the use case of searching shopping areas in London, UK. Finally, Section 4 concludes and points out directions for future work.

2 Methodology

In this section, we first analyze the required extensions that will improve the composition patterns of place with flexibility. Then, we detail how the process of extracting patterns is (1) adjusted to conform to the extended formalization and (2) enhanced in order to allow automations and patterns with finer details based on spatial analysis and statistics.

2.1 Flexible Composition Pattern

The initial model of composition patterns of place is extended by introducing its flexible counterpart, the model of *flexible composition patterns*. This extension is made possible by applying the principles of modal logic [1]. Note that modalities are chosen, instead of quantities, to represent necessity and possibility in a concise and natural manner, and to preserve the model's generality. In the remainder of this document, we refer to these flexible composition patterns simply as patterns.

The newly proposed patterns conform to the fundamental assumption that “place is space with ascribed functions” and are formalized as a collection of three sets ($Comp, C_R, F_R$). The first two stand for the possible components and composition rules (Figure 2) that frame the composition of a place, respectively. F_R contains logical implications between functions and logical formulas comprised of composition rules. The latter are extended with modal operators that allow the expressions “necessarily” and “possibly” (denoted as \square and \diamond ,

respectively) in order to attribute a certainty value for every rule. Considering the above, a particular function is enabled, if the combination of necessary or possible rules holds.

2.2 Enhanced Pattern Extraction

In order to achieve automated creation of more realistic patterns, we propose an extraction process that utilizes both theoretical and empirical knowledge. According to this approach, a pattern of place is no longer a strict reflection of the written word, but a combination of text-based and experiment-based information acquired through the phases of *theoretical design* and *empirical revision*, respectively.

The phase of theoretical design uses text analysis to extract a *theoretical* pattern, which includes the textually derived knowledge about the composition of a place. This pattern is regarded as a collection of “echoes”, after Alexander’s 15 structural properties [12] and it describes the expected features that would enable the functions of the place under question. Since text-based knowledge is usually designed with generalization in mind, it is safe to assume that the composition rules included in this pattern are marked as possible (and not necessary) in terms of the level of certainty.

The empirical revision focuses on the analysis of regions that are considered as the ideal candidates of the place for which the theoretical pattern was created. More specifically, spatial and semantic data are acquired for a wide range of ideally defined instances of the place under question. Considering the latter as anchors, additional data is collected about adjacent components conforming to requirements listed in the theoretical pattern.

The next step aims to extract and describe the most significant components that characterize the ideal places under question. This is achieved by classifying the aggregated data into context-specific categories by conducting statistical and spatial analysis. Statistical analysis includes extraction of the population count and the average frequency of occurrences per category. Spatial analysis, on the other hand, focuses on the mean distance between components and the centroids of the ideal candidates of place. By the end of the analysis, an *empirical* pattern is constructed that includes the required and optional information that describes the composition of the place under question. A context-specific significance threshold is employed in order to classify which rules are considered as necessary or optional. The aforementioned threshold is chosen empirically and is calculated based on the statistical importance. Particularly, we assume the following convention: aggregated data that exceed this threshold introduce necessary composition rules, while the rest imply possible rules.

It is worth noting that there are cases where necessity or possibility rules depend on the possible or necessary existence of some components, respectively. These scenarios imply to possible necessity and conditional possibility and so forth. Figure 3 illustrates all the possible cases of interrelation between necessity rules and conditional existence of components along with the corresponding descriptions.

3 Experiment

This section demonstrates the proposed methodology using the example of shopping malls in London, UK. The objective of the described experiment is to create a pattern which can enable a place search system to locate all places that offer functions similar to a shopping mall, even if they are not explicitly defined as one. By convention, we refer to these places as shopping areas, for which the ideal representatives are the standard shopping malls.

Before proceeding, we list some basic assumptions that underline our experiment. We consider a simple version of shopping areas that support the functions of shopping experience,

Dependency Condition	Explanation
Let C be a component and R a composition rule	
$\square C \text{ AND } \square R$	C exists and R must hold
$\square C \text{ AND } \diamond R$	C exists and R can hold
$\diamond C \text{ AND } \square R$	if C exists then R must hold
$\diamond C \text{ AND } \diamond R$	if C exists then R can hold

■ **Figure 3** Dependency between necessity rules and existence of components.

Components	
Shop	Amenity
Road Junction	Road
Bus station	
Functions	
Shopping experience (F_S)	Existence of Shops
Leisure (F_L)	Existence of Amenities
Walkability (F_W)	Shops and Amenities are within a walkable distance.
Accessible to drivers (F_{AD})	Existence of Road junctions and Roads within min. driving distance.
Accessible to non drivers (F_{AND})	Existence of bus stations within walkable distance.

■ **Figure 4** Functions and components of a shopping area.

leisure, walkability, accessibility to drivers and accessibility to non-drivers (Figure 4). In addition, we assume that the maximum walkable distance is 500m and the minimum driving distance is no more than 5000m. We use a subset of the composition rules in Figure 2 that includes *Occurrence*, *Correlation* and *Proximity*.

Considering the assumptions above, textual analysis is performed on the following sources: Wikipedia reference, dictionary definition and architectural guidelines of shopping malls. This results in the theoretical pattern depicted in Figure 5.

Empirical revision is then conducted using data acquired from OpenStreetMap. We collected a set of 63 polygons, outlining shopping malls in London. Using the centroids of the latter we aggregate: (1) point geometries of shops, amenities and public transport stops within a 500m radius, and (2) junction points along with line geometries of primary and secondary highways within a 5000m radius. Figure 7 illustrates indicative results of the spatial and statistical analysis applied on the acquired components for all the ideal instances of shopping areas.

For the construction of the empirical pattern, we assume that a variable is significant and hence it implies a necessary composition rule if the coefficient of variation for the corresponding mean value is no more than 25%. Values more than this level result to insignificant variables and, hence, refer to possible rules. The empirical pattern is shown in Figure 6.

Our method is evaluated by conducting and comparing two function-based search processes for shopping areas: one relying on the theoretical pattern, and one on the empirical one. Pattern matching is realized by converting each pattern to a sequence of spatial queries and procedures, implemented using PostGIS and QGIS. Particularly, every function, included in the pattern, is expressed as a query that reflects the implied composition rules. Afterwards, the generated queries are issued on the database. The theoretical pattern is evaluated by aggregating the results of each query in a conjunctive manner. The empirical pattern, on the other hand, is evaluated in two steps. Initially, queries based on the necessity rules suggest candidate regions of the place under question; then the possibility rules are checked

Function	Formula	Function	Formula
{F ₁ }	Occurrence(Shop, [2,])	{F ₁ }	□ Occurrence(Shop, [4,])
{F ₂ }	Occurrence(Amenity, [1,]) AND Correlation(Shop, Amenity, [2,])	{F ₂ }	◇ Occurrence(Amenity, [1,]) AND ◇ Correlation(Shop, Amenity, [7,])
{F _w }	Proximity(Shop, Amenity, [, 500m]) AND Proximity(Shop, Shop, [, 500m]) AND Proximity(Amenity, Amenity, [, 500m])	{F _w }	□ Proximity(Shop, Amenity, [, 63m]) AND □ Proximity(Shop, Shop, [, 63m]) AND □ Proximity(Amenity, Amenity, [, 63m])
{F _{rd} }	Proximity(Shop, Road Junction, [, 500m])	{F _{rd} }	□ Proximity(Shop, Road Junction, [, 3496m])
{F _{st} }	Proximity(Shop, Bus station, [, 500m])	{F _{st} }	◇ Proximity(Shop, Bus station, [, 291m])

■ **Figure 5** Theoretical pattern.

■ **Figure 6** Empirical pattern.

	Count(shop)	Count(amenity)	Proximity(Shop, Amenity)	Proximity(Bus st., Shop)	Count(Bus st)
min	4	0	5.8	47	0
median	64	17	63	261	4
coefficient of variation	19%	49%	22%	47%	8%

■ necessary
■ possible

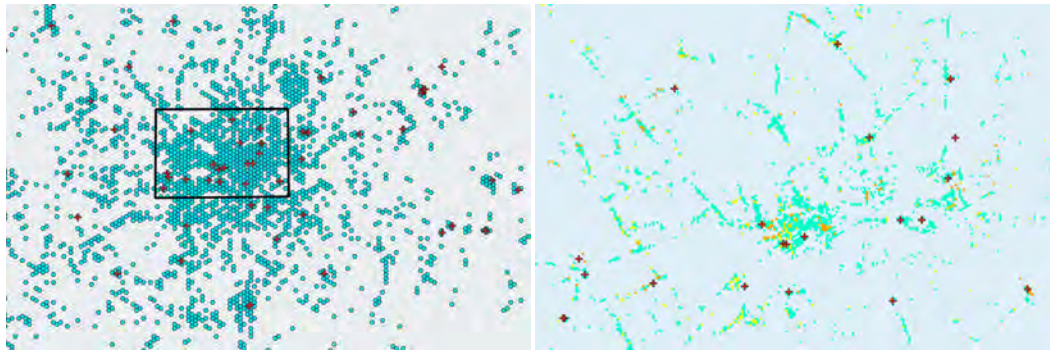
■ **Figure 7** Indicative results of spatial and statistical analysis.

in order to mark which among the selected candidates better fit the initial functionality. The algorithms used for the function-based search go beyond the scope of this work, and the methods used are not considered optimal but used for demonstration purposes only.

Figure 8 illustrates the results retrieved by the theoretical pattern, with blue cells representing all candidate shopping areas and cross symbols indicating the locations of the shopping malls. It should be clear that all ideal places are included, however there is no indication as to which cells better support the required functions. In contrast, the results shown in Figure 9 indicate that the empirical pattern enables a finer delineation of shopping areas, as well as a clear indication of the level of functions support (illustrated using a heat map where blue represents least support and red represents highest support). Note that due to a much smaller grid size, Figure 9 corresponds only to the rectangle area in Figure 8.

4 Conclusion

This study contributes to formalizing place and place search. In particular, we introduce a more flexible formalization of place capable of capturing what is necessarily or possibly included in the composition of a particular place. Furthermore, we propose a pattern extraction process that combines the theoretical, text-based design of composition patterns of place with empirical revision based on statistical and spatial analysis. The resulting pattern provides a detailed description of place that is closer to reality and can lead to more accurate results in function-based search of space, as evidenced by the conducted experiment of locating shopping areas in London, UK. This work indicates that place can be treated as a functional region and be formalized as a system using both narratives and spatial data, however further development is necessary. The dependency of theoretical patterns on narratives raises important obstacles; indicatively, natural language processing has many technical difficulties and usually the extracted information is context-dependent and highly vague. In addition, a more formal definition of the composition rules is required, which will allow the introduction of new rules or the modification of existing ones. Furthermore, although modal logic seems a convenient solution, when it comes to reasoning, it hinders quantification, which in return limits the model's ability to provide grading on places or functionality rating. Interesting directions of future work include the integration of probabilistic models, which will quantify the possible knowledge about places, and the automation of the pattern extraction process utilizing deep learning techniques.



■ **Figure 8** Results using theoretical pattern.

■ **Figure 9** Results using empirical pattern.

References


- 1 Alexander Chagrov. *Modal logic*. Oxford Logic Guides, 1997.
- 2 Michael R Curry. *The work in the world: geographical practice and the written word*. U of Minnesota Press, 1996.
- 3 Martin Doerr. The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI magazine*, 24(3):75, 2003.
- 4 Michael F Goodchild and Linda L Hill. Introduction to digital gazetteer research. *International Journal of Geographical Information Science*, 22(10):1039–1044, 2008.
- 5 Nicola Guarino, Daniel Oberle, and Steffen Staab. What is an ontology? In *Handbook on ontologies*, pages 1–17. Springer, 2009.
- 6 Linda L Hill. Core elements of digital gazetteers: placenames, categories, and footprints. In *International Conference on Theory and Practice of Digital Libraries*, pages 280–290. Springer, 2000.
- 7 Emmanuel Papadakis and Thomas Blaschke. Place-based GIS: Functional Space. *Proceedings of the 4th AGILE PhD School*, 2088, 2017.
- 8 Emmanuel Papadakis and Thomas Blaschke. Composition of Place: Components and Object Properties. *International Journal of Geo-information*, 2018. Submitted, under review.
- 9 Simon Scheider and Ross Purves. Semantic Place Localization from Narratives. In *Proceedings of The First ACM SIGSPATIAL International Workshop on Computational Models of Place*, pages 16:16–16:19, New York, NY, USA, 2013. ACM.
- 10 Gao Song, Janowicz Krzysztof, and Couclelis Helen. Extracting urban functional regions from points of interest and human activities on location-based social networks. *Transactions in GIS*, 21(3):446–467, 2017.
- 11 Yi-Fu Tuan. Space and Place: Humanistic Perspective. In *Philosophy in geography*, pages 387–427. Springer, 1979.
- 12 Maria Vasardani, Martin Tomko, and Stephan Winter. The Cognitive Aspect of Place Properties. In *International Conference on GIScience Short Paper Proceedings*, volume 1, 2016.

Novel Models for Multi-Scale Spatial and Temporal Analyses

Yi Qiang

University of Hawaii - Manoa, Honolulu, HI, the United States

yiqiang@hawaii.edu

 <https://orcid.org/0000-0002-6872-8837>

Barbara P. Buttenfield

University of Colorado - Boulder, Boulder, CO, the United States

babs@colorado.edu

Nina Lam

Louisiana State University, Baton Rouge, LA, the United States

nlam@lsu.edu

Nico Van de Weghe

Ghent University, Ghent, Belgium

nico.vandeweghe@ugent.be

Abstract

Multi-scale analysis for spatio-temporal data forms a fundamental challenge for many analytic systems. In geographic information systems, analysis and modeling at pre-defined spatial and temporal scales may miss critical relationships in other scales. Previous studies have investigated the uses of the triangle model as a multi-scale framework in analyzing temporal data. This article demonstrates the utilities of the triangle model and pyramid model for multi-scale spatial analysis through real-world analytical tasks and discusses the potential of developing a unified modeling framework that integrates the two models.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases Triangle Model, Pyramid Model, multi-scale spatial and temporal analysis, GIS

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.55

Category Short Paper

1 Introduction

Currently, the increasing diversity of geospatial data collected at different resolutions (e.g. satellite, UAV, field sampling, and census data) poses serious challenges for data integration and analyses. The choice of analytic scale to a large extent determines the insights that can be gained, due to the nature of geospatial information and due to its sensitivity to spatial and temporal resolution. The importance of scale has been epitomized in the well-known Modifiable Areal Unit Problem (MAUP). Ideally, geospatial data should be analyzed at multiple scales to reveal the nested interactions at different scales and decisions should be made at the level where the spatial and temporal relationships are maximized. However, the scale of analysis that is best suited for a given problem is not always immediately evident, which raises a compelling justification for exploring data solutions supporting multiple-scale analyses.



© Yi Qiang, Barbara P. Buttenfield, Nina Lam, and Nico Van de Weghe;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 55; pp. 55:1–55:7

Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

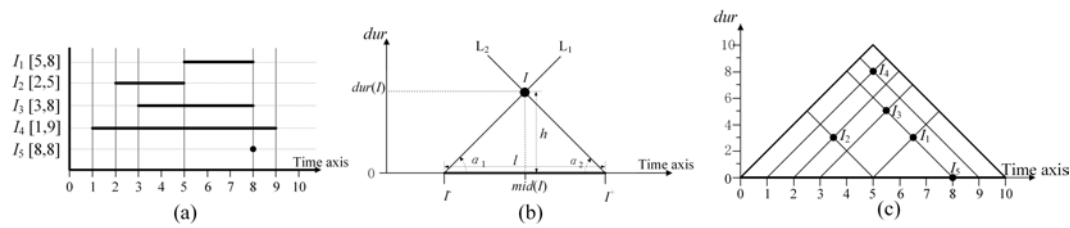
Multi-scale analysis for spatio-temporal data forms a longstanding challenge for analytic systems in different disciplines. In GIS, spatial data are represented as flat layers and temporal data are represented as linear sequences at pre-determined resolution, with spatial analysis tools operating usually at a single scale. For instance, kernel density can display clusters of features only at a single spatial scale. To discover clusters concealed in other scales, the analysis needs to adjust the bandwidth using an inefficient “trial-and-error” approach that repeats the density calculation at one or more alternative scales. Similar issues exist in image classification and land cover change modeling, which usually are based on pixel-centered single-scale methodologies that can ignore or obscure the impact of scale and hierarchy in landscape processes that drive pattern creation. To fully understand the complexity of coupled natural and human systems, the interactions and competitions among different systems need to be analyzed and modeled at multiple scales.

To address the issue of multi-scale temporal analysis, Qiang et al. [4][5] proposed a Triangle Model (TM) that projects linear temporal data onto a 2D space and demonstrated how variation of data across multiple temporal scales can be represented in a continuous 2D space [3]. The TM was later applied in analyzing movement data [7]. This paper demonstrates the utility of the triangle model in evaluating surface-adjusted distance measurements in digital elevation models and the utility of its extension (the pyramid model) in analyzing land fragmentation. Finally, we will discuss the potential of building a unified framework for integrating the triangle and pyramid model for multi-scale spatio-temporal analyses.

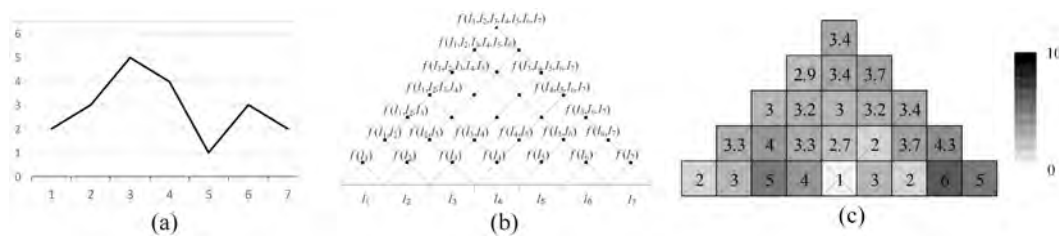
2 Triangle Model

Time intervals are conventionally represented as linear segments in a one-dimensional space (Figure 1(a)). Alternatively, a 2D representation of intervals was originally introduced by Kulpa [2] as a diagrammatic tool for mathematical reasoning. Later, Qiang et al. [4][5][3] extended the model for spatio-temporal analysis, and implemented it into a GIS. In Qiang et al’s approach, a time interval (starting at I^- and ending at I^+ can be mapped to a point at $((I^+ + I^-)/2, (I^+ - I^-)/2)$ in a 2D Cartesian coordinate system (Figure 1(b)). The position of the point in the horizontal axis $((I^+ + I^-)/2)$ indicates the midpoint of the interval, while the vertical position $((I^+ - I^-)/2)$ is proportional to the length of the interval. Using this approach, which is termed the Triangle Model (TM), all time intervals can be represented as unique points in a 2D coordinate space. Figure 1(c) demonstrates a TM depiction of the five intervals shown in Figure 1(a), illustrating its facility for representing temporal properties (e.g. start, end, midpoint and duration) in a compact view. One of the advantages of the TM is that by converting temporal relations into a spatial representation, the TM permits temporal analysis to be conducted seamlessly across multiple scales using simple GIS operations.

In addition to time intervals, the TM can be used to represent sequential time series data [3]. A time series (e.g. daily temperature) consists of a sequence of intervals (e.g. days) associated with an attribute (e.g. average temperature of the day) (Figure 2(a)). Daily intervals, which are the finest granularity, can be represented as points at the lowest level in a TM. Intervals of every two days can be represented as points at 2nd level and intervals of every three days can be represented as points at the 3rd level. The point on the top represents the interval of the entire time series (Figure 2(b)). Each point in the TM is associated with an aggregated value (e.g., the mean or standard deviation) of attributes of the intervals it represents. Figure 2(c) shows the values of each day shown in Figure 2(a) along its lowest level; and a nested average for each interval is easily computed in the TM. Interpolation of



■ **Figure 1** The transformation from the linear model to the triangle model. (a) Time intervals in the linear model. (b) Projecting a time interval into a point in the triangle model. (c) Time intervals in (a) shown in the triangle model.



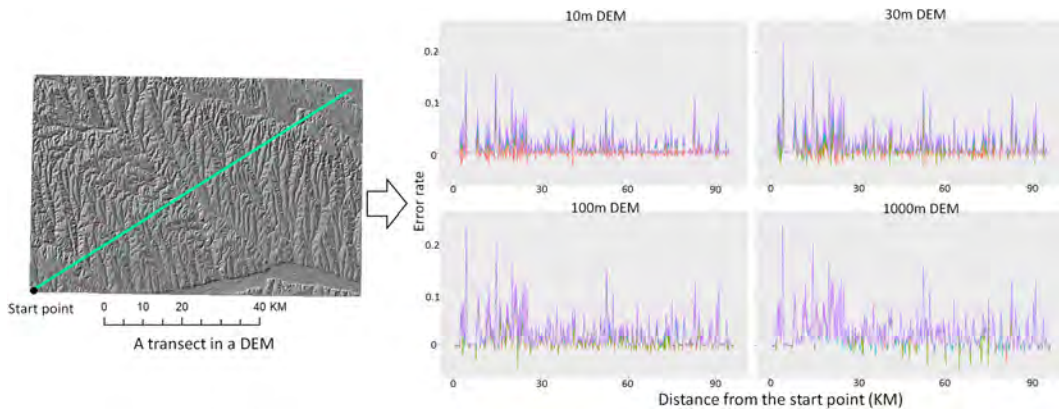
■ **Figure 2** Representation of time series in the TM: (a) a time series represented in a conventional line chart; (b) representing time intervals in the time series in (a) as points in the TM; (c) a rasterized TM showing nested means for the time series.

the nested values within the TM can form a continuous field representing all intervals in the time series, providing an explicit view of the hierarchy and nested relations of patterns across different scales. Using conventional spatial analysis methods, (e.g. classification, overlay, and Map Algebra), multiple time series can be compared to support multi-criteria decision-making at different temporal scales.

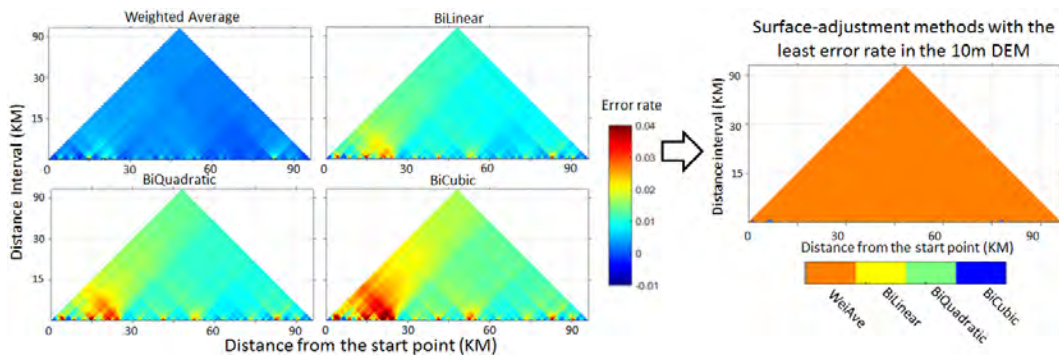
3 Surface-Adjusted Distance in the Triangle Model

In GIScience, distance is the most fundamental spatial metric that anchors proximity analysis, spatial pattern detection, and spatial interpolation, and, indeed, underlies detection of nearly every type of geospatial pattern. Similar to time series data, distance measurement is a linear process based on aggregating distances in small intervals. Current distance measurement on terrain assumes that Digital Elevation Model (DEM) pixels are rigid and flat, as tiny facets of ceramic tile approximating a continuous terrain surface. It is still unclear how the measurement errors using different approaches propagate over scales in all types of terrain. As the measured distance increases, the errors introduced by the assumption of rigid pixels can propagate dramatically and increase overall error [1], or cancel each other out and result in coarse-scale accuracy.

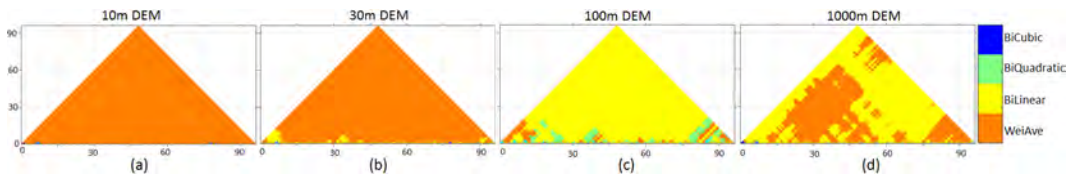
Distance measurements of four surface-adjustment approaches, including weighted average (WeiAve), and three polynomial approaches (i.e. Bilinear, Biquadratic and Bicubic) are compared in this section. Please refer to [1] for the details of these surface-adjusted approaches. The distance of a transect is measured using the four surface-adjustment approaches on DEMs at different resolutions (10, 30, 100 and 1000m). Then, the measured distances are compared with the benchmark distance measured on a 3m LiDAR DEM to evaluate their accuracies. Using traditional visualization methods, accuracies of the surface-adjustment distance measurements can only be examined at a single scale. For instance, Figure 3 shows



■ **Figure 3** A transect in a DEM (left) and error rates of distance measurement in 100m intervals along the transect (right). The colors of the lines indicate different surface-adjusted methods.



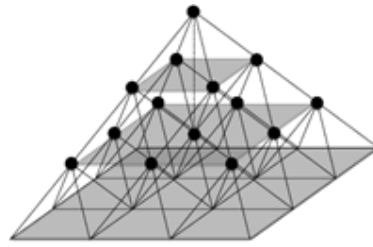
■ **Figure 4** Error rates of distance measurements in the 10m DEM represented in TM (left) and the overlaid TM showing the most accurate approach (right).



■ **Figure 5** Comparison of the four surface-adjustment methods for measuring different lengths of intervals in the TM. Colors indicate the most accurate approach.

error rate of distance (computed as (benchmark distance – measured distance)/benchmark distance) in every 100m interval along the transect, using different surface-adjustment methods and on DEMs of different resolutions. From the linear charts, it is also difficult to analyze the propagation of errors at different scales and to compare the surface-adjustment approaches in measuring different interval lengths.

The TM provides a compact view of error rates of measuring different lengths of intervals (Figure 4). By overlaying the TMs of different surface-adjustment approaches, we can identify the most accurate (lowest error rate) approach in measuring different lengths of intervals along the transect. The left side of Figure 4 demonstrates the error rates of the four surface-adjustment approaches in the 10m DEM. The right side is the result of overlaying the four TMs in the left panel, where the colors denote the most accurate surface-adjustment



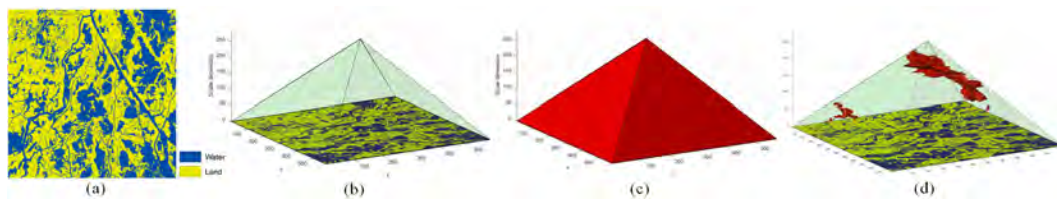
■ **Figure 6** A raster in a Pyramid Model (PM).

approach in measuring different lengths of intervals. The single color (orange) of the overlaid TM indicates that weighted average is the most accurate approach in measuring all intervals in the 10m DEM. However, as shown in Figure 5(b), the Bilinear approach outperforms weighted average in some short intervals (yellows in the bottom of the TM) in the 30m DEM. In the 100m DEM, Bilinear approach becomes the most accurate approach for long intervals, while Bilinear and Biquadratic have better accuracy in short intervals (Figure 5(c)). In the 1000m DEM, Bilinear and weighted average have competing accuracies in measuring distances in different intervals (Figure 5(d)). The measurement errors represented in the TMs inform the best surface-adjustment approach for measuring distance in different lengths of intervals. Next, the relationship between the measurement accuracy and terrain roughness will be explored in the framework of TM, which will be presented in the conference.

4 Multi-Scale Spatial Analysis in the Pyramid Model

The concept of the Pyramid Model (PM) is similar to an image pyramid, which represents a raster image across multiple resolutions by smoothing and resampling. Image pyramids were originally developed in computer vision, image processing and signal processing, but are now used more widely to enhance the efficiency of multi-scale raster rendering in GIS. In our approach, the construction of a PM is similar to that of the TM in the sense of developing a hierarchy; but the PM represents 2D space across scales instead of linear time. To construct a PM, every pixel in the base raster can be represented as a point at the lowest level in the pyramid. Points at the second level represent square region of four pixels (2×2) in the raster. Points at the n th level represent square regions of n^2 pixels. In constructing this hierarchy, all square regions of different sizes in the base raster are represented as a pyramid containing uniformly arranged points in a 3D space (Figure 6). The pyramid can be represented as a 3D raster that consists of numerous equal-size voxels, each of which is associated with an aggregated statistic of the attributes or spatial metrics (e.g. density, texture, spatial dependence) for the square region it represents. The PM can also represent vector features (e.g. point, line, and polygon), in which points represent square regions in the base layer.

The utility of the PM is demonstrated in analyzing wetland fragmentation in coastal Louisiana. Published evidence shows that fragmented wetland habitats may accelerate wetland erosion and wetland loss (e.g. Lam et al. 2018). However, fragmentation indices calculated for different sizes of focal windows may lead to different results. Similar issues exist in other spatial pattern indices such as density, spatial autocorrelation, and terrain roughness. Figure 7 illustrates PM representations of fractal dimension (a commonly used fragmentation index) of a binary land cover raster clipped from coastal Louisiana. Using the land cover raster as the base (Figure 7(b)), local fractal dimensions for different sized focal windows can be stacked into a 3D PM (Figure 7(c)) in which lower layers represent



■ **Figure 7** Representing local fractal dimensions of land cover raster in a PM. (a) A binary land cover raster. (b) The land cover raster in the base of a PM. (c) A 3D Pyramid Model built from the land cover raster. (d) Voxels with a fractal dimension in the 99th percentile.

fractal dimensions for smaller regions, and higher layers represent fractal dimensions for larger regions. The internal variation of the PM can be visualized using ‘spatial query’. For instance, Figure 7(d) displays the voxels in the 99th percentile (i.e. the highest 1%) fractal dimension calculated in different sizes of focal windows, indicating the most fragmented regions at different scales. Extending map algebra into the 3D space, the variation in the PM can be further analyzed and multiple PMs can be compared or correlated. For instance, subtracting PMs of fractal dimension calculated at two time points, one can identify land areas where fragmentation has accelerated significantly. Moreover, regression analysis can be conducted between the PM of the land fragmentation and density of man-made structures to discover the scales at which human activities have most impact on wetland erosion.

5 Summary and Future Work

This study demonstrates the applications of the TM and the PM in multi-scale spatial analyses. Compared with traditional analytic tools that are limited to a single scale, the PM and TM can represent spatial patterns and relationships in a full dimension of continuous changing scales, and facilitate queries across spatial and temporal scales. This study demonstrates the use of two multi-scale models in evaluating distance measurement in DEMs and analyzing landscape fragmentation respectively. Beyond these, our future research plan includes developing a unified modeling framework that integrates TM and PM to fully support multi-scale spatio-temporal analyses. Within the unified modeling framework, an atomic element (x) consists of four dimensions including spatial location (s), spatial scale (s'), temporal location (t), and temporal scale (t') [6]. Compared with prevalent GIS that focus on spatial analysis (i.e. $f(s)$) and single-scale spatio-temporal analysis (i.e. $f(s,t)$), the unified framework will fundamentally resolve the issue of multi-scale spatio-temporal analysis by providing 15 types of analyses using one or more of the four dimensions (i.e. $\binom{4}{1} + \binom{4}{2} + \binom{4}{3} + \binom{4}{4} = 15$).

References

- 1 Barbara Battenfield, Mehran Ghandehari, Stefan Leyk, Lawrence Stanislawski, Meg Brantley, and Yi Qiang. Measuring Distance “As the Horse Runs”: Cross-Scale Comparison of Terrain-Based Metrics. *International Conference on GIScience Short Paper Proceedings*, 1(1):37–41, 2014. doi:10.21433/B31118rh987cz.
- 2 Zenon Kulpa. Diagrammatic Representation of Interval Space in Proving Theorems about Interval Relations. *Reliable Computing*, 3(3):209–217, 1997. doi:10.1023/A:1009919304728.


- 3 Yi Qiang, Seyed H Chavoshi, Steven Logghe, Philippe De Maeyer, and Nico Van De Weghe. Multi-scale analysis of linear data in a two-dimensional space. *Information Visualization*, 13(3):248–265, 2014. doi:10.1177/1473871613477853.
- 4 Yi Qiang, Matthias Delafontaine, Katrin Asmussen, Birger Stichelbaut, Guy De Tré, Philippe De Maeyer, and Nico Van De Weghe. Modelling imperfect time intervals in a two-dimensional space. *Control and Cybernetics*, 39(4), 2010.
- 5 Yi Qiang, Matthias Delafontaine, Mathias Versichele, Philippe De Maeyer, and Nico Van de Weghe. Interactive Analysis of Time Intervals in a Two-Dimensional Space. *Information Visualization*, 11(4):255–272, 2012. doi:10.1177/1473871612436775.
- 6 Nico Van de Weghe, B. De Roo, Y. Qiang, M. Versichele, T. Neutens, and Philippe De Maeyer. The continuous spatio-temporal model (CSTM) as an exhaustive framework for multi-scale spatio-temporal analysis. *International Journal of Geographical Information Science*, 28(5):1047–1060, 2014. doi:10.1080/13658816.2014.886329.
- 7 Pengdong Zhang, Jasper Beernaerts, Long Zhang, and Nico Van de Weghe. Visual exploration of match performance based on football movement data using the Continuous Triangular Model. *Applied Geography*, 76:1–13, 2016. doi:10.1016/j.apgeog.2016.09.001.

Geosocial Media Data as Predictors in a GWR Application to Forecast Crime Hotspots

Alina Ristea

Department of Geoinformatics – Z_GIS, Doctoral College GIScience, University of Salzburg, Austria


mihaela-alina.ristea@sbg.ac.at

 <https://orcid.org/0000-0003-2682-1416>

Ourania Kounadi

Department of Geo-information Processing, Faculty of Geo-Information Science and Earth Observation, University of Twente, Enschede, Netherlands


o.kounadi@utwente.nl

 <https://orcid.org/0000-0002-5998-7343>

Michael Leitner

Department of Geography and Anthropology, Louisiana State University, Baton Rouge, LA, USA

mleitne@lsu.edu

 <https://orcid.org/0000-0002-1204-0822>

Abstract

In this paper we forecast hotspots of street crime in Portland, Oregon. Our approach uses geosocial media posts, which define the predictors in geographically weighted regression (GWR) models. We use two predictors that are both derived from Twitter data. The first one is the population at risk of being victim of street crime. The second one is the crime related tweets. These two predictors were used in GWR to create models that depict future street crime hotspots. The predicted hotspots enclosed more than 23% of the future street crimes in 1% of the study area and also outperformed the prediction efficiency of a baseline approach. Future work will focus on optimizing the prediction parameters and testing the applicability of this approach to other mobile crime types.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases spatial crime prediction, street crime, population at risk, geographically weighted regression, geosocial media

Digital Object Identifier 10.4230/LIPICs.GIScience.2018.56

Category Short Paper

Funding This research was funded by the Austrian Science Fund (FWF) through the Doctoral College GIScience at the University of Salzburg (DK W 1237-N23).

1 Introduction

Crime occurrences are complex phenomena studied from an interdisciplinary path, including criminology, law, psychology, geography, or economy. An important factor in understanding crime patterns relates to their spatial and temporal attributes. Some of the methods that have been used to explore these attributes in crime analysis are hot spot detection [7], spatial regression [8], retrospective forecasting, machine learning, near-repeat concept [10], and



© Alina Ristea, Ourania Kounadi, and Michael Leitner;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 56; pp. 56:1–56:7

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

risk terrain analysis [14]. However, many prediction models and their strategies are defined and modeled for places with disparate regional conditions. Also, crime types have different spatiotemporal distributions because they are affected by different factors. For example, robberies increase during nights and weekends [14], while assaults are frequent around liquor outlet areas on weekends [4]. The current study aims to forecast crime in three different future periods by considering GWR models from precedent similar periods. Additionally, we consider only street crimes and integrate information about their particular spatial and temporal patterns to predict areas where crimes are more likely to occur.

1.1 Predictors of crime & population at crime risk

Some elements of the build environment that are strongly correlated to crime and have been used for prediction include hospitals, schools, police stations, and population. Additionally, Twitter data have been integrated in crime analysis by considering their location [2], their topic [16, 9, 1], their sentiment [6], and by using a dictionary to select tweets that include specific words [15]. Regarding population information, census data are commonly used in calculating population at crime risk. However, population is not random in space and has varying patterns during working days and hours compared with home or leisure times. Hence, recent studies integrate dynamic population models. The ancillary dynamic information can be extracted from social media data [11, 12], mobile phone data [13], or spatial data and imagery analysis like LandScan Global Population Database, provided by Oak Ridge National Laboratory.

1.2 Research objective

The objective of this study is to integrate and test geosocial media data as variables in GWR crime prediction models. Geosocial media data are free and easier to obtain than authoritative data. In addition, they can be used to produce ambient models compared to the static nature of census data. Furthermore, there are at least two cases in which retrospective methods that require historical data of more than one past period cannot be used in prediction and thus regression-based approaches are promising alternatives. The first case is to estimate the crime occurrence in an area for which data are not available. An area with similar profile and availability of data can be used to deliver GWR regressions for the “unknown” area. The second case is when there are crime and predictor data for time $t-1$ and we want to estimate the crime prevalence in t by assuming that slight variations in time can be better represented in a generalized model of $t-1$ than the actual crimes of $t-1$.

2 A social media based GWR application for the prediction of crime

GWR is a modeling approach for spatially heterogeneous processes [3]. In the last decade, implementing GWR as a predictor increased substantially even if still controversial. GWR technique has the advantage of considering non-stationary variables and modeling local relationships between dependent and independent variables. Our strategy is to use recent past variables to create a model over the study area and predict crimes in the future. As for the parameters need to set for this tool, we used the “adaptive” bandwidth, as recommended in literature. For crime analysis, researchers are using a palette of interdisciplinary variables in regression models to understand crime distribution, not only spatially. We define explanatory variables from social data to examine, if being the only predictors in a model can favor the understanding of spatial crime distribution. Two types of variables are defined and tested:

Population at crime risk (PopCR) and crime-related tweets (CrimeTW). We used geolocated Twitter data in calculating PopCR, adapting the methodology described by Kounadi et al. 2017 for translating density of tweets into population density with the point-based areal interpolation method. Residential population is calculated for large geometries, such as census tracts or neighborhoods. Density-weighted interpolation disaggregates the data included in large geometries (source zones, i.e. residential population values) by using control point data (i.e. the spatial distribution of tweets) in order to obtain a new variable for target zones, which are smaller polygon geometries (i.e. grid cells). In addition, to define relevant geolocated tweets for our analysis, we first analyzed the temporal distribution of street crimes. Then, we chose the days and times of crime peaks and only for those timeframes, geolocated tweets were extracted and introduced in PopCR models. Second, we extracted CrimeTW by filtering the entire geolocated Twitter dataset for crime related terms. After preprocessing the data, we noticed four sources that post constantly about crimes: City of Portland 911 feed, City of Portland Fire/EMS feed, TTN POR traffic, Multnomah County Sheriff feed. The last source is an unofficial posting of the East County using a scanner feed from police information. Practically, we are using the intensity per polygon for these two independent variables, PopCR and CrimeTW, in order to explain the dependent crime counts. The analysis was performed for the three periods (one week, two months, three months), as well as for two cell sizes (0.006 km^2 small size called cell A and 0.023 km^2 large size called cell B).

3 Case study: Portland

The study area is in Portland, the largest city from the state of Oregon in the USA. The size of the study area is 382.6 km^2 and includes an estimated population of 640,000 people in 2016. Data for this case study contain crime occurrences in 2015 and 2016 from call-for-service data from the Portland Police Bureau, and Twitter data from 2015. We only consider street crime types that affect the mobile population, which include assault, disturbance, gang related crimes, robbery, shooting, stabbing, drugs, liquor, prostitution, and gambling. We tested three periods from the two years for which we downloaded the crime data: One week (1st week of March: 559 crimes in 2015 and 538 in 2016); two months (March to April: 5,129 crimes in 2015 and 5,386 in 2016); and three months (March to May: 7,987 crimes in 2015 and 8,417 in 2016). Twitter data were obtained using the Twitter API. We only used tweets that had the geolocation activated, so that we know the exact coordinates of the message. For the PopCR variable, we extracted the tweets from the three periods in 2015 and kept only those that showed a peak of crime events (e.g. weekend nights). Namely, we processed the temporal information from the tweets and we extracted the ones which have a corresponding time slot with crime at its peak (e.g. street crime type has temporal peaks during weekend nights, so we extracted the tweets from weekend nights to be control points for PopCR). For the second variable, CrimeTW, the entire filtered data set was used in all three periods (the tweets from the four aforementioned users).

4 Results

The analysis was performed for three time periods (one week, two months, three months), as well as for two cell sizes. The first size, called cell A, covers a rather small area of 0.006 km^2 (total number of cells: 66,841), while the second one, called cell B, covers an area of 0.023 km^2 (total number of cells: 16,753). We applied the analysis six times, one for each combination of time period and cell size, so that we can have a first exploration on the effects

■ **Table 1** Evaluation of GWR models (three prediction periods and two cell sizes).

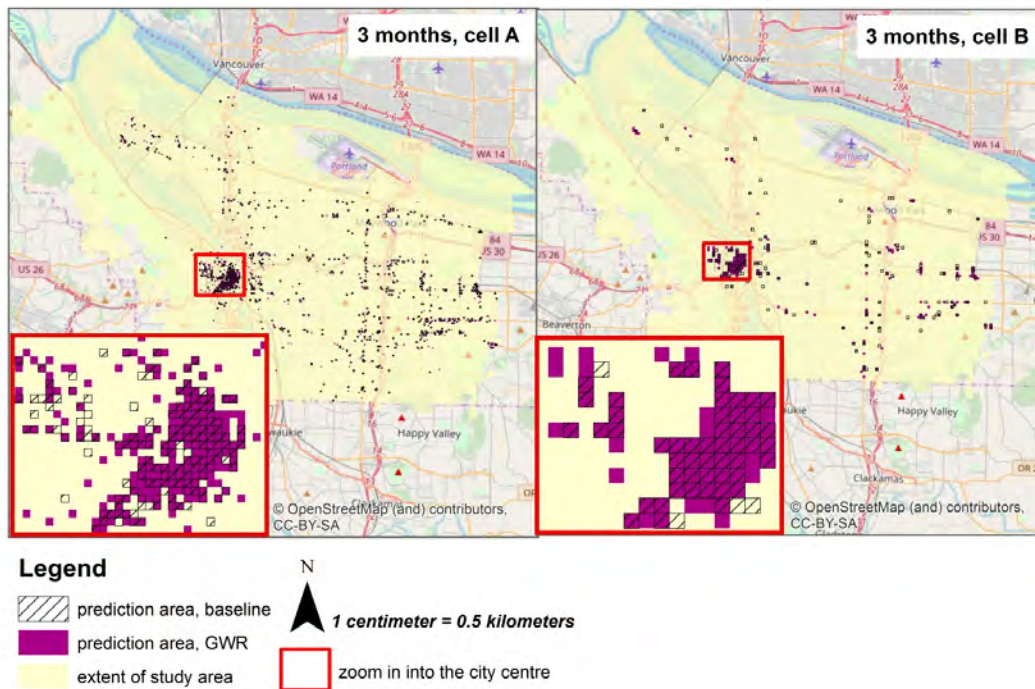
<i>Prediction period</i>	Cell A			Cell B		
	<i>1 week</i>	<i>2 months</i>	<i>3 months</i>	<i>1 week</i>	<i>2 months</i>	<i>3 months</i>
<i>AICc</i>	123,245.67	55,119.85	96,639.98	22,994.95	37,001.81	23,664.39
<i>R-squared</i>	0.13	0.44	0.49	0.54	0.67	0.61

of these parameters on the estimation models and prediction accuracy. Table 1 shows the Akaike Information Criteria (AICc) and R^2 values of the six GWR models. The AICc scored the lowest value for cell B and a three month period and R^2 has the highest value for cell B and a two month period. In general, we observe that larger cell sizes and longer time periods (two or three months) give considerably better results compared to smaller cell sizes and shorter prediction periods. The fact that different spatial aggregation chosen in analysis produce different results is one well-known issue in geography, named modifiable areal unit problem (MAUP), and in many cases the aggregation to a larger cell size can yield to better results. To identify hotspot areas in 2016 using data from 2015, we selected areas with high prediction values. Also, to compare the results among the models we standardized the size of the prediction area to approximately 1% of the total number of cells. The size of the total area is 382.6 km² and the size of the prediction area is 3.9 km². This amounts to 668 A cells and 168 B cells (selected based on their prediction values). Since the prediction area among models was the same and about 1, the denominator of the prediction accuracy index (PAI), which represents a common accuracy index in crime analysis [5], was canceled and thus we considered only the denominator, which is essentially the hit rate (i.e. success rate). Furthermore, we compared the GWR models with baseline models. As a baseline model we define a simple exploratory approach, where cells with the highest crime intensity in 2015 define hotspots in 2016. Again, the amount of cells was 668 A cells and 168 B cells so as to compare only hit rates. Figure 2 shows the results of one period for cells A and B and a zoomed-in section where we observe the divergence between baseline and GWR hotspots.

Street crimes in 2016 were used to calculate the hit rate. Although the prediction area was defined to be quite small, all GWR models resulted in a hit rate between 23.2% and 27.8% (Table 2). In particular, GWR models identified more than 20% of the crimes being located in 1% of the area. On the other hand the hit rate of the baseline models ranges between 12.4% and 28.8%. In Table 3, we provide a summary of the predictive efficiency analysis by calculating the mean hit rate by predictive period, cell size, and method. Although in Table 1 it is apparent that the larger cell size creates better GWR models, in terms of the prediction efficiency the smaller cell size predicts more crimes than the larger cell size (Table 3). Additionally, the larger the prediction period the better the results are. Finally, hotspot areas defined from GWR models predict a significantly higher number of crimes than areas defined from baseline models.

5 Discussion

The success of a prediction that employs spatial regression analysis depends on explanatory variables. In the current study we used variables from one main data source, Twitter, and we calculated how much of the spatial distribution of crimes is explained by tweets. We based our analysis on the 2015 information in order to get hotspots for 2016. Six GWR models were created using as dependent variables street crime data and independent variables from two tweet subsets (i.e. population at crime risk and crime related tweets) both from the year



■ **Figure 1** Overlapping of prediction areas resulting from GWR and baseline methods for a period of three months and two different cell sizes (cell A = 0.006 km² and cell B = 0.023 km²).

■ **Table 2** Predictive efficiency grouped by period, cell size, and method (GWR vs Baseline).

Prediction period	Cell size	Method	Hit Rate
1 week	A	GWR	25.4
		Baseline	13.5
	B	GWR	23.4
		Baseline	12.4
2 months	A	GWR	27.8
		Baseline	26.3
	B	GWR	23.8
		Baseline	14.1
3 months	A	GWR	27.6
		Baseline	28.8
	B	GWR	23.2
		Baseline	23.2

■ **Table 3** Average Hit Rate by the three parameters (predictive period, cell size, and method). * Indicates higher Hit Rate among comparisons.

Mean values of Hit Rate		
cell size	cell A*	24.9
	cell B	20.0
length of prediction period	1 week	18.7
	2 months	23.0
	3 months*	25.7
method	GWR*	25.2
	Baseline	19.7

2015. The predictive efficiency of GWR outcomes was higher than a baseline model that considered the past areas of high crime density as being the same for the next period. In order to account for effects of the spatial resolution and temporal differences, we selected three testing periods (one week, two months, and three months) and two different grid cell sizes, namely small and large. Results for the year 2015 show that by using the larger cell size the GWR models explain more variance of crime distribution patterns than by using the smaller cell size. However, when it comes to prediction efficiency the smaller cell size

yielded higher accuracy than the larger one. This may be a characteristic of the crime type in question and possibly high number of repeats and/or near-repeats. Also, the accuracy varies by prediction period with the longest analyzed period (i.e. three months) having the highest prediction efficiency. The main limitations of our approach is the under-representativeness of Twitter sample data (not each person is tweeting; not all users are using geolocation actively), the possibility of having non-reported crime occurrences that we did not evaluate, multicollinearity issues for GWR and the MAUP, which is not sufficiently addressed by two different cell sizes. To compensate for these limitations, our future work on this topic will employ the next three additions. First, we want to use additional types of social media platforms (e.g. Foursquare or Flickr) for the development of predictors. Second, we will perform multiple case studies for which the accuracy and completeness of crime data will be tested. Last, we will extensively and empirically test the parameters of our approach, including but not limited to the spatial resolution of the models and the temporal resolution of the prediction periods.

References

- 1 Meshrif Alruily. *Using text mining to identify crime patterns from arabic crime news report corpus*. Thesis, DeMontfort University, 2012. URL: <http://hdl.handle.net/2086/7584>.
- 2 Johannes Bendler, Tobias Brandt, Sebastian Wagner, and Dirk Neumann. Investigating crime-to-twitter relationships in urban environments-facilitating a virtual neighborhood watch. In *ECIS 2014*, 2014. URL: <http://aisel.aisnet.org/ecis2014/proceedings/track11/10/>.
- 3 Chris Brunsdon, A Stewart Fotheringham, and Martin E Charlton. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical analysis*, 28(4):281–298, 1996.
- 4 Joel M Caplan and Leslie W Kennedy. Risk terrain modeling compendium. *Rutgers Center on Public Security, Newark*, 2011.
- 5 Spencer Chainey, Lisa Tompson, and Sebastian Uhlig. The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal*, 21(1):4–28, 2008.
- 6 Xinyu Chen, Youngwoon Cho, and Suk young Jang. Crime prediction using twitter sentiment and weather. In *Systems and Information Engineering Design Symposium (SIEDS), 2015*, pages 63–68. IEEE, 2015.
- 7 JE Eck, S Chainey, JG Cameron, M Leitner, and RE Wilson. Mapping crime: Understanding hot spots. national institute of justice. *Washington, DC*, 2005.
- 8 A Stewart Fotheringham, Chris Brunsdon, and Martin Charlton. *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons, 2003.
- 9 Matthew S Gerber. Predicting crime using twitter and kernel density estimation. *Decision Support Systems*, 61:115–125, 2014. doi:10.1016/j.dss.2014.02.003.
- 10 Philip Glasner and Michael Leitner. Evaluating the impact the weekday has on near-repeat victimization: A spatio-temporal analysis of street robberies in the city of vienna, austria. *ISPRS International Journal of Geo-Information*, 6(1):3, 2016.
- 11 Ourania Kounadi, Alina Ristea, Michael Leitner, and Chad Langford. Population at risk: using areal interpolation and twitter messages to create population models for burglaries and robberies. *Cartography and Geographic Information Science*, pages 1–15, 2017. doi:15230406.2017.1304243.
- 12 Nick Malleon and Martin A Andresen. The impact of using social media data in crime rate calculations: shifting hot spots and changing spatial patterns. *Cartography and Geographic Information Science*, 42(2):112–121, 2015. doi:10.1080/15230406.2014.905756.


- 13 Nick Malleson and Martin A Andresen. Exploring the impact of ambient population measures on london crime hotspots. *Journal of Criminal Justice*, 46:52–63, 2016.
- 14 Walt L Perry. *Predictive policing: The role of crime forecasting in law enforcement operations*. Rand Corporation, 2013.
- 15 Alina Ristea, Justin Kurland, Bernd Resch, Michael Leitner, and Chad Langford. Estimating the spatial distribution of crime events around a football stadium from georeferenced tweets. *ISPRS International Journal of Geo-Information*, 7(2):43, 2018. URL: <http://www.mdpi.com/2220-9964/7/2/43>.
- 16 Xiaofeng Wang, Matthew S Gerber, and Donald E Brown. Automatic crime prediction using events extracted from twitter posts. In *International conference on social computing, behavioral-cultural modeling, and prediction*, pages 231–238. Springer, 2012.

Who Masks? Correlates of Individual Location-Masking Behavior in an Online Survey

Dara E. Seidl

Department of Geography, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182, USA


dseidl@sdsu.edu

 <https://orcid.org/0000-0001-8737-7115>

Piotr Jankowski

Department of Geography, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182, USA and Institute of Geocology and Geoinformation, Adam Mickiewicz University, Poznań, Poland

pjankows@sdsu.edu

 <https://orcid.org/0000-0002-6303-6217>

Abstract

Geomasking traditionally refers to a set of techniques employed by a data steward to protect the privacy of data subjects by altering geographic coordinates. Data subjects themselves may make efforts to obfuscate their location data and protect their geoprivacy. Among these individual-level strategies are providing incorrect address data, limiting the precision of address data, or map-based location masking. This study examines the prevalence of these three location-masking behaviors in an online survey of California residents, finding that such behavior takes place across social groups. There are no significant differences across income level, education, ethnicity, sex, and urban locations. Instead, the primary differences are linked to intervening variables of knowledge and attitudes about location privacy.

2012 ACM Subject Classification Security and privacy → Human and societal aspects of security and privacy

Keywords and phrases privacy, geoprivacy, geomasking, obfuscation, accuracy

Digital Object Identifier 10.4230/LIPICs.GIScience.2018.57

Category Short Paper

Funding This material is based upon work supported by the National Science Foundation under Grant No. 1657610. This work was supported in part by a American Association of Geographers (AAG) dissertation grant and an American Geographical Society (AGS) fellowship.

1 Introduction

While a large body of research is dedicated to protecting the privacy of human subjects, there has been less documentation on the efforts of individuals to protect their own privacy. The set of procedures known as geomasking typically refers to the alteration of point data to protect both spatial distributions and privacy of data subjects [2]. Common geomasking techniques include random perturbation [6], donut masking [4], and grid masking [12]. The typical use scenario for these top-down strategies is for researchers who wish to share geospatial data with others, but must protect privacy. Masking behavior at an individual level, such as by responding to location requests with false or imprecise address data, can also serve to protect an individual's geoprivacy. This study tests the correlates of bottom-up or individual-level



© Dara E. Seidl and Piotr Jankowski;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 57; pp. 57:1–57:6



Leibniz International Proceedings in Informatics

LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

location masking in an online survey, finding that intervening variables of hacking exposure, social media use, and geoprivacy knowledge and attitudes are significantly correlated with masking behavior.

1.1 Related work

In their geoprivacy manifesto, [5] argue that location privacy stands apart from information privacy, in part because of the range of inferences that can be drawn from location, the ubiquity of location-collecting technology, and the incentives which draw consumers to share their locations. Compared to information privacy, which has been reported on by [8] and [1], not much is known about specific geoprivacy-related attitudes and behaviors. Obfuscation at the individual level is characterized as an act of resistance to surveillance [3], an idea seconded by [13] and [14], who argue that Tor, an onion routing technology that masks location by altering IP address, is a prime example of resistance to geosurveillance. Compared to the technologically-advanced location masking of Tor, this study focuses on the masking behavior internet users exhibit when faced with the explicit location request: “Please enter your home location.” Both the precision and participant-reported truthfulness of entered location are collected as outcome variables in determining “who masks”.

2 Methods

An online survey testing location masking behavior was deployed to California residents between October 2017 and March 2018. Participants were drawn from two samples: a random address-based sample obtained from Survey Sampling International (SSI) and contacted by postal mail, and a non-probability online open sample, reached by paid ad placement on Facebook and free advertising on Craigslist. A primary concern in the survey design was to avoid social desirability bias, which results in inflated privacy concerns by participants in studies advertised as privacy-related [11]. Therefore, this survey was designed to omit use of the word “privacy” and to capture location masking as it might occur in a routine online setting. Participants were told they were participating in a study about “online information sharing” and were debriefed about the true purpose of the study at its conclusion, at which time, they were also given the option to withdraw their responses.

2.1 Conceptualization

This study follows a knowledge-attitudes-behavior framework to predict participant location masking, a model commonly used to predict behavioral outcomes in health and environmental studies [9][7]. Hypothesized background variables included age, education, sex, income, ethnicity, and urban location. Given that previous negative privacy experience online increases perceived risk of sharing on social media [15], hypothesized intervening variables included recent identity theft or hacking, social media use, and employment experience with personal data. It was hypothesized that location masking behavior would be most closely correlated with high geoprivacy knowledge and concern for geoprivacy. Each of these variables was measured in a series of Likert-type questions in the survey.

2.2 Survey design

The primary test of location masking was participants’ response to “please enter your home location,” for which they were given text boxes for street, cross street, city, state, and zip code. If respondents entered a text-based location, they would then have the option to open

■ **Table 1** Differences between mail and online sample in Mann-Whitney U tests for background variables (* $p < 0.05$).

Variable	Mail Sample	Online Sample	Sig
Female	55%	76%	*
White	66%	55%	
College degree	69%	44%	*
Median age group	45-54	25-34	*
Median income tax bracket	25% (38,000–92,000)	15% (9,000–38,000)	*
Somewhat or very urban	62%	56%	
Total participants	113	101	

up a map and adjust a pin to their chosen coordinates. By default, the map pin was placed at the geocoded coordinates of the entered street address with the Google geocoding API. Respondents then selected their level of agreement on a five-point Likert scale (strongly disagree to strongly agree) to the statements, “I intentionally provided incorrect information on my home location” and “I intentionally moved the pin on the map away from my home location.” The remainder of the survey tested geoprivacy knowledge, attitudes, and the other background variables with similar Likert-type items, asking participants to respond with their level of agreement. The survey was hosted on the Qualtrics platform and fully encrypted.

2.3 Analysis

Differences between the two samples were analyzed with Mann-Whitney U tests, a non-parametric test for differences between two categorical variables [11]. Due to the ordinal nature of the majority of the study variables, Spearman’s correlations were calculated between each of the variables and tested for significance [10]. To determine geographical patterns, global and local Moran’s I were applied as tests of spatial autocorrelation for survey participation rates, location masking behavior, and geoprivacy-related attitudes.

3 Results

The questionnaire had a total of 214 respondents with 113 in the mail sample and 101 in the online open sample. The two samples differed significantly in age, income, education level, and gender composition, based on Mann-Whitney U tests (Table 1). The online open sample was more female, younger, and had lower education levels and incomes compared to the mail-based sample. The mail sample self-reported on average as more urban, though this did not reach significance. The mail sample was also significantly more likely to have employment experience working with personal data. In terms of location masking, the online sample was significantly less likely to provide a numbered street address for home location ($p < 0.05$), compared to the mail sample, although the majority of participants in both cases provided home location at this highest precision (73% of mail sample respondents and 56% of open sample respondents). When it came to factuality of reported home location, however, there were no significant differences between the two samples (Figure 1). About 15% of respondents somewhat or strongly agreed that they intentionally provided an incorrect home address, and 11% of respondents who interacted with the map function agreed that they intentionally moved the pin away from their home location.

When tested with global Moran’s I, there was no global clustering of the respondents from the two samples at the county level when normalized by population. This suggests that a randomly distributed sample was achieved in both cases. Location masking behavior was

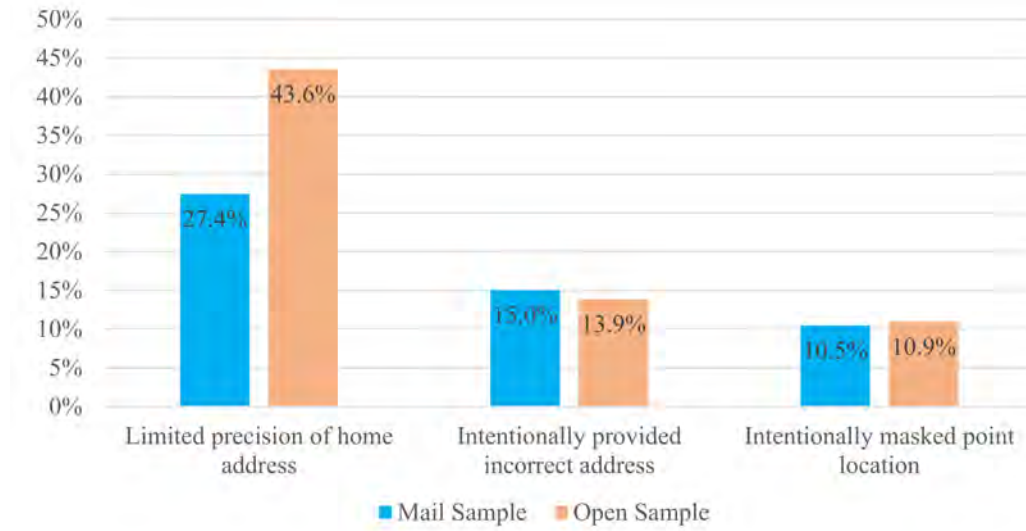


Figure 1 Results by sample for three location masking behaviors.

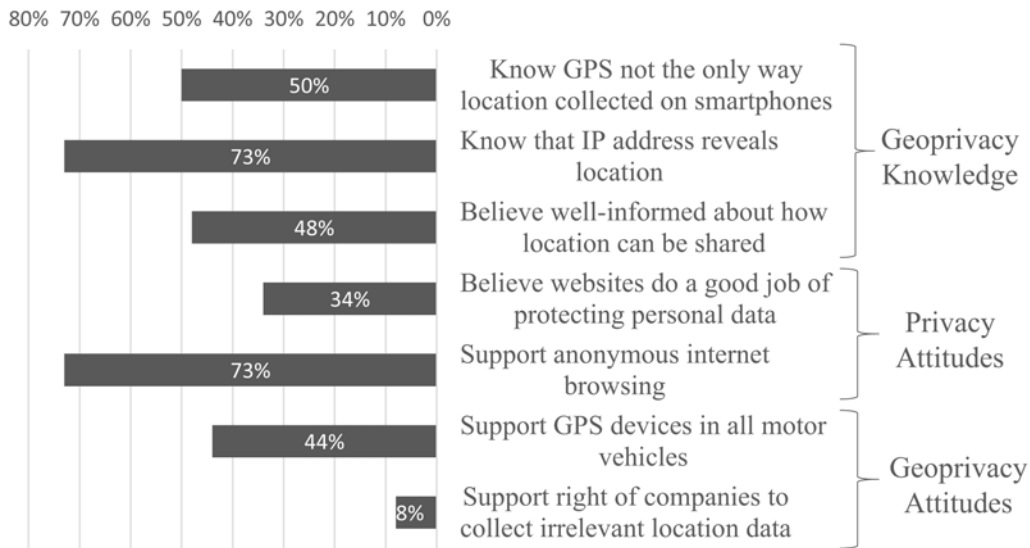
not globally clustered when tested with Moran's I, however, two of the attitude variables, trust in websites to protect personal data and support of GPS devices in all vehicles, were globally and locally clustered ($p < 0.05$).

Overall knowledge about location privacy was low to average, with just 50% aware that smartphones collect location outside of GPS, and 73% aware that IP address reveals location (Figure 2). Self-reported knowledge about how location is shared was also low, with 48% believing themselves to be well-informed. The attitude results demonstrated overall concern for privacy, with only 34% believing websites to do a good job of protecting personal data, and just 8% supporting the right of companies to collect irrelevant location data.

The Spearman's tests (Table 2) revealed that no demographic background variables were significantly correlated with the three indicators of location masking. Location precision had the highest frequency of significant correlates. Respondents were more likely to mask location by providing lower address precision if they were part of the open sample, if they had a recent hacking experience, if they had more knowledge about smartphone location collection, and if they did not trust websites to protect their personal data. Lower precision was also correlated with other masking behavior, including use of technology to alter IP address and provision of incorrect address information to retailers. The two intentional masking outcome variables were not correlated with knowledge or attitudes, but again with other location masking behaviors. Enjoyment of social media was the one intervening variable significantly correlated with providing accurate home location.

4 Conclusion

With 15% of participants admitting to providing incorrect address information, location masking behavior is a small but present minority among participants, and it takes place across demographic lines. The precision of location respondents provide appears to be dependent on context, trust, and knowledge, rather than background variables. The open online sample, respondents who do not trust websites to protect their personal data, and respondents who know that location can be collected in smartphones outside of GPS were more likely to



■ **Figure 2** Percent of participants exhibiting geoprivacy-related knowledge and attitudes.

■ **Table 2** Spearman’s rho between predictor variables and location masking behavior. Only significant correlations shown ($p < 0.05$).

	Correlates	Provided higher home location precision	Intentionally provided incorrect home location	Intentionally moved pin away from home location
Background	Sample (1=Mail Sample, 2=Open Sample)	-0.169		
Intervening	Enjoy contributing to social media		-0.227	
	Had unauthorized user on online account	-0.155		
Knowledge	Believe GPS only way location collected on smartphone	0.139		
Attitudes	Believe websites do a good job of protecting personal data	0.238		
Other masking behavior	Use technology to alter IP address	-0.169		
	Give inaccurate or misleading address information to retailers	-0.194	0.227	
	Turn location services off on smartphone			0.194
	Intentionally provided incorrect home location			0.406

provide a lower precision of home address. Location masking measured as truthfulness of location has fewer clear correlations with the hypothesized background variables than location precision does, but is significantly correlated with other location masking behaviors and lower enthusiasm for social media. The results demonstrate that in California, a U.S. state with a large high-tech sector, there is still relatively limited exercise of geoprivacy protection measures at an individual level.

References

- 1 Alessandro Acquisti and Jens Grossklags. Privacy attitudes and privacy behavior. In *Economics of information security*, pages 165–178. Springer, 2004.
- 2 Marc P Armstrong, Gerard Rushton, Dale L Zimmerman, et al. Geographically masking health data to preserve confidentiality. *Statistics in medicine*, 18(5):497–525, 1999.
- 3 Finn Brunton and Helen Nissenbaum. *Obfuscation: A user's guide for privacy and protest*. Mit Press, 2015.
- 4 Kristen H Hampton, Molly K Fitch, William B Allshouse, Irene A Doherty, Dionne C Gesink, Peter A Leone, Marc L Serre, and William C Miller. Mapping health data: improved privacy protection with donut method geomasking. *American journal of epidemiology*, 172(9):1062–1069, 2010.
- 5 Carsten Keßler and Grant McKenzie. A geoprivacy manifesto. *Transactions in GIS*, 22(1):3–19, 2018.
- 6 Mei-Po Kwan, Irene Casas, and Ben Schmitz. Protection of geoprivacy and accuracy of spatial information: how effective are geographical masks? *Cartographica: The International Journal for Geographic Information and Geovisualization*, 39(2):15–28, 2004.
- 7 Debra Siegel Levine and Michael J Strube. Environmental attitudes, knowledge, intentions and behaviors among college students. *The Journal of social psychology*, 152(3):308–326, 2012.
- 8 Mary Madden and Lee Rainie. *Americans' attitudes about privacy, security and surveillance*. Pew Research Center, 2015.
- 9 Susan Morgan and Jenny Miller. Communicating about gifts of life: The effect of knowledge, attitudes, and altruism on behavior and behavioral intentions regarding organ donation. *Journal of Applied Communication Research*, 30(2):163–178, 2002.
- 10 Susan A Nolan and Thomas Heinzen. *Essentials of statistics for the behavioral sciences*. Macmillan, 2010.
- 11 Erin Ruel, William Edward Wagner III, and Brian Joseph Gillespie. *The practice of survey research*. Sage, 2015.
- 12 Dara E Seidl, Gernot Paulus, Piotr Jankowski, and Melanie Regenfelder. Spatial obfuscation methods for privacy protection of household-level data. *Applied Geography*, 63:253–263, 2015.
- 13 David Swanlund and Nadine Schuurman. Mechanism matters: Data production for geosurveillance. *Annals of the American Association of Geographers*, 106(5):1063–1078, 2016.
- 14 David Swanlund and Nadine Schuurman. Resisting geosurveillance: A survey of tactics and strategies for spatial privacy. *Progress in Human Geography*, 2018.
- 15 Hongwei Yang and Hui Liu. Prior negative experience of online disclosure, privacy concerns, and regulatory support in chinese social media. *Chinese Journal of Communication*, 7(1):40–59, 2014.

Dynamically-Spaced Geo-Grid Segmentation for Weighted Point Sampling on a Polygon Map Layer

Kelly Sims

Oak Ridge National Laboratory, One Bethel Valley Rd, Oak Ridge, TN 37831, U.S.A.
simskm@ornl.gov

Gautam Thakur

Oak Ridge National Laboratory, One Bethel Valley Rd, Oak Ridge, TN 37831, U.S.A.

Kevin Sparks

Oak Ridge National Laboratory, One Bethel Valley Rd, Oak Ridge, TN 37831, U.S.A.

Marie Urban

Oak Ridge National Laboratory, One Bethel Valley Rd, Oak Ridge, TN 37831, U.S.A.

Amy Rose

Oak Ridge National Laboratory, One Bethel Valley Rd, Oak Ridge, TN 37831, U.S.A.

Robert Stewart

Oak Ridge National Laboratory, One Bethel Valley Rd, Oak Ridge, TN 37831, U.S.A.

Abstract

Geo-grid algorithms divide a large polygon area into several smaller polygons, which are important for studying or executing a set of operations on underlying topological features of a map. The current geo-grid algorithms divide a large polygon in to a set of smaller but equal size polygons only (e.g. is ArcMaps Fishnet). The time to create a geo-grid is typically proportional to number of smaller polygons created. This raises two problems - (i) They cannot skip unwanted areas (such as water bodies, given about 71% percent of the Earth's surface is water-covered); (ii) They are incognizant to any underlying feature set that requires more deliberation. In this work, we propose a novel dynamically spaced geo-grid segmentation algorithm that overcomes these challenges and provides a computationally optimal output for borderline cases of an uneven polygon. Our method uses an underlying topological feature of population distributions, from the LandScan Global 2016 dataset, for creating grids as a function of these weighted features. We benchmark our results against available algorithms and found our approach improves geo-grid creation. Later on, we demonstrate the proposed approach is more effective in harvesting Points of Interest data from a crowd-sourced platform.

2012 ACM Subject Classification Theory of computation → Divide and conquer

Keywords and phrases geofence, geo-grid, quadtree, points of interest (POI), volunteered geographic information (VGI)

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.58

Category Short Paper

1 Introduction

Obtaining land use data at a global scale can be arduous with respect to data availability, coverage, resolution, accuracy, computational power and storage. The emergence of Volunteered Geographic Information (VGI) exploited from open sourced platforms, however,



© Kelly Sims, Gautam Thakur, Kevin Sparks, Marie Urban, Amy Rose, and Robert Stewart; licensed under Creative Commons License CC-BY

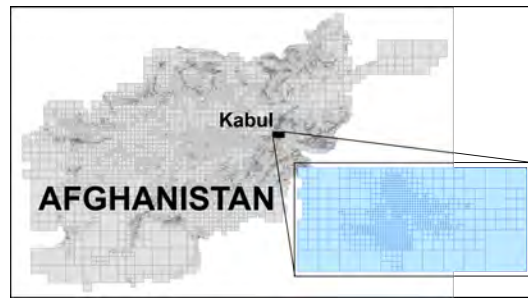
10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 58; pp. 58:1–58:7

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



■ **Figure 1** An example of dynamically-spaced geo-grid segmentation for weighted point sampling on a polygon map layer.

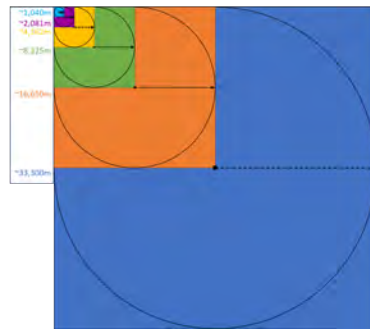
can provide rich attributes world-wide from Points Of Interest (POI) data. This data can facilitate many mapping and modeling techniques related to human dynamics, which is vital to emergency management and response, urban planning, and energy use. In order to optimize the collection of such available data, it is a recommended practice to segment a given geographical region into smaller grid cells for a more focused collection effort. ArcMap's fishnet[5], and other spatial indexing techniques that implement triangular mesh gridding are commonly used geoprocessing tools for this geo-grid segmentation. However, these tools are agnostic to the underlying topology and evenly segments areas into equally sized cells and/or triangular facets. Such approaches cannot skip unwanted regions (such as oceans) and their computational time is proportional to the size of individual cells, eventually, it consumes needless time and computational resources.

These challenges led us to develop a new approach for geo-grid segmentation that make use of underlying topology such as population distribution, building settlement extractions, etc. The proposed method creates dynamically spaced geo-grids as a function of underlying topological data. For example, as shown in Figure-1, using population distribution data to influence the location for request calls, the proposed method generates bigger cells for sparsely populated areas, or several smaller cells for densely populated regions.

2 Related Work

Multiple algorithms have been developed to generate a geocoded index to describe an exact or general location on Earth. In fact, many of these methods implement a hierarchical dissecting system that follows a tree structure concept to index places [7], or use a reverse geocoding practice of interpreting actual latitude/longitude coordinates to a single array [1], or develop a continuous map at a specific spatial unit labeled with new and random naming conventions [6] [8] [10]. Other approaches use a triangular mesh grid for analyzing geographic data within equally sized facets with minimal distortion [4] [3], or to identify coverage areas for database retrieval purposes [9] [11].

While the previously mentioned practices influenced this research, our approach is not to create a universal addressing system or a new projected referencing system. Instead, we propose a gridding system partitioned by a given topological requirement to help collect data from open source platform graph APIs that require a lat/long and search radius. Specifically, we produced a hierarchically gridded map of the world based on the underlying population within each grid to maximize our retrieval of Points of Interest (POI) data from VGI. By considering population distributions to subdivide a grid, we produce a map that can allocate additional computational resources where higher populations reside.



■ **Figure 2** GeoHashed Grid Sizes with their respective search radii.

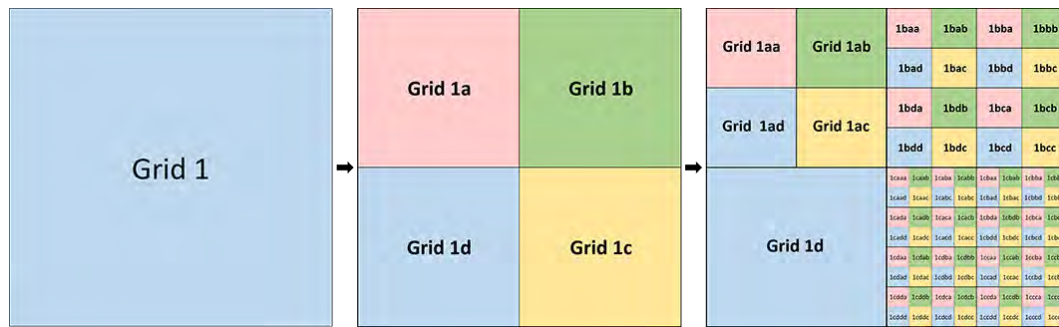
3 Methodology

In this section, we discuss the proposed method for creating dynamically-spaced geo-grid segmentation for weighted point sampling on a polygon map layer. And with our efforts focused on optimizing the number of POIs fundamental to human dynamics, we assume population is an indicator of places. We use the LandScan Global population dataset [2], developed at Oak Ridge National Laboratory (ORNL), for weighted point sampling purposes. This model depicts an “ambient” population distribution (average over 24 hours) at 30 arc-seconds resolution (roughly 1km at the equator).

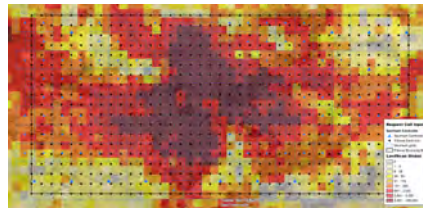
3.1 Spatial Analysis

We began by gridding the world in “.6 x .6” degree increments for two reasons. The first reason was to begin with a small enough grid size so that the largest possible search radius (from the center of a grid to the nearest edge) would be less than 50,000 meters. With that said, a .6 degree grid at the equator has a radius of roughly 33,300 meters. The rationale for not using a search radius just under 50,000m (which would actually be a .9 degree grid at the equator), leads us to our second reasoning. Our geohashing technique splits each grid into 4 equal parts, until the last grid becomes smaller than a LandScan Global cell size, which is ~1km at the equator. When a geohashed grid is smaller than a LandScan Global cell, we can no longer sum the population within the single grid without missing at least one LandScan Global cell centroid. By using a .6 degree grid initially, we can geohash until the second to last geohashed grid is roughly 1% bigger than a LandScan Global population grid, allowing us to sum the population one last time. If the threshold is still not met, we can then geohash one final time to make the smallest request grid possible based on population (see Figure 2).

The population threshold for this study was 5,000. Therefore, when a grid’s underlying population sum was above 5,000 people, subdividing took place into 4 new quadrants. From the top-left quadrant clockwise, we labeled the four quadrants A, B, C, and D, respectively. These quadrant labels were then appended to the original ID of the grid being GeoHashed. For example, Figure 3 represents the iterations of geohashing grids and the product of each new quadrant’s ID. Grid 1 has a population over 5,000 and is replaced with four new grids labeled Grid 1a, Grid 1b, Grid 1c, and Grid 1d. When necessary, these divisions continue through each new quadrant(s) (Grid 1a, 1b, and 1c) constantly replacing the previous grid with new appropriately sized grids. If a grid’s summed population is under 5,000, the iterations end and no additional divisions occur (Grid 1d). The rest of the section discusses the algorithms aspect of our method.



■ **Figure 3** An example of how Grids are geohashed and labeled.



■ **Figure 4** Request points calculated through proposed algorithms and a standard 1km fishnet grid over Kabul, Afghanistan. The background is LandScan Global's population distribution.

3.2 Algorithm

The spatial analysis algorithm performs two important tasks that results in deciding whether to split the current grid cell in equal size small sub-grid cells (or move to next grid cell. This decision is made based on the population count in the current grid cell. If the population count is higher than the threshold, the algorithm (Algorithm:1) segments the grid cell. In this algorithm, first we calculate the extent of the current grid cell. An extent defines the geographic boundaries that contains a population data frame. These boundaries contain top, bottom, left, and right coordinates, which are the edges of the map extent. For the purpose of this work, we rely on fixed extent calculation of the cell. Later, we calculate the total grid cell population count from the raster centroid of this grid cell (Algorithm: 2). The algorithm returns the value of population count that main algorithm use to decide whether to segment or maintain the current extent of the cell. In this algorithm, parameters are passed as reference to calculate the values of grid cell extent and cell population count.

4 Application and Summary of Results

To showcase the usability of our proposed method, we curated Points of Interest (POI) data over Kabul, Afghanistan. Influenced by population distributions from LandScan Global, we generated dynamically-spaced geo-grid cells for weighted point sampling on Kabul's polygon map layer. We then benchmarked this method against the traditional 1-km fishnet generated from ESRI ArcGIS and quantified the overall performance on three different measures. These measures included - (i) total number of POIs curated; (ii) total number of requests made to collect the POIs; and (iii) duration to collect.

4.1 Approach

A Point of Interest (POI) is a feature on a map that has a unique latitude and longitude coordinate. Some examples include - church, school, and hospital. Several mapping sites, such as OpenStreetMap, provide APIs to search POIs in a vicinity. These APIs input a

Algorithm 1: Algorithm to calculate grids using LandScan Global.

```

Function CalculateGeoHash(inputGrid, landScan_global, threshold)
  Data: LandScan Global raster population layer, grid, and threshold values
  Result: grid cells
  /* Split the original grid in 0.6° grid blocks. The radius from
    center is 33,000 mtrs at the equator. API threshold 50,000 mtrs
    */
  intpuGrid ← calculate_fishnet(inputGrid, 0.6°) ;
  overlay_landscan_global();
  while true do
    /* Traverse each grid block in sequential fashion */
    RunSpatialAnalysis(gridId, GeoPoint geoPoint[4], long populationCount) ;
    if populationCount > 5000 then
      Split the cell in to four equal parts;
      replace old cell with four new cells ;
      if current_grid == len(grid) then
        | break();
      end
    end
    else
      | continue; /* move to next grid cell */
    end
  end
end

```

geocoordinate with a radius, and returns POIs in that extent. Mapping websites impose a limit on the number of POIs returned, so its vital to keep a small radius. However, that increases the scanning time and number of API calls needed to maximize POIs collection. We assume population distribution dictates the distribution of POIs. A densely populated area may have more POIs than a sparsely populated area. In this experiment, we attempt to collect POIs for Kabul, Afghanistan. As shown in Figure-4, we have generated a set of 740 request points and distance between adjacent points using the proposed method and 790 request points at 1-km distance using the ArcMap fishnet algorithm. Next, we will make API calls using these two request points dataset and compare the results of total POIs collected.

4.2 Results Discussion

The proposed geohash approach outperforms the traditional 1-km fishnet approach on all three measures. As shown in Figure-5a, when the proposed approach request points dataset was used, it took us 3520 seconds to collect 2548 POIs using 741 API call. When the 1-k dataset was used, it took us 3990 seconds to collect only 2495 POIs using 790 calls. These numbers become significant for large scanning area such as at country scale. In Figure-5b, we have shown the distribution of collected POIs from the two datasets.

4.3 Limitations and Future Work

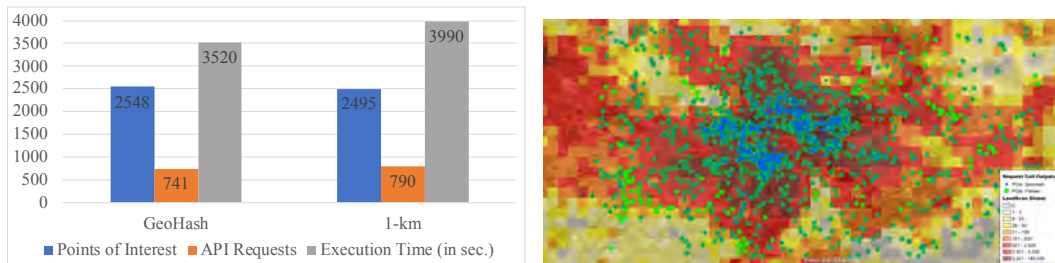
One limitation to our proposed method is our use of LandScan Global, which restricts our segmentation abilities considering its resolution of 30 arc-seconds. In the future, we plan to implement LandScan HD, also developed at ORNL, because of its resolution at 3 arc-seconds

Algorithm 2: Algorithm to calculate grid cell extent and population.

```

Function RunSpatialAnalysis(gridId, GeoPoint geoPoint[4], long
populationCount)
  Data: LandScan Global raster population layer, grid, and threshold values
  Result: grid cell extent, populationCount of the grid cell
  if cell !=null then
    /* Genrate coordinate values for top_left, top_right,
      bottom_left, bottom_right corners of the grid cell          */
    geoPoint ← calculate_boundaries_spatial_extent();
    /* Calculate population count for this grid cell                */
    populationCount ← extract_landscan_spatial_statistics(geoPoint) ;
  end
end

```



(a) The histogram shows the distribution of Points of Interest curated using the two methods, time, and the total number of request made to the server. (b) Curated Points of Interest from two different methods, overlaying LandScan Global's population distribution.

■ **Figure 5** Results of Points of Interest curation.

(~90m at the equator). With a more spatially refined population distribution, our model can be partitioned further with an ending grid size just shy of 90 square meters, instead of the current 600 meters. Furthermore, the processing time to develop our dynamically spaced geo-grids will also need to be improved. Countries with wide-spread population distributions, like India and Pakistan, produced hundreds of thousands of grids that were appended after each iteration, thus slowing down the overall processing time. Future work will explore how to improve this workflow as it will be necessary when we refine our input measurements to analyze population distributions at 3 arc-seconds.

5 Conclusion

In this paper, we proposed a method for gridding the world into varying geofenced grids based on a given measurement threshold to optimize search requests against social media APIs. While our research used ORNL's LandScan Global population dataset to designate the requirements, this algorithm can be augmented for other geospatial analysis and with other datasets. In fact, raster and point data are best suited with this method for generalizing or identifying areas of interest in vector data at multiple spatial representations. For example, crime data can be used to provide emergency personnel a map for allocating resource coverage. While, we recognize this method of geohashing is time consuming on the front end, we have observed an increase in the amount of data exploited, thus validating the necessity up front.

Copyright. This manuscript has been authored by employees of UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy. Accordingly, the United States Government retains, and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

References

- 1 Ziyad S. AL-Salloum. What Is Your Makaney Code?, 2011. URL: <http://www.makaney.net/>.
- 2 Budhendra Bhaduri, Edward Bright, Phillip Coleman, and Jerome Dobson. *Landscan. Geoinformatics*, 5(2):34–37, 2002.
- 3 Jun Chen, Xuesheng Zhao, and Zhilin Li. An algorithm for the generation of voronoi diagrams on the sphere based on qtm. *Photogrammetric Engineering & Remote Sensing*, 69(1):79–89, 2003.
- 4 Geoffrey Dutton. Encoding and handling geospatial data with hierarchical triangular meshes. In *Proceeding of 7th International symposium on spatial data handling*, volume 43. Citeseer, 1996.
- 5 ESRI. Create Fishnet—Data Management toolbox | ArcGIS Desktop. URL: <http://pro.arcgis.com/en/pro-app/tool-reference/data-management/create-fishnet.htm>.
- 6 Google. Google Plus Codes. URL: <https://plus.codes/>.
- 7 Gustavo Niemeyer. Geohash - Wikipedia, 2008. URL: <https://en.wikipedia.org/wiki/Geohash>.
- 8 Open Street Maps. DE:Browsing - OpenStreetMap Wiki. URL: <https://wiki.openstreetmap.org/wiki/QuadTiles><http://wiki.openstreetmap.org/wiki/DE:Browsing>.
- 9 Patrik Ottoson and Hans Hauska. Ellipsoidal quadtrees for indexing of global geographical data. *International Journal of Geographical Information Science*, 16(3):213–226, 2002.
- 10 Chris Sheldrick, Jack Waley-Cohen, Mohan Ganesalingam, and Michael Dent. what3words. URL: <http://what3words.com>/<https://map.what3words.com/palace.things.talking>.
- 11 Kentaro Toyama, Ron Logan, and Asta Roseway. Geographic location tags on digital images. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 156–166. ACM, 2003.

The Landform Reference Ontology (LFRO): A Foundation for Exploring Linguistic and Geospatial Conceptualization of Landforms

Gaurav Sinha

Department of Geography, Ohio University, Athens, Ohio, USA
sinhag@ohio.edu

Samantha T. Arundel

US Geological Survey, Center of Excellence for Geospatial Information Science, Rolla, MO, USA
sarundel@usgs.gov

Torsten Hahmann

School of Computing and Information Sciences, University of Maine, Orono, Maine, USA
torsten.hahmann@maine.edu

E. Lynn Usery

US Geological Survey, Center of Excellence for Geospatial Information Science, Rolla, MO, USA
usery@usgs.gov

Kathleen Stewart

Department of Geographical Sciences, University of Maryland, College Park, Maryland, USA
stewartk@umd.edu

David M. Mark

(Emeritus) Department of Geography, The University at Buffalo, Amherst, New York, USA
dmark@buffalo.edu

Abstract

The landform reference ontology (LFRO) formalizes ontological distinctions underlying naïve geographic cognition and reasoning about landforms. The LFRO taxonomy is currently based only on form-based distinctions. In this significantly revised version, several new categories have been added to explicate ontological distinctions related to material-spatial dependence and physical support. Nuances of common natural language landform terms and implications for their mapping are discussed.

2012 ACM Subject Classification Information systems → Ontologies

Keywords and phrases landform, reference ontology, terrain reasoning, dependence, support

Digital Object Identifier 10.4230/LIPICs.GIScience.2018.59

Category Short Paper

Acknowledgements “Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.” The authors acknowledge feedback from three anonymous and two USGS reviewers.



© Gaurav Sinha, Samantha T. Arundel, Torsten Hahmann, E. Lynn Usery, Kathleen Stewart, and David M. Mark;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).
Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 59; pp. 59:1–59:7



Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Motivation and background

The Landform Reference Ontology (LFRO) is being developed as a domain reference ontology for knowledge representation and reasoning about landforms. Its immediate purpose is to guide automated landform mapping from imagery data, but it is carefully being designed as a more generally applicable reference ontology, independent of any specific culture, language, or scientific discipline. LFRO is *not* being proposed as a universal ontology. Landforms have been previously argued to be mind-dependent or fiat [7] or quasi-objects [1] because their demarcation and categorization is not independent of human cognition. Ethnophysiographic research clearly shows there are many alternative ways of describing the domain of landforms.[9] As one extreme case, in the Lokono language, there is only one scale and size independent general term *horhorho* for landforms, and all distinctions are made through a complex vocabulary of lexical phrases that classify landforms as networks of connected places.[11]

While LFRO clearly cannot cover every possible conceptualization of landforms, there are still many consistent patterns in how people from diverse backgrounds conceptualize landforms. This makes LFRO a worthy, albeit ambitious, ontology engineering initiative to unify, through linguistic and formal ontological approaches, those fundamental categories and relationships that typically and generally (but not necessarily) seem to underlie most people's common sense (naïve geographic) conceptualization and reasoning about landforms.[8] The expectation is that, others can use this as a foundation to represent and compare more specialized linguistic, cultural and geo-scientific concepts.

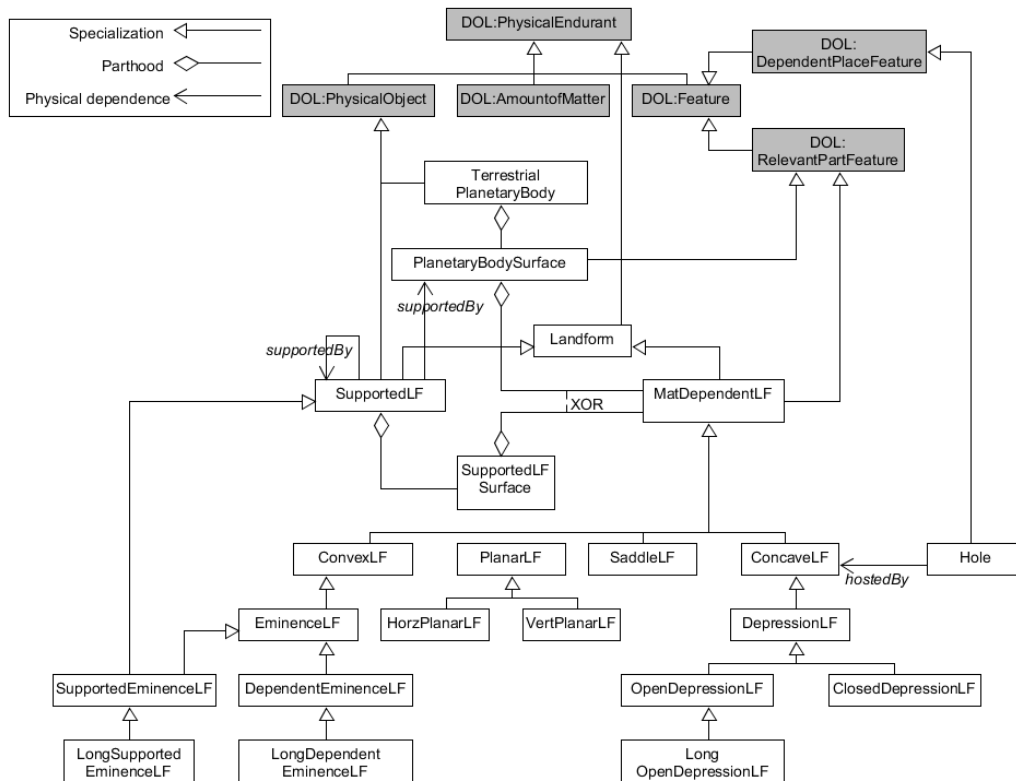
In the current phase, only the most important criterion of (three-dimensional) form is relied upon, because, landforms are apprehended as unitary entities primarily (but not exclusively) based on their characteristic form. Partitioning of the surface arbitrarily based on non-morphological criteria will yield regions with other unifying characteristics, but not landforms with a coherent, characteristic shape. An initial, simpler version of LFRO based on form considerations was introduced in a short paper.[12] In this substantially revised version, several new categories are introduced to explicate critical ontological notions of material-spatial dependence and support.[4]

Many scientific, administrative, and folk landform classification systems and vocabularies have obviously informed the conceptual development of LFRO. Here, for lack of space, only a few directly relevant ontology design efforts are acknowledged. The *surface network* ontology [13], which formalizes the well-established theory of surface networks, was the first step in identifying the critical shape elements of the terrain surface. When further aligned and integrated with the ontology of spatial regions [2] and contours [6], it will also complement LFRO as an automated terrain feature extraction and reasoning ontology (the primary inspiration for designing LFRO). The *surface water features* ontology pattern provided insights about depressions [14], while work on hydro domain formal ontology (HyFO) provided essential insights about holes and physical containment, which strongly influenced how concave landforms should be represented in LFRO.[2],[3],[5]

2 Design and rationale for the Landform Reference Ontology

2.1 The ontology of landforms

Figure 1 presents all the categories and relationships recognized in LFRO. Grounding of LFRO in the DOLCE upper level ontology [10] is now used to explicitly declare that all landforms are of type *Physical Endurant*, which are physical entities that wholly (i.e., with



■ **Figure 1** The categories and primary relationships of the Landform Reference Ontology (LFRO), with top level categories grounded in DOLCE categories (prefixed with DOL:).

all their proper parts) exist (typically as three-dimensional entities) in physical space at any time they exist. DOLCE specializes physical endurants into *physical objects*, *features* and *amount of matter*. Features depend on (are hosted by) other physical endurants and are of two types: relevant parts (materially constituted bumps, edges, surfaces) or dependent places (immaterial holes, shadows, etc.).

While LFRO is designed primarily for landforms on Earth, it can also be applied to any other terrestrial/telluric planetary body – i.e., whose surface region is materially constituted of rocky (silicate) or metallic material (like the bedrock and regolith composing the Earth’s surface). LFRO is also intentionally designed to remain neutral regarding what qualifies as the surface of a planetary body – whether it is only the bedrock or also includes all or some of the overlying regolith material (e.g., sand, soil, alluvium, glacial till).

Intuitively, a *landform* is a physical endurant that is physically dependent on the solid surface of a terrestrial planetary body (*TerrestrialPlanetaryBody*) in two ways: it is either part of the surface (*MatDependentLF*) or physically supported by/on the surface (*SupportedLF*). The *MatDependentLF* category includes landforms that are *materially* and *spatially* dependent on either the solid surface of a planetary body or the surface of a *SupportedLF* landform. As explained in [4], material-spatial interdependence (*mat-dep*) is a type of physical dependence that requires the physical extents of two entities to be necessarily and mutually contingent (e.g. an object and a material part thereof or its matter, or a hole and its host). Thus, *MatDependentLF* landforms are (DOLCE) relevant part features of the surface. The location and identity of *MatDependentLF* landforms are also intrinsically tied to their location on

the host surface. Most commonly known landform categories will be categorized under the *MatDependentLF* category.

In contrast to *MatDependentLF* landforms, *SupportedLF* landforms are not features of the host surface, but independent, physical object landforms supported on/by the surface. While the formalization of a support is still being worked out, intuitively, support is the relation between two material, physical endurants (objects or features) where one significantly contributes to maintaining the other one in one (or more) specific location(s). *SupportedLF* landforms can be supported directly on the planetary body surface, or another *SupportedLF* landform surface. *SupportedLF* landform surfaces can support both dependent and supported landforms. A special challenge in formalizing *SupportedLF* semantics arises because some surface entities (e.g., covered landfill and burial mounds) can be treated as landforms in some contexts, but other artificial structures (e.g., bridges, buildings), even if formed by naturally occurring rocks and soil, can never be.

2.2 Shape based categorization of landforms

In the first version of LFRO, *Landform* was the direct parent category for all shape-based subcategories. However, now, it can be said more specifically that all material landforms inherit their shape from certain characteristically shaped parts of the host surface. Based on generic landform shape-based categories proposed in [8], five, mutually exclusive shape-specific landform subcategories are specialized from the *MatDependentLF* category: *ConvexLF*, *ConcaveLF*, *HorzPlanarLF*, *VertPlanarLF* and *SaddleLF*. Convex and concave landforms comprise an overwhelming number of commonsense landform categories. While a convex landform protrudes out from the host surface, a concave landform is an indentation in the host surface, and necessarily hosts a hole feature. Planar surfaces are now further subcategorized as vertical or horizontal because such surfaces are experienced and lexicalized quite differently. The saddle subcategory was added to model passes, notches and gaps.

The semantics of concave landforms requires modeling of multiple possible perspectives of the negative spaces (holes) associated with such landforms. While any concave surface part necessarily encloses some hole, when people think of a concave-shaped *landform*, they may associate it variously with i) *only* the immaterial hole hosted by the concave part of the surface; ii) *only* the material concave part *ConcaveLF* of the surface, or iii) *both* the immaterial hole and the enclosing concave part of the surface. Every *ConcaveLF* landform must host some hole – regardless of whether the hole itself is viewed as a landform or not. Examples and implications of these three choices are discussed briefly in section 3.

An topographic eminence is a convex landform that stands completely above its surroundings. The corresponding *EminenceLF* category in LFRO is now split into two subcategories. *DependentEminenceLF* eminences (e.g., mountain, plateau, hill) inherit their characteristic convex shape from a host part of the planetary surface while *SupportedEminenceLF* eminences (e.g. landfills, mounds) are independent physical objects. Both these eminence categories are further specialized as *LongDependentEminenceLF* or *LongDependentSupportedLF* subcategories to explicitly cover elongated eminences such as ridge forms.

The subcategories for concave landforms and their various specializations remain unchanged from the previous version.[12] *ConcaveLF* landforms that are surrounded completely by higher land are specialized as *DepressionLF*, which is further specialized as *closed* and *open*. *ClosedDepressionLF* landforms have a rim marking the upper edge, the constant elevation of which is determined by the location of the closed depression's pour-point. *Closed-DepressionLF* landforms are special because they can store water for prolonged periods, thereby acting as containers for water bodies (e.g., puddle, lake, sea), provided enough water

is available. In contrast, *OpenDepressionLF* landforms are “open” because they lack a unique enclosing rim, and/or have holes and openings such that they cannot be store water, but only allow it to flow through. Open and closed depression landforms can also be specialized based on planimetric shape (of their spatial region) to distinguish elongated depressions. Only the *LongOpenDepressionLF* category is currently recognized because elongated open depressions (e.g., stream channel, valley, canyon, ravine, canal, and trench) are quite frequently recognized across the world. Such landforms are commonly perceived as a concave part of the surface with a primary, sloping longitudinal axis, sloping sides, and generally open at both ends of the longitudinal section to allow water to flow through.

3 Exploring semantics and mapping of linguistic landform categories using LFRO

While some linguistic categories are easier to associate with one LFRO category, many others can be interpreted in multiple ways. For instance, a mountain landform will be a surface-dependent eminence for most people, but the terms hill or ridge can be used for both surface-supported and dependent eminences. If the surface is defined as the bedrock only, cinder cones, drumlins, and sand dunes will all be categorized as *SupportedEminenceLF*. However, if the earth’s surface is not bedrock, but the exposed land surface (ground) that is directly accessible to us, the above-mentioned landforms should be modeled as (*DependentEminenceLF*) landforms. Similarly, people often assume craters to be like lake basins, which are closed depressions, but if any part of the rim is eroded to base level, the crater transforms into an open depression, that cannot contain water bodies. Considering a language other than English, the Yindjibardi term *marnda* can be applied to a variety of eminences including mountains, hills, ridges, and ranges.[9] Marnda is, therefore, almost (but not perfectly) synonymous with *EminenceLF*. Another Yindjibardi language term *bantha* refers to artificial or piled up eminences [9], and, will, therefore, be a subcategory of *SupportedEminenceLF*.

LFRO also helps illustrate practical reasoning implications of different conceptualizations of landforms. For example, if valley (or any other term) refers to just the immaterial *Hole* in the surface, then it can contain a water body, but it must be the concave part of the adjacent host surface that provides the material support for the water body and all other things that are “in” the valley. Alternatively, if the valley is just the material *ConcaveLF* part of the surface, it can only support, but not contain water bodies (which can only be contained in the hosted hole). Also, unlike the immaterial valley, a material concave valley can also share a part with the bordering mountain or hill. Finally, if a valley is conceived as a landform that has both *ConcaveLF* and *Hole* as necessary constituent parts, then people holding such a view would consider a valley to have all the above-mentioned properties. Note that it is not even necessary that people use the same interpretation for all concave landforms – for valley, they might choose the compound landform interpretation, while sink holes may be treated as holes, ignoring the materiality of their bottoms and sides.

LFRO can also be used to construct decision trees to choose appropriate mapping algorithms and construct semantic queries. For example, a semantic search for “*landforms that can store water*” would return all closed depression landforms, while searching for “*landforms where streams can flow*” would return all open longitudinal depression landforms. Analysis of linguistic terms and their alignment with LFRO also suggests that automated systems might be better off starting with methods to make generalized categorical distinctions. So, differences between mountain/hill/plateau/butte or valley/canyon/gorge or gully/gulch/ri

are probably quite difficult to tease out. It might be better to first define methods to delineate eminences, elongated eminences, closed depressions, and open longitudinal depressions.

While LFRO is still, primarily, a taxonomy, a comprehensive axiomatic formalization will be undertaken only after some existing limitations are resolved by adding new LFRO form categories. Then LFRO will be integrated with the hydro-domain ontology HyFO and the (suitably enhanced) surface network ontology to support semantic reasoning and guide automated mapping methods. For example, a request for mapping a valley floor can be recast as a query to find the area within a certain distance and/or height of a *surface network courseline*.^[13] A search for lake boundaries can be automatically inferred to require delineation of a closed depression landform, which in turn can be linked to finding pits and their basins from surface network theory.^[13] Finally, LFRO needs to be expanded to support reasoning with other non-morphological commonsense criteria such as size, material, color, geomorphological origin, and culturally significant factors.

References

- 1 Niclas Burenhult and Stephen C. Levinson. Language and landscape: a cross-linguistic perspective. *Language Sciences*, 30(2):135–150, 2008.
- 2 Torsten Hahmann and Boyan Brodaric. The void in hydro ontology. In *Proceedings of FOIS 2012*, page 45–58. IOS Press, 2012.
- 3 Torsten Hahmann and Boyan Brodaric. Kinds of full physical containment. In *Proceedings of COSIT 2013, LNCS 8116*, page 397–417. Springer-Verlag New York, Inc., 2013.
- 4 Torsten Hahmann, Boyan Brodaric, and Michael Grüninger. Interdependence among material objects and voids. In *Proceedings of FOIS 2014*, page 37–50. IOS Press, 2014.
- 5 Torsten Hahmann and Shirly Stephen. Using a hydro-reference ontology to provide improved computer-interpretable semantics for the groundwater markup language (gwm12). *International Journal of Geographical Information Science*, 32(6):1138–1171, 2018.
- 6 Torsten Hahmann and E. Lynn Usery. What is in a contour map? In *Proceedings of COSIT 2015, LNCS 9368*, page 375–399. Springer-Verlag New York, Inc., 2015.
- 7 David M. Mark and Barry Smith. Do mountains exist? Towards an ontology of landforms. *Environment and Planning B: Planning and Design*, 30(3):411–427, 2003.
- 8 David M. Mark and Barry Smith. A science of topography: Bridging the qualitative quantitative divide. In Michael P. Bishop and John F. Shroder, editors, *Geographic Information Science and Mountain Geomorphology*, page 75–100, Chichester, England, 2004. Springer-Praxis.
- 9 David M. Mark and Andrew G. Turk. Landscape categories in yindjibarndi: Ontology, environment, and language. In *Proceedings of COSIT 2003, LNCS 2825*, pages 28–45. Springer International Publishing, 2003.
- 10 Claudio Masolo, Stefano Borgo, Aldo Gangemi, Nicola Guarino, and Alessandro Oltramari. Wonderweb deliverable d18 - ontology library (final report). Technical report, IST Project 2001-33052 WonderWeb: Ontology Infrastructure for the Semantic Web, 2003.
- 11 Konrad Rybka. Between objects and places: The expression of landforms in Lokono (Arawakan). *International Journal of American Linguistics*, 81(4):539–572, 2015.
- 12 Gaurav Sinha, Samantha T. Arundel, Kathleen Stewart, David M. Mark, Torsten Hahmann, Boleslo E. Romero, Alexandre Sorokine, E. Lynn Usery, and Grant MacKenzie. A reference landform ontology for automated delineation of depression landforms from dems. In *Proceedings of Workshops and Posters, COSIT 2017*, page 111–116. Springer International Publishing, 2017.
- 13 Gaurav Sinha, Dave Kolas, David M. Mark, Boleslo E. Romero, E. Lynn Usery, and Gary Berg-Cross. Surface network ontology design patterns for linked topographic data. *Semantic*

Web, 2014. Manuscript to be re-submitted to Semantic Web. Draft version available online @ <http://www.semantic-web-journal.net/system/files/swj675.pdf>.


- 14 Gaurav Sinha, David M. Mark, Dave Kolas, Dalia Varanka, Boleslo E. Romero, Chen-Chieh Feng, E. Lynn Usery, Joshua Liebermann, and Alexandre Sorokine. An ontology design pattern for surface water features. In *Proceedings of GIScience 2014, LNCS 8728*, page 187–203. Springer International Publishing, 2014.

Abstract Data Types for Spatio-Temporal Remote Sensing Analysis

Martin Sudmanns


University of Salzburg, Department of Geoinformatics - Z_GIS, Schillerstraße 30, Salzburg, Austria

martin.sudmanns@sbg.ac.at

 <https://orcid.org/0000-0002-0473-1260>


Stefan Lang

University of Salzburg, Department of Geoinformatics - Z_GIS, Schillerstraße 30, Salzburg, Austria

 <https://orcid.org/0000-0003-0619-0098>

Dirk Tiede

University of Salzburg, Department of Geoinformatics - Z_GIS, Schillerstraße 30, Salzburg, Austria

 <https://orcid.org/0000-0002-5473-3344>

Christian Werner

University of Salzburg, Department of Geoinformatics - Z_GIS, Schillerstraße 30, Salzburg, Austria

Hannah Augustin

University of Salzburg, Department of Geoinformatics - Z_GIS, Schillerstraße 30, Salzburg, Austria

Andrea Baraldi

Italian Space Agency (ASI), Rome, Italy.

Abstract

Abstract data types are a helpful framework to formalise analyses and make them more transparent, reproducible and comprehensible. We are revisiting an approach based on the space, time and theme dimensions of remotely sensed data, and extending it with a more differentiated understanding of space-time representations. In contrast to existing approaches and implementations that consider only fixed spatial units (e.g. pixels), our approach allows investigations of the spatial units' spatio-temporal characteristics, such as the size and shape of their geometry, and their relationships. Five different abstract data types are identified to describe geographical phenomenon, either directly or in combination: coverage, time series, trajectory, composition and evolution.

2012 ACM Subject Classification Information systems → Search interfaces

Keywords and phrases Big Earth Data, Semantic Analysis, Data Cube

Digital Object Identifier 10.4230/LIPICs.GIScience.2018.60

Category Short Paper

Funding The study was supported by the Austrian Science Fund (FWF) through the Doctoral College GIScience [DK W1237-N23] and by the Austrian Federal Ministry of Transport, Innovation and Technology (BMVIT) under the program “ICT of the Future“ within the project SemEO [contract no: 855467].



© Martin Sudmanns, Stefan Lang, Dirk Tiede, Christian Werner, Hannah Augustin, and Andrea Baraldi;

licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 60; pp. 60:1–60:7

Leibniz International Proceedings in Informatics



Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction & Motivation

In the context of big Earth data, users do not seem to struggle mainly with technical problems, such as the provision of hardware (e.g. disk space or computing power), but are challenged by conceptual problems. These include decisions on how to observe phenomena on Earth (e.g. see [6]), store and analyse observations (e.g. see [3]), or replicate studies (e.g., see [19] or [14]). The value of big data, other than their volume, variety, and velocity, is challenging to leverage not based on inherent data characteristics, rather by how the data will be used [13]. For example, many data storage systems perform well when inputting data (i.e. saving raw EO images), but perform poorly when outputting data (i.e. finding relevant data and producing information from them) [13, 22]. Not knowing how data are structured and how they will be used on a generic level does not only challenge the general use of big Earth data, but also the replication of studies and reuse of workflows, because tools are not clearly distinguished from methods and data are not separated by semantic type [19].

Regular, free provision of Landsat and Sentinel data makes analyses of the temporal dimension increasingly important. Therefore, 3D Earth observation (EO) geospatial data cubes [18, 17] are becoming an increasingly popular tool. They do not treat images as temporally isolated, but index and reference them in a data structure where all axes (e.g. spatial and temporal dimensions) can be integrated and accessed equally [18]. It is necessary to know what types of queries are expected in order to decide on an optimal tiling scheme to optimize a geospatial data cube [8].

Increased data availability allows for analysis of high-resolution images, like Sentinel-2, on a continental or global scale, therefore opening new application domains such as serving the information needs of intergovernmental agreements, e.g. the United Nations sustainable development goals (SDGs). In this context, EO data and analysis methods spread into 'new' domains and confront new user communities with their complexity and particularities without providing a guiding and logical understanding of the representation of the geographical reality.

With all the technical preconditions available, analyses still aim to produce information relevant to questions posed by humans. The translation from questions to queries and results to answers is difficult, necessitates more than increasing data volumes and computing power, and goes beyond pure technical achievements. Recent developments are often technology-driven and are not necessarily tied to user requirements, where user groups are also non-experts from various application domains. For example, terminologies like 'big Earth data', 'data cube' or 'analysis ready data' are used before a proper definition or a common understanding is achieved. Inexperienced users struggle to become familiar with tools for reasons which might include a lack of common core terminology [15] and gaps between the user domain and the technical EO image domain [22, 21, 4]. This is especially complicated because a consistent conceptual model of space-time (e.g. consisting of continuants and occurrents (events) [9] and their relationships), as a representation of a mental model of the physical world (i.e. world model or world ontology), is still missing.

While the definition and a formalisation of a world model goes beyond the scope of this short paper, a certain level of understanding of at least continuants is necessary as a first step. A continuant can be seen as an entity in the physical world, parameterised by a unique continuant-identifier and an inner state, consisting of three types of attributes in the modelled 4D physical world: (a) positional, 3D geospatial attributes in geospatial units (e.g. lat-long coordinates and height in meters); (b) time attribute in a physical unit of time; and (c) "theme" [20]. We define theme as the combination of: (I) a theme type (i.e. geo-objects, geo-fields, and field-objects according to [10]); (II) a theme name (e.g.

any symbolic geo-object has a theme name belonging to a finite and discrete hierarchical, structured taxonomy of concepts or classes of real-world objects); and (III) appearance properties in the 4D physical world, expressed as either quantitative/numeric variables or qualitative/categorical sub-symbolic theme attributes in physical units [16]. These are: (1) photometric properties, expressed as either numeric colour values in spectral reflectance units (e.g. mean reflectance) or categorical colour names (e.g. red) belonging to a community-agreed discrete and finite vocabulary of colour names, related to a partition of a numeric colour space into quantization bins [11]); (2) shape (i.e. geometric) variables [2] such as compactness, rectangularity, elongatedness, straightness of boundaries, simple connectivity and orientation; and (3) size variables, like length and width in metres. Occurents, as events, are able to change the inner state of a continuant, its relationship to other continuants, or the emergence of new continuants. To stick with the examples given above, we may conceive occurents as rotating crop types on an agricultural field, or the vanishing of a lake. The latter changes its size and thereby also its relationship to other continuants (patches of vegetation or open soil), which emerge simultaneously as new continuants.

For defining abstract data types for the application on Earth observation data, our conjecture is:

1. *The variety of phenomena in the focus of Earth observation can be represented and categorised by a limited set of abstract data types.*
2. *Having a set of defined abstract data types and knowing their behaviour can make remote sensing analyses more comprehensive and reproducible.*

2 State-of-the-art and research gap

A set of generic data types for spatio-temporal data was proposed by [7] based on three dimensions (i.e. spatial, temporal and thematic dimensions) inherent to any geospatial data [20]. Observations can be analysed by keeping one attribute fixed, controlling another and measuring the third. For example, in an EO image, fixing time, but controlling space and measuring the theme yields a land cover map. Similarly, fixing space (e.g. the location of a temperature sensor), controlling time and measuring the theme represents a temperature curve throughout a year. In total, [7] identified three out of nine possible data types as relevant:

- Coverage: fixing time, controlling space, measuring theme
- Time series: fixing space, controlling time, measuring theme
- Trajectory: fixing theme, controlling time, measuring space

Another method for separating geospatial data types from their physical organisation is comprehensively described by [1], where "spatial lenses" provide software-based views as a way to interpret datasets. The interpretations, based on a specific view of the world, include a network, objects, fields and events, as well as refer back to the core concepts of spatial data [15].

In the remote sensing domain, geographic object-oriented image analysis (GEOBIA) uses image segments (i.e. objects) instead of pixels as target analysis units [5]. Therefore, GEOBIA applies object-oriented data models to geographic image data. Since the segments have inherent spatial characteristics (e.g. size, shape, topological arrangement) and can be temporally associated with each other, GEOBIA allows spatial and temporal analyses. Typically, the objects' semantics are modelled using ontologies or a rule-based approach, such as implemented in the eCognition software. However, the ontologies or rule-sets are usually tied to a virtual 2D map legend domain and not to the 4D physical world domain [4].

Separating the virtual image domain from the physical world domain in EO image analysis was introduced in [16] and was then later taken up and applied as a GEOBIA-based approach by [12] and [22, 21, 4].

Although some previous work is available, a set of universally applicable, comprehensive, abstract data types for EO data have not yet been developed. Such a set could serve as a framework for mapping spatial, temporal and thematic attributes of observations in EO data cubes. Existing approaches and implementations lack either generality (e.g. specific GEOBIA implementations), or are limited to fixed analysis units (e.g. pixels). We suggest abstract data types to be used as a logical, intermediate layer between EO data cubes and the 4D physical world domain, thus adopting a clear distinction from the physical organisation of data [1] as well as the 2D virtual image domain [22, 21, 4]. Our proposed abstract data types adapt the ideas of [7] and extend them with the more differentiated understanding of space-time phenomena and their spatial, temporal or semantic relations in GEOBIA required for spatial image analysis [2]. Space in an EO image context has multiple meanings since it: (1) refers to the absolute or relative location of an object (e.g. represented by a coordinate tuple) and its spatial relation to other objects; and (2) also refers to inherent spatial characteristics of an object (e.g. size and shape). In a more complex situation, e.g. observing the expansion of a city, the object itself is the result of a spatial arrangement of other objects, including houses and streets.

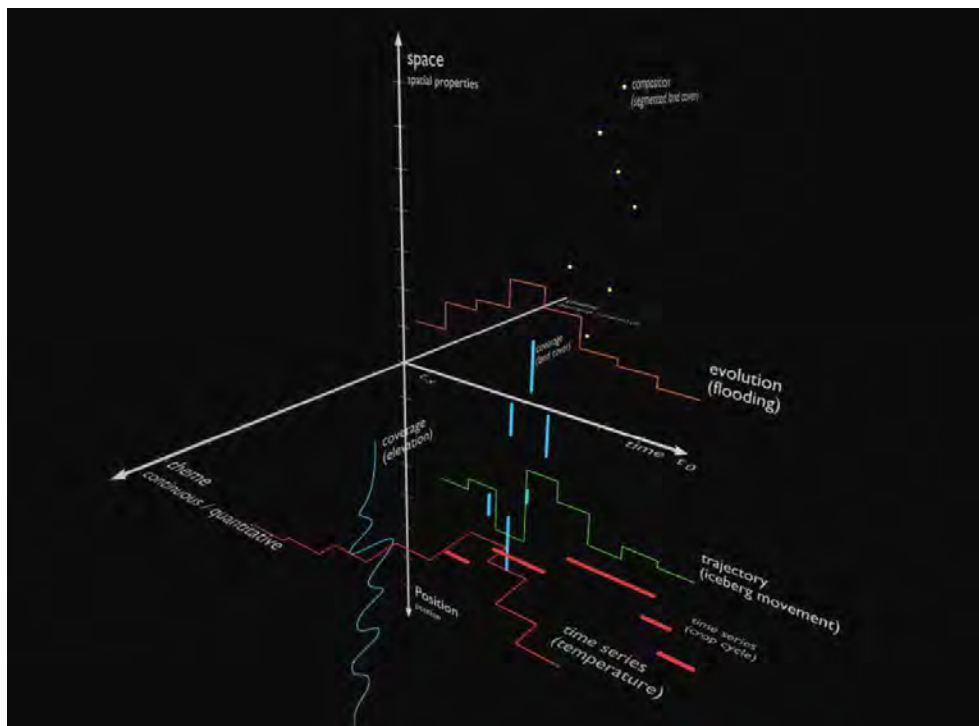
3 Proposed abstract data types

We differentiate between *position* (or *location*) and *space*, which are inherent spatial properties of objects. Further, a position of an object might not only be the absolute position, but also its relative location within a topological arrangement. We also differentiate between continuous (i.e. quantitative) and discrete (i.e. categorical) variables. The temporal dimension has its upper limit in t_0 and goes back until t_{-x} as this approach is intended for querying an archive and not for projecting processes in the future. The following abstract data types can be selected, and are illustrated in Figure 1:

- **Coverage:** constructed by fixing time, controlling position, measuring theme (continuous or discrete)
- **Composition:** constructed by fixing time, controlling theme, measuring space
- **Time series:** constructed by fixing the position, controlling time, measuring theme (continuous or discrete)
- **Trajectory:** constructed by fixing theme, controlling time, measuring position
- **Evolution:** constructed by fixing theme, controlling time, measuring space.

4 Conclusion & Outlook

Challenges of big Earth data go beyond technical issues. We suggest a limited, yet defined and tangible set of abstract data types, which are specifically selected for use as a framework for query primitives within EO data cubes. While existing solutions rely on fixed spatial units, such as pixels, in GEOBIA the space properties do not only refer to the position, but also to the spatial arrangement of objects and to properties such as extent, shape and size of the object under consideration. Based on the state-of-the-art review, we found the necessity to extend the original set of abstract data types with two new ones to account for the differentiated view on space within the GEOBIA domain. While defining this framework is an ongoing process and this contribution is a first step towards it, in this short paper



■ **Figure 1** Example phenomena relevant to Earth observation visualised in a 3D space. Here, the axes provide an ordering principle for EO spatio-temporal phenomena. Note that space can be conceived as position (e.g. 0-dimensional, coordinate tuple or tripel) and the spatial relationship to other entities, or as geometric features (e.g. set of coordinate tuples, size, compactness). Although the attributes are represented on single, individual axis, the semantics of the axes differ between theme or time (monodimensional) and space or position (multidimensional). An interactive visualisation is available as online visualisation (<http://cf000008.geo.sbg.ac.at/adt/>).

we aim to highlight the necessity of having it for formalising queries. Future work will align this framework with the definition of a world model as a conceptual description of geospatial phenomena, e.g. using a rigorous formalisation of continuants, occurrences and their relationships. Further, this also includes revisiting the original and suggested terms and a discussion of whether they are appropriate for this purpose. Being in a preliminary stage, the framework and the abstract data types are presented here in a rather informal manner. Therefore, the focus will lie on the formalisation of the data types and their methods as well as an example implementation in an EO data cube.

Abstract data types allow for semantic annotations and workflow exchanges by separating methods from tools and the image domain from the physical world domain. They can be considered as a logical, intermediate layer between the conceptual world model and the data storage engine, e.g. geospatial data cubes. Therefore, they can be used to answer questions such as “what data are used?” or “what are they useful for?” and are linked to big Earth data relevant decisions. These include but are not limited to how certain phenomena can be observed, how a system can be designed to provide analysis results with reasonable response times and how the result can be interpreted and deemed trustworthy. Further, they help non-EO experts to express their questions in formalised terms. It is increasingly relevant to analyse EO data together with non-EO data, where abstract data types might also play a significant role.

References

- 1 Christopher Allen, Thomas Hervey, Sara Lafia, Daniel W. Philips, Behzad Vahedi, and Werner Kuhn. Exploring the notion of spatial lenses. In *Geographic Information Science 2016*, pages 260–274. Springer, Montréal, Canada, 2016. doi:10.1007/978-3-319-45738-3.
- 2 Andrea Baraldi, Dirk Tiede, Martin Sudmanns, and Stefan Lang. Systematic esa eo level 2 product generation as pre-condition to semantic content-based image retrieval and information/knowledge discovery in eo image databases. In Publications Office of the European Union, editor, *Proceedings of the 2017 conference on Big Data from Space*. Publications Office of the European Union, 2017.
- 3 Peter Baumann, Paolo Mazzetti, Joachim Ungar, Roberto Barbera, Damiano Barboni, Alan Beccati, Lorenzo Bigagli, Enrico Boldrini, Riccardo Bruno, Antonio Calanducci, Piero Campalani, Oliver Clements, Alex Dumitru, Mike Grant, Pasquale Herzig, George Kakalettris, John Laxton, Panagiota Koltsida, Kinga Lipskoch, Alireza Rezaei Mahdiraji, Simone Mantovani, Vlad Merticariu, Antonio Messina, Dimitar Misev, Stefano Natali, Stefano Nativi, Jelmer Oosthoek, Marco Pappalardo, James Passmore, Angelo Pio Rossi, Francesco Rundo, Marcus Sen, Vittorio Sorbera, Don Sullivan, Mario Torrisi, Leonardo Trovato, Maria Grazia Veratelli, and Sebastian Wagner. Big data analytics for earth sciences: the earthserver approach. *International Journal of Digital Earth*, 9(1):3–29, 2016. doi:10.1080/17538947.2014.1003106.
- 4 Mariana Belgiu, Martin Sudmanns, Dirk Tiede, Andrea Baraldi, and Stefan Lang. Spatiotemporal enabled content-based image retrieval. In *Ninth International Conference on GIScience, Short Paper Proceedings*, volume 9. University of California, 2016.
- 5 Thomas Blaschke, Geoffrey J Hay, Maggi Kelly, Stefan Lang, Peter Hofmann, Elisabeth Addink, Raul Queiroz Feitosa, Freek van der Meer, Harald van der Werff, Frieke van Coillie, et al. Geographic object-based image analysis—towards a new paradigm. *ISPRS Journal of Photogrammetry and Remote Sensing*, 87:180–191, 2014.
- 6 M Drusch, U Del Bello, S Carlier, O Colin, V Fernandez, F Gascon, B Hoersch, C Isola, P Laberinti, P Martimort, et al. Sentinel-2: Esa’s optical high-resolution mission for gmes operational services. *Remote Sensing of Environment*, 120:25–36, 2012.
- 7 Karine Reis Ferreira, Gilberto Camara, and Antônio Miguel Vieira Monteiro. An algebra for spatiotemporal data: From observations to events. *Transactions in GIS*, 18(2):253–269, 2014.
- 8 Paula Furtado and Peter Baumann. Storage of multidimensional arrays based on arbitrary tiling. In *Data Engineering, 1999. Proceedings., 15th International Conference on*, pages 480–489. IEEE, 1999.
- 9 Antony Galton. Experience and history: Processes and their relation to events. *Journal of Logic and Computation*, 18(3):323–340, 2008.
- 10 Michael F Goodchild, May Yuan, and Thomas J Cova. Towards a general theory of geographic representation in gis. *International journal of geographical information science*, 21(3):239–260, 2007.
- 11 Lewis D Griffin. Optimality of the basic colour categories for classification. *Journal of the Royal Society Interface*, 3(6):71–85, 2006.
- 12 Stefan Growe. Knowledge based interpretation of multisensor and multitemporal remote sensing images. *Int. Arch. Photogramm. Remote Sens*, 32(pt 7):4–3, 1999.
- 13 Adam Jacobs. The pathologies of big data. *Communications of the ACM*, 52(8):36, 2009. doi:10.1145/1536616.1536632.
- 14 Christian Knoth and Daniel Nüst. Reproducibility and practical adoption of geobia with open-source software in docker containers. *Remote Sensing*, 9(3):290, 2017.


- 15 Werner Kuhn. Core concepts of spatial information for transdisciplinary research. *International Journal of Geographical Information Science*, 26(12):2267–2276, 2012. doi:10.1080/13658816.2012.722637.
- 16 T. Matsuyama and V.S.S. Hwang. *SIGMA: A Knowledge-Based Aerial Image Understanding System*. Advances in Computer Vision and Machine Intelligence. Springer US, 1990.
- 17 Stefano Nativi, Paolo Mazzetti, and Max Craglia. A view-based model of data-cube to support big earth data systems interoperability. *Big Earth Data*, pages 1–25, 2017.
- 18 Peter Strobl, Peter Baumann, Adam Lewis, Zoltan Szantoi, Brian Killough, Matthew Purss, Max Craglia, Stefano Nativi, Alex Held, Trevor Dhu. The six faces of the data cube. In Publications Office of the European Union, editor, *Proceedings of the 2017 conference on Big Data from Space*, pages 32–35. Publications Office of the European Union, 2017.
- 19 Simon Scheider, Frank O. Ostermann, and Benjamin Adams. Why good data analysts need to be critical synthesists. determining the role of semantics in data analysis. *Future Generation Computer Systems*, 72:11–22, 2017. doi:10.1016/j.future.2017.02.046.
- 20 David Sinton. The inherent structure of information as a constraint to analysis: Mapped thematic data as a case study. *Harvard papers on geographic information systems*, 6:1–17, 1978.
- 21 Martin Sudmanns, Dirk Tiede, Stefan Lang, and Andrea Baraldi. Semantic and syntactic interoperability in online processing of big earth observation data. *International Journal of Digital Earth*, 11(1):95–112, 2018. doi:10.1080/17538947.2017.1332112.
- 22 Dirk Tiede, Andrea Baraldi, Martin Sudmanns, Mariana Belgiu, and Stefan Lang. Architecture and prototypical implementation of a semantic querying system for big earth observation image bases. *European journal of remote sensing*, 50(1):452–463, 2017. doi:10.1080/22797254.2017.1357432.

Towards Vandalism Detection in OpenStreetMap Through a Data Driven Approach

Quy Thy Truong

Univ. Paris-Est, LASTIG COGIT, IGN, ENSG, F-94160 Saint-Mande, France


quy-thy.truong@ign.fr

 <https://orcid.org/0000-0002-3413-482X>

Guillaume Touya

Univ. Paris-Est, LASTIG COGIT, IGN, ENSG, F-94160 Saint-Mande, France


gguillaume.touya@ign.fr

 <https://orcid.org/0000-0001-6113-6903>

Cyril de Runz

Modeco, CReSTIC, University of Reims Champagne-Ardenne, CS 30012, Reims cedex 2, France

cyril.de-runz@univ-reims.fr

 <https://orcid.org/0000-0002-5951-6859>

Abstract

Vandalism is a phenomenon that has affected by now the digital domain, in particular in the context of Volunteered Geographic Information projects. This paper aims at proposing a methodology to detect vandalism in the OpenStreetMap project. First, an analysis of related works sheds light on the lack of consensus when it comes to defining vandalism in VGI from both conceptual and practical points of view. Second, we present experiments on the use of clustering-based outlier detection methods to identify vandalism in OSM. The outcome of this study focuses on choosing the right variables when it comes to detecting vandalism in OSM.

2012 ACM Subject Classification Human-centered computing → Collaborative content creation, Computing methodologies → Anomaly detection

Keywords and phrases Vandalism, Volunteered Geographic Information, Outlier detection

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.61

Category Short Paper

1 Introduction

Skepticism toward the use of Volunteered Geographic Information (VGI) stems from the lack of data qualification in VGI datasets despite the likelihood of poor quality contribution occurrences. In the case of the OpenStreetMap (OSM) project, allowing anyone to map also adds the risk of welcoming ill-intentioned contributors who impoverish the quality of the data through acts of vandalism. For instance, some of the Pokemon Go players who signed up as OSM contributors wrongly mapped geographic elements in order to boost the development of Pokemon nests¹. But how to distinguish actual vandalism from unintended mistakes? And how to automatically detect real vandalism in OSM? This paper's contribution is twofold: first we highlight the various definitions of vandalism that were adopted to automatically detect vandalism. Then we investigate the ability of an unsupervised method to detect vandalism in OSM by using a clustering-based outlier detection.

¹ <http://resultmaps.neis-one.org/osm-discussion-comments?uid=6310073&commented>



© Quy Thy Truong, Guillaume Touya, and Cyril de Runz;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 61; pp. 61:1–61:7

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

2 Understanding vandalism: related work

Historically, vandalism comes from Germanic barbarians, called Vandals, who were reputed for sacking artworks and monuments during their invasions in Western Europe [7]. Over times, its meaning broadened and nowadays vandalism refers generally to material defacement made by human beings. However, a degradation does not necessarily fall under vandalism because depending on the context, an act will not bear the same label. For instance, animal slaughter can be labeled as vandalism unless the killer has a license to hunt [7]. Actually, labeling an action as vandalism requires to assess the damage caused, the author's motives and the context of the incident [2]: these notions are already difficult to evaluate juridically, as each case has its own elements of context and oftentimes the author's motives are not directly accessible. Actually, vandalism definition is quite clear but as it relies on elements that are hard to assess for human beings, detecting it automatically remains a challenge.

Automatic detection of vandalism has been widely studied in Wikipedia [1] and Wikidata [4], but as these papers dealt with vandalism detection using supervised machine learning, they did not focus on giving a clear definition of vandalism. Actually, the existence of a corpus of labeled data on Wikidata/Wikipedia enabled them to evacuate the question. [6] developed a rule-based vandalism detection system for OSM data. The rules mainly take into account user reputation and object history. Therefore OSM newbies' created objects are at a disadvantage as they are more prone to be detected as vandalism. Unlike Wikidata, no corpus of OSM vandalism data is available. This is why our experiment attempts to detect OSM vandalism with the use of an unsupervised method and considering other vandalism metrics that were not tackled in [6].

Through the analysis of intentional vandalism incidents on Wikimapia and OSM, [2] proposed a typology for carto-vandalism composed of six categories: play carto-vandalism, ideological carto-vandalism, fantasy carto-vandalism, artistic carto-vandalism, industrial carto-vandalism and spam. This typology is drawn from experimental observations so it is quite realistic, however it can be difficult in some cases to label vandalism straight away in one of these categories. For instance, artistic carto-vandalism can be seen as a sub-category of fantasy carto-vandalism: mapping polygon art is necessarily a fictional data. In fact, the proposed typology implies knowing the contributor's intentions, which is a research problem in itself. On this intentionality issue, [6] solves the problem by stating that “ *in the case of OSM, vandalism can occur intentional and unintentional, contradicting the traditional definition of the term ‘vandalism’* ”. However, the OSM Wiki page about vandalism² does mention the difference between vandalism and bad editing which lays in the contributor's purpose, although both of them require data repairs. Actually, OSM vandalism and bad editing may both result in the same defacement of the dataset. This is why OSM Wiki page on vandalism does not provide a definition of what vandalism is but how it manifests in OSM dataset together with bad editing. Thus another challenge for vandalism detection in OSM is to steer clear of mistaking bad edits with true vandalism (i.e. minimizing false positives).

Like in [6], we analyze some cases of OSM user blocks in the light of the context, the user's motive and the caused damage in order to better understand what belongs to true vandalism and what does not. The case depicted in Figure 1 is true vandalism on versions 6 and 7, as the contributor³ completely defaced the nature of a Russian island by changing its name tags and turning it into a park. These edits are obviously made on purpose. We also

² <https://wiki.openstreetmap.org/wiki/Vandalism>

³ https://www.openstreetmap.org/user_blocks/1598

Version	5	6	7	8
Time	September 10, 2017 10:10 AM	October 21, 2017 10:19 PM	October 21, 2017 10:19 PM	October 21, 2017 10:20 PM
Changeset	51900382	53137491	53137504	53137513
User	nyuriks	higuy	higuy	higuy
Tags				
leisure		park	park	
name	остров Геральд	Jerryland	Jerryland	Jerryland
name:en	Herald Island	Herald Island	Jerryland	Herald Island
name:ru	остров Геральд	остров Геральд	остров Геральд	остров Геральд
natural	coastline	coastline	coastline	coastline
place	island			island
source	bing	bing	bing	bing
wikidata	Q1586913	Q1586913	Q1586913	Q1586913
wikipedia	ru:Геральд (остров)	ru:Геральд (остров)		

■ **Figure 1** Tag history of an OSM fantasy vandalism case (source: OSM Deep History application). Each column gives to the state of an OSM object’s version concerning its metadata and its tag values. Key tags are on the left. The changes are coded by colors: green stands for tag addition, yellow for tag-value edit and red for tag delete.

Time	April 24, 2017 6:05 PM	April 25, 2017 9:54 AM	May 9, 2017 4:22 PM	May 11, 2017 7:52 AM	May 16, 2017 12:17 AM	May 31, 2017 2:01 PM
Changeset	48095750	48114086	48533533	48580533	48714592	49133727
User	Diana777	SomeoneElse_Revert	Diana777	noteka	richiv	Diana777
Tags						
alt_name	Anņinmuižas parks	Anņinmuižas parks	Anņinmuižas parks	Anņinmuižas parks	Anņinmuižas parks	Anņinmuižas parks
landuse	forest	forest	forest		forest	
leisure	park		park	park		park
name	Anņinmuižas mežs	Anņinmuižas mežs	Anņinmuižas mežs	Anņinmuižas mežs	Anņinmuižas mežs	Anņinmuižas mežs

■ **Figure 2** Tag history of an ambiguous area in Latvia (source: OSM Deep History application).

note that the changes made on version 8 are not vandalism as the same user brings back the values of some previously altered tags. The edit war on an area in Latvia depicted on Figure 2 shows a disagreement about the real nature of this place. The banned user⁴ is the one who added the ‘leisure=park’ tag to the area. However, further research shows that local people do consider this place as a park⁵. Consequently, this case is not truly vandalism but rather highlights the ambiguity of the geographic object. Lastly, adding unconventional tags⁶ can be seen as an abnormal contribution but in this case it is not vandalism. Some of the tag values are understandable for humans and actually add valuable information to the objects (Figure 3). These examples show that vandalism – according to a data-oriented traditional definition – is less regular than contributions being non-compliant to OSM policy. Due to the scarcity of vandalism in OSM and the difficulty to enumerate all of the possible cases, this study tackles the vandalism issue following an outlier detection approach.

3 Methodology

Assuming that vandalized data form outliers in a dataset, our experiment aims at finding out whether using a clustering-based outlier detection enables to identify vandalized data in an OSM dataset. As vandalism does not often occur on OSM and we do not know where it

⁴ https://www.openstreetmap.org/user_blocks/1328

⁵ <http://www.spottedbylocals.com/riga/anninmuizas-mezs/>

⁶ https://www.openstreetmap.org/user_blocks/1455

Tags	
ls_in	3 BOULEVARD GALLIENI
Ref:FR:SIREN	56209154601009
addr:housenumber	3
addr:street	Boulevard Gallieni
building	office
building:levels	7
description1	IMMOBILIER
description2	Promotion immobilière
name	Bouygues Immobilier
note:qadastre	
phone	01 55 38 25 25
ref:FR:SIREN	
smoking	no
source	https://www.data.gouv.fr/fr/datasets/entreprises-de-grand-paris-seine-ouest-de-plus-de-50-salaries-2013/
website	http://www.bouygues-immobilier.com

■ **Figure 3** OSM oddly-tagged contribution (source: OSM Deep History application).

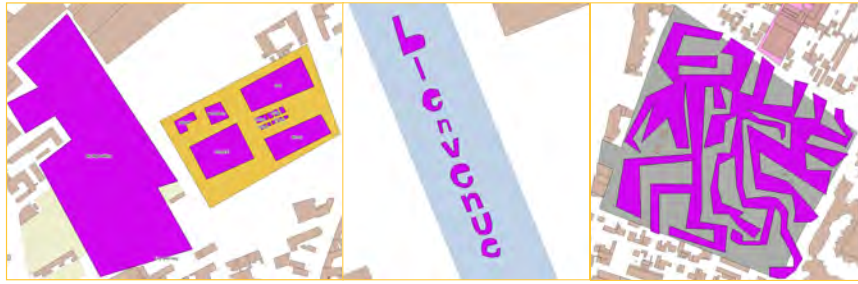
■ **Table 1** Overview of OSM building geometric variables that are used for the experiments. N.B. : MBR stands for Minimal Bounding Rectangle.

Variable name	Description
<i>perimeter_out_of_max</i>	$\frac{\text{perimeter}(\text{building})}{\max_{\text{dataset}}(\text{perimeter})}$
<i>area_out_of_max</i>	$\frac{\text{area}(\text{building})}{\max_{\text{dataset}}(\text{area})}$
<i>shortest_length_out_of_perimeter</i>	$\frac{\text{length}(\text{shortest_edge}_{\text{building}})}{\text{perimeter}(\text{building})}$
<i>median_length_out_of_perimeter</i>	$\frac{\text{length}(\text{median_edge}_{\text{building}})}{\text{perimeter}(\text{building})}$
<i>elongation</i>	$\frac{\text{width}(\text{MBR}_{\text{building}})}{\text{length}(\text{MBR}_{\text{building}})}$
<i>convexity</i>	$\frac{\text{area}(\text{building})}{\text{area}(\text{MBR}_{\text{building}})}$
<i>compactness</i>	Miller's index: $\frac{4 * \pi * \text{area}(\text{building})}{\text{perimeter}(\text{building})^2}$

happened, we cannot choose a study area where we would be assured to find vandalism cases. Therefore, we need to purposely add vandalized data so that the outliers to be detected are known in advance.

Then, every OSM element should be described by variables that will be used as inputs for the clustering algorithm. In the first place, the experiment will be limited to the detection of vandalism on buildings. This implies retrieving OSM ways and OSM relations that contain the ‘building’ key tag. To best describe OSM data, several types of descriptors may be contemplated: geometric variables [3], topological variables [3], historic variables [6] and user variables have been used in the literature to qualify OSM and crowdsourcing data in general [1, 6]. In this study, we employ fantasy and artistic vandalism to deface our dataset so at the moment only geometric variables were input into the clustering algorithm, as artistic vandalism is characterized with oddly shaped objects (Table 1). Eventually, the clustering algorithm will group similar objects according to their input attributes while setting aside buildings having particular values.

■ **Figure 4** Fantasy vandalism (left image) and artistic vandalism in Aubervilliers, France.



■ **Table 2** Outliers detected in OSM vandalized dataset using DENCLUE clustering algorithm ($\sigma = 0.05, m = 10250$)

	Number of outliers	Number one-size clusters
Artistic vandalism (total: 10)	1	9
Fantasy vandalism (total: 17)	0	15
Non-vandalism (total: 10315)	115	6080

4 Experiment and initial results

In this study, the dataset is composed of OSM buildings that are located in Aubervilliers, a suburban town of Paris. Vandalism committed in this dataset includes (Figure 4): 17 fictional buildings of different sizes which were mapped in a blank space (the yellow polygon in Figure 4 indicates that this space is currently an area under construction) and 10 artistically shaped buildings that were mapped in the middle of a river and over the town’s graveyard.

The outlier detection was run using the DENCLUE clustering algorithm (Java Smile library) because it is noise-invariant and remains efficient for high dimensional datasets [5]. It takes a smoothing parameter σ that describes the influence of a data point in the data space, and a parameter m that corresponds to the noise threshold. The algorithm starts by building a clustering model based on the input variables, then predicts the class of each element according to the clustering model. At this point, buildings whose descriptors are totally inconsistent with the clustering model are classified as outliers. The others are classified into clusters. However, some clusters contain only one element, meaning the values of these buildings descriptors fit into the clustering model but no building was similar enough regarding its attributes’ values to belong to the same cluster. Thus, in a certain way, these one-size clusters can be considered as outliers too but to a lesser degree. Table 2 summarizes the number of outliers and one-size clusters that were detected for each kind of data (vandalism or not).

The first ‘e’ letter-shaped building was the only outlier-labeled vandalism while the remainder of artistically vandalized buildings – including the other two ‘e’ letter-shaped ones – was classified into one-size clusters. We note that 25 cases of vandalism out of 27 – that represents 92% of known vandalism – could be retrieved either in the outlier class or a one-size cluster, which is quite outstanding. Nevertheless, 60% of normal buildings have been also classified into outliers or one-element clusters. By taking a look at the variables of the vandalized buildings, we notice that the geometric descriptors do not bring out the geometric peculiarities of the artistic vandalism that was committed into our dataset. Maybe considering a polygon density variable which accounts for a polygon’s number of vertices

would have brought out all of the committed artistic vandalism. OSM French buildings have been mostly imported through mass imports from the French cadaster, so a lot of OSM building elements actually map small and weirdly shaped pieces of building. This is why the tiniest fictional building was not seen as an outlier given the strong presence of small sized elements in the dataset. Therefore we should reconsider geometric attributes that would not bring out the geometric specificity of geographic objects. Eventually, our input variables did not take into account the building's spatial relations with other elements. Here, some vandalized buildings are contained inside a river, a construction area and intersect a cemetery. Considering additional topological variables that express these peculiar situations might improve the detection of uncommonly located vandalized building. Actually, we did not expect to successfully detect all our vandalism cases – without any false positive – by simply using a clustering method on geometric features, so this first result is fairly encouraging.

5 Conclusion and future work

Our work focused on the definition of vandalism and the aspects that challenge its automated detection, such as the contributor's purpose, the context and the harm done. Initial experimental results showed that detecting OSM vandalism using an unsupervised method requires a wiser choice of the attributes to be input in the clustering algorithm. These attributes cannot be simple data quality assessment features but they have to be specifically designed for vandalism detection.

Future work includes exploring the influence of the σ and m parameters of DENCLUE clustering algorithm on the outlier detection predictions. Other clustering algorithms (*e.g.* DBScan, BIRCH) should also be tested to check if they perform better on detecting vandalism. Besides, we intend to carry out the same experiment on OSM German buildings because most of them have been mapped by hand, so unlike OSM French buildings, they should not be divided up into small pieces: maybe in this dataset our vandalism cases would be detected. We also intend to deal with other types of vandalism, for instance vandalism through tag edits or object delete. In this case, other relevant variables should be contemplated to enrich our dataset – as mentioned previously, the set of input clustering variables should be extended with topological, semantic and historical features, as well as contributor-oriented descriptors and reference data matching indicators. However we will then have to address the curse of dimensionality issue. Eventually, in the same way as with Wikipedia vandalism, supervised learning classification techniques may be contemplated to detect vandalism in OSM.

References

- 1 Adler, Luca Alfaro, Santiago M. Mola-Velasco, Paolo Rosso, and Andrew G. West. Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 6609 of *Lecture Notes in Computer Science*, chapter 23, pages 277–288. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. doi:10.1007/978-3-642-19437-5_23.
- 2 Andrea Ballatore. Defacing the map: Cartographic vandalism in the digital commons. *The Cartographic Journal*, 51(3):214–224, 2014. doi:10.1179/1743277414y.0000000085.
- 3 Jean-François Girres and Guillaume Touya. Quality assessment of the french OpenStreetMap dataset. *Transactions in GIS*, 14(4):435–459, aug 2010. doi:10.1111/j.1467-9671.2010.01203.x.
- 4 Stefan Heindorf, Martin Potthast, Benno Stein, and Gregor Engels. Vandalism detection in wikidata. In *Proceedings of the 25th ACM International on Conference on Information*


- and Knowledge Management - CIKM '16*, pages 327–336. ACM Press, 2016. doi:10.1145/2983323.2983740.
- 5 Alexander Hinneburg and Hans H. Gabriel. DENCLUE 2.0: Fast clustering based on kernel density estimation. In Michael R. Berthold, John S. Taylor, Nada Lavrac, Michael R. Berthold, John S. Taylor, and Nada Lavrac, editors, *IDA*, volume 4723 of *Lecture Notes in Computer Science*, pages 70–80. Springer, 2007. doi:10.1007/978-3-540-74825-0_7.
 - 6 Pascal Neis, Marcus Goetz, and Alexander Zipf. Towards automatic vandalism detection in OpenStreetMap. *ISPRS International Journal of Geo-Information*, 1(3):315–332, nov 2012. doi:10.3390/ijgi1030315.
 - 7 Philip G. Zimbardo. A Social-Psychological analysis of vandalism: Making sense of senseless violence. Technical report, Stanford University, Department of Psychology, 1971.

A Conceptual Framework for Representation of Location-based Social Media Activities

Xuebin Wei

Department of Integrated Science and Technology, James Madison University, 701 Carrier Dr, Harrisonburg, VA 22807, USA


weixx@jmu.edu

 <https://orcid.org/0000-0003-2197-5184>

Xiaobai Angela Yao

Department of Geography, University of Georgia, 210 Field St., Athens, GA 30602, USA

xyao@uga.edu

 <https://orcid.org/0000-0003-2719-2017>

Abstract

This research develops a conceptual framework for the representation and analysis of location-based social media activities (LBSMA) in GIS. With increasing popularity of location-based social networking, social media platforms have become new channels to observe human activities in physical and virtual worlds. At the same time, there is a shift of some human interactions from the physical space to the virtual social space. Traditional geographical representation in GIS is not sufficient to handle the increased sophistication of human activities related to, or embedded in, location-based social media data. This research proposes an ontology for the location-based social media activity data and a conceptual framework for them to be modeled in a GIS environment so that interconnections of human activities in spatial-temporal-social dimensions can be represented, organized, retrieved, analyzed, and visualized in the system.

2012 ACM Subject Classification Information systems → Data management systems

Keywords and phrases GIS, Social Media, Ontology, Location-based Social Media Activity

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.62

Category Short Paper

Supplement Material Study Website: www.lbsocial.net

1 Introduction

Understanding human dynamics through human activities has been an important geographic inquiry in the literature. Researchers have studied human behavior from various perspectives. Behavior geography concerns the cognitive process of human behavior and draws on works in other fields such as psychology, physiology and economics. Another thread of research examines human activities through visualization, analysis, and modeling of human dynamics. Our research attempts to contribute to the latter, motivated by two fundamental issues. First, the growing popularity of computer network-based social media and the availability of data from these social media provide an unprecedented opportunity to study human activities in new lights. However, the new types of data require new conceptualization, new methodologies, and new tools to make the best out of them. Secondly, it has been well recognized that social connections play an important role in human behavior. However, social network has been ignored or oversimplified in current representations of human activities in current off-the-shelf GIS programs. Therefore, this research aims to develop a GIS conceptual



© Xuebin Wei and Xiaobai Angela Yao;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 62; pp. 62:1–62:7

Leibniz International Proceedings in Informatics



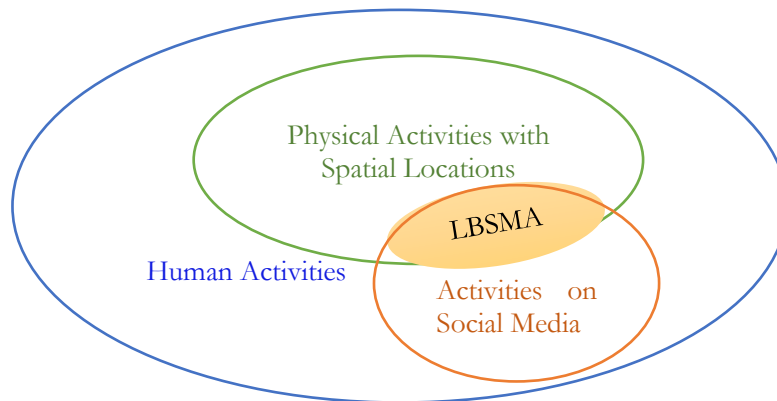
LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

framework and associated logical models to represent space, time, and social connections from location-based social media activity (LBSMA) data in GIS.

By nature of the topic, studying human activities ideally requires data at the individual level, or so called disaggregated level, with fine spatial and temporal granularities. However, commonly available census data are usually aggregated. Thus the availability of location-based social media (LBSM) data provides an unprecedented opportunities for this type of research, as the LBSM data are inherently entered on individual basis and are of high granularities in space and time. In the information age, a message from social media is considered an extension of human mind [10]. Furthermore, details of human activities can now be extracted from the social media to reveal when and where people interact with others. Collections of such interactions can be used to develop social networks among people. This is particularly advantageous for research on human activities, as the context of social connection is particularly important to human activities. It has been argued that time, space and social differentiation should be coupled in the study of practices or phenomena [7]. From the relationalism-idealism perspective, the assumed existence of social networks sets the scope to which space and time should be conceptualized and analyzed in human activity analysis [12]. Different types of social media allow for different types of connections. For example, Twitter fosters an asymmetric network structure that people prefer to broadcast individual activities, while LinkedIn and Facebook capture pre-existing ties by focusing on social interactions among friends [9]. Previous studies have investigated the content and friendship structure on Twitter [9][5], Facebook [3] and Weibo [4]. The spatial distribution of location-based social activities from different social media has also been explored in recent studies.

There is a long tradition that human activities are visually represented and analyzed, particularly in GIS. Starting from the space-time prism [1], trajectories of human activities are visually represented as a series of locations in space-time dimensions. Because human activities have innate spatial component, geographic information system (GIS) is naturally the most desirable environment for the visualization and analysis of it. Sui and Goodchild [8] suggest that GIS is a media for communicating and sharing knowledge and supporting location-based social networking. Meanwhile, the convergence of geographic information systems (GIS) and social media has resulted in a data avalanche that creates new challenges in GIScience [8]. Although location-based social media activities have been examined in many studies, a structured GIS representation is still absent for all three dimensions of space, time, and the context of social connections in which activities take place. GIS representation of space and time alone is already a critical research theme in the literature [11], adding more dimensions obviously is not a trivial issue. The goal of this paper is to fill the gap by developing an ontological framework and a conceptual model for the representation location-based social media activity (LBSMA) data in GIS, so that the space, time, and social connections associated with LBSM activities can be represented and analyzed further. In this research, as shown in Figure 1, the LBSMA refer to the subset of human activities of which locations can be georeferenced in the geographical space and contents are advertised in the networked social media. The scope of the study is limited to human activities that are recorded explicitly or implicitly in the LBSMA data, which is illustrated as the yellow-highlighted area in Figure 1.

The paper is organized as follows. The next section presents an ontological framework for LBSMA and a conceptual model for the representation of LBSMA in GIS. Section 3 introduces a pilot implementation based on the framework. A case study is conducted in the prototype. The paper is concluded with discussions in Section 4.



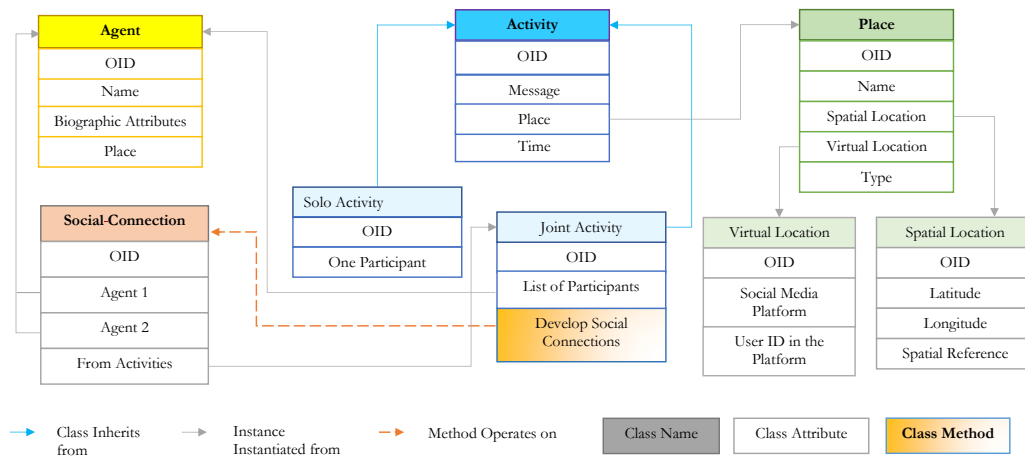
■ **Figure 1** Location-based Social Media Activity.

2 LBSMA Ontological Framework and Representation in a GIS environment

Traditional GIS conceptual models use either object-based or field-based representations. The former distinguishes each spatial object with delineated spatial boundaries, while the latter enumerates all spatial locations systematically and stores attribute values for each location. However, none of them is able to directly account for social network (or social associations) or human activities in the context of such a network. Aiming to have a conceptual underpinning for later technical deployment to fill the gap, this paper first develops an ontological framework that identifies four primary categories for the LBSMA. They are Agents, Activities, Places, and Social Connections. Following the ontological framework, a conceptual data model is designed in the paradigm of object-oriented modeling. The model is illustrated in Figure 2. The purpose of conceptual model is to organize the data in a reasonable and retrievable way, so as to maximize the possibility to study hidden relationships and patterns in the spatio-temporal big data of growing size. The most important aspect is to allow information in the spatio-temporal-social dimensions, expressed either explicitly or implicitly, to be identified and represented in the system.

2.1 Agents

Here an agent refers to a person or a collective entity with a group of individuals as long as the entity has a unique ID in a social media platform. A person may have one or more active accounts in social media platforms and thus may be associated with multiple agents. An agent can participate in any number of activities, some of which show social connections with other agents. It is also possible that through analytical methods, multiple agents are found to be the same person in the real world. A collective entity, such as a restaurant, a university, or an association, which has an official account in a social media platform can be identified as an agent too. It is often more likely for such an agent to identify itself with its official name and thus can be readily associated with its real world identity.



■ **Figure 2** Conceptual Model of LBSMA.

2.2 Places

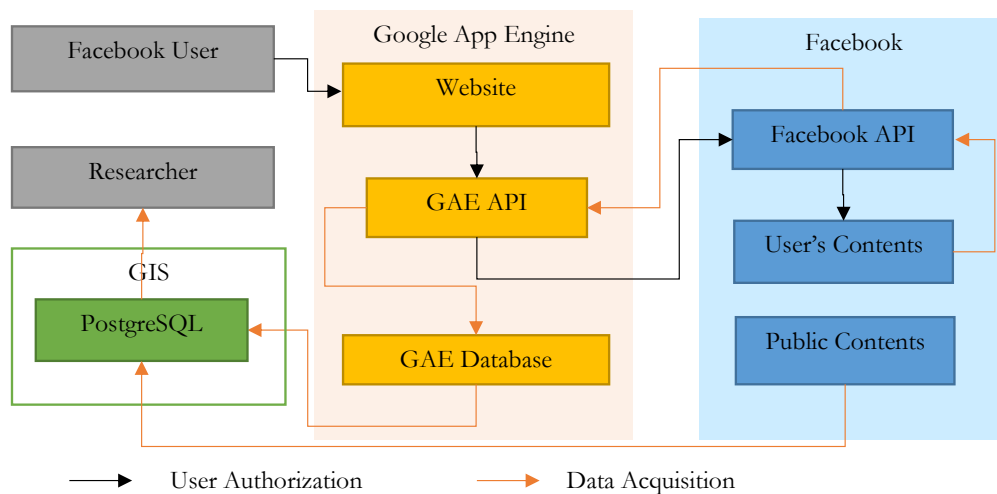
An agent may exist in a physical geographic space, a virtual social media space, or in both spaces. Uniquely identifiable locations in either type of frame of references are called places in this ontology. Therefore two types of locations are distinguished in the framework: the geographic location (or spatial location) and the virtual location. For example, just like an address can refer to the geographic location of a person’s home, a uniform resource locator (URL) of a user’s profile page refers to the person’s virtual location on social media. Sometimes, information of both types of locations may be available for an agent. For instance, a restaurant can have its footprints in the geographical world, while its URL is a location in the virtual world where its menu, public reviews or other types of information can be retrieved. Virtual locations are equally important in the framework because they not only facilitate the organization of human activities in the virtual world, but also provide the source of rich information about people, activities and the context of environment.

2.3 Activities

An activity in this ontology refers to any action of an agent in either the physical space or the virtual space. For example, visiting, commuting, reading, participating in a party are examples of activities in the physical world, while posting a message and following another account on Facebook or Twitter are examples of activities in the virtual space. Activities can be further classified into solo activities and joint activities based on the number of participants. Because the scope of this study is the intersection area of human activities in the physical space and the social media space, the activities of concern are those reflected in a social media regardless which space the actual activity took place.

2.4 Social Connections

In this framework, social connections are personal relationships expressed via social interactions. It is a mind-dependent construct that can be reflected by mind-independent human activities. Social connections can be explicitly expressed or identified through their self-reported relationships such as kinship, workplace connection, friendship, and so on, which can be explicitly indicated in the profiles or implicitly revealed via connections between



■ **Figure 3** LBSMA Data Collection.

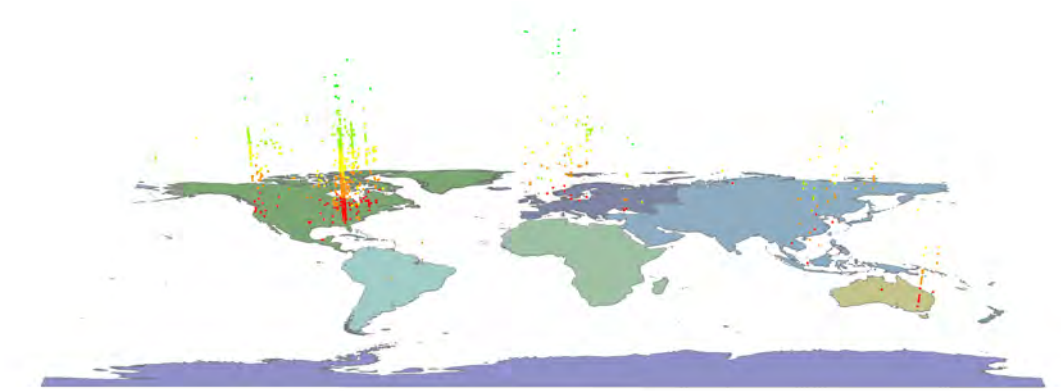
profiles on social media platforms. However, many additional social connections can be identified through spatial-temporal reasoning. For instance, two agents may already have, or are potentially developing, a social connection if they participate in joint-activities. Frequent joint-activities at the same home address suggests close family ties or friendship.

3 Pilot Prototype Implementation and Case Study

Based on the proposed conceptual model, a pilot prototype has been implemented and tested. A case study is performed to validate the prototype and most importantly to evaluate the usefulness of the proposed framework of LBSMA representation in GIS. This research has widely recruited students in the University of Georgia to collect their Facebook data. The extracted Facebook data are organized in the implemented LBSMA data model. About 500 unique Facebook accounts and 2,500 posts have been collected in this case study.

This website obtained the Institutional Review Board (IRB) and Facebook App approval, and has gathered participants' information through the Facebook Application Programming Interface (API) with explicit authorizations of Facebook users. The website is running on the Google Application Engine (GAE). The collected LBSMA data is then organized and maintained in a PostgreSQL database with the PostGIS plugin to provide GIS functions. Figure 3 shows the process. When a post is received, the name and user account ID are recorded in an Activity table. Other people who are tagged in the same post are also kept in the Participants field. Since the number of the tagged people is not predictable, this data filed utilize the JSON format to record all the participants.

The prototype has developed a set of visualization and analysis tools for the LBSMA in ArcGIS, including visualize activities and places, query people-based social network, create location-based social network and identify spatial-temporal interactions of activities. The demo of the developed tools can be found at <https://www.youtube.com/watch?v=aJnmOGTqV5w&list=PLHutrxqbP1BxvYmCOGX5fDQkLYUrbqsS5>



■ **Figure 4** Visualization of Activities in Space and Time Dimensions.

3.1 Visualize Activities and Places in GIS

This tool reads the activity table and place table from PostgreSQL in ArcGIS, and displays the places and activities on a map. In addition, since the activity records have the time stamps, the activities can also be visualized on a 3D map in which the z coordination represents the time an activity took place. The map in Figure 4 is such a map showing the data collected from the case study.

3.2 Create a Place-Based Social Network

This tool allows users to interactively select the places in ArcGIS, and create a social network of the visitors from those places. Participants in the same activity are connected to each other in the social network. In addition to visualize the location-based social network, some network measures, e.g., number of nodes, number of cliques, average clustering coefficient and etc. are also reported in the output.

3.3 Query People-Based Social Network

This tool allows users to query the people-based social network based on a user-defined query which will be translated into an SQL sentence. The user can also visualize and analyze the social networks for the selected people.

3.4 Identify Spatial-Temporal Interactions

The spatial-temporal interactions are identified with the Knox test [2] by using the Pysal python library [6]. This tool reports the identified spatial-temporal interactions based on the user-defined spatial (δ) and temporal (τ) intervals.

4 Conclusion

Current GIS environment is not suitable for the representation and analysis of rich information embedded in location-based social media data due to its lack of capability to represent some of the key components of the data. This research aims to fill the blank by developing an ontological framework and a conceptual data model for multiple-dimensional representation

of geography, time, and social connections. The prototype of the conceptual model is implemented and a case study is carried out. The effectiveness of the LBSMA model is evidenced by a case study which collects and analyzes Facebook data.

The findings of this research yield new insights regarding human activities in virtual and physical space, and will enhance technical capabilities for social media analysis in GIS. The developed methods can help identify place-based or people-based strategies, e.g., urban planning, traffic planning, commercial advertising or energy communicating. The proposed framework paves new avenues for future research, such as public health, transportation, urban geography and social science. Based on the proposed model and prototype, we believe there are many more potential ways to mine the organized datasets. This study has only provided a case study with a few application examples, both of which asked questions that are only related to two dimensions of space, time, and social network. Starting from here, many exciting future research avenues should be explored. Examples include development of new analytical methods and explorations of new application studies, particularly those involve all three dimensions of the LBSMA data.

References

- 1 Torsten Hägerstrand. What About People in Regional Science? *Papers in Regional Science*, 24(1):7, 1970.
- 2 E. G. Knox and M. S. Bartlett. The Detection of Space-Time Interactions. *Applied Statistics*, 13(1):25, 1964.
- 3 Kevin Lewis, Jason Kaufman, Gonzalez A. Marco, Wimmer B. Andreas, and Christakis A. Nicholas. Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks*, 30:330–342, 2008.
- 4 Yuan Li, Haoyu Gao, Mingmin Yang, Wanqiu Guan, Haixin Ma, Weining Qian, Zhigang Cao, and Xiaoguang Yang. What are Chinese Talking about in Hot Weibos? *arXiv preprint arXiv:1304.4682*, 2013.
- 5 Mor Naaman, Jeffrey Boase, and Chih-Hui Lai. Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 189–192. ACM, 2010.
- 6 Sergio J. Rey and Luc Anselin. PySAL: A Python Library of Spatial Analytical Methods. *The Review of Regional Studies*, 37(1):5–27, 2007.
- 7 Tim Schwanen and Mei-Po Kwan. Critical Space-Time Geographies: Guest Editorial. *Environment and Planning A*, 44(9):2043–2048, 2012.
- 8 Daniel Sui and Michael Goodchild. The convergence of GIS and social media: challenges for GIScience. *International Journal of Geographical Information Science*, 25(11):1737, 2011.
- 9 Yuri Takhteyev, Anatoliy Gruzd, and Barry Wellman. Geography of Twitter networks. *Social Networks*, 34(1):73–81, 2012.
- 10 Ming-Hsiang Tsou and Michael Leitner. Visualization of social media: seeing a mirage or a message? *Cartography and Geographic Information Science*, 40(2):55, 2013.
- 11 Xiaobai Yao. Modeling and analyzing cities as spatio-temporal places. In Bin Jiang and Xiaobai Yao, editors, *Geospatial Analysis and Modelling of Urban Structure and Dynamics*, GeoJournal Library, pages 311–328. Springer, Dordrecht, the Netherlands, 2010.
- 12 May Yuan, Atsushi Nara, and James Bothwell. Space–time representation and analytics. *Annals of GIS*, 20(1):1–9, 2014.

Towards the Statistical Analysis and Visualization of Places

René Westerholt

Heidelberg University, Institute of Geography, Heidelberg, Germany
westerholt@uni-heidelberg.de

Mathias Gröbe

TU Dresden, Institute for Cartography, Dresden, Germany
mathias.groebe@tu-dresden.de

Alexander Zipf

Heidelberg University, Institute of Geography, Heidelberg, Germany
zipf@uni-heidelberg.de

Dirk Burghardt

TU Dresden, Institute for Cartography, Dresden, Germany
dirk.burghardt@tu-dresden.de

Abstract

The concept of *place* recently gains momentum in GIScience. In some fields like human geography, spatial cognition or information theory, this topic already has a longer scholarly tradition. This is however not yet completely the case with statistical spatial analysis and cartography. Despite that, taking full advantage of the plethora of user-generated information that we have available these days requires mature place-based statistical and visualization concepts. This paper contributes to these developments: We integrate existing place definitions into an understanding of places as a system of interlinked, constituent characteristics. Based on this, challenges and first promising conceptual ideas are discussed from statistical and visualization viewpoints.

2012 ACM Subject Classification Information systems → Geographic information systems, Theory of computation → Probabilistic computation, Applied computing → Cartography

Keywords and phrases Spatial Analysis, Visualization, Statistics, Geosocial Media

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.63

Category Short Paper

Funding This work was generously supported by the German Research Foundation (DFG) through the priority programme ‘Volunteered Geographic Information: Interpretation, Visualisation and Social Computing’ (SPP 1894).

1 Introduction

People utilize place for the mental representation of geographic phenomena, to verbalize locations in colloquial conversations, and to orientate themselves geographically [25, 28, 23]. A place may thereby refer to either material or immaterial entities [19] and, most generally, describes a location together with a set of attached meanings [5]. While place has been of recurring importance (e.g., place was crucial to Aristotle, to German geographers of the late 19th century, and to human geographers since the 1970s [5]), the concept has only



© René Westerholt, Mathias Gröbe, Alexander Zipf, and Dirk Burghardt;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 63; pp. 63:1–63:7

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

recently gained greater attention in GIScience. For instance, *patial*¹ approaches are still uncommon in spatial analysis and geovisualization. Due to its importance, Mike Goodchild has anticipated a place-based account of geographical information systems (GIS) enabling us to benefit from the latest wealth of subjective user-generated (and thus largely *patial*) geographical information [13, 12]. This paper contributes to these developments by discussing conceptual *patial* statistical and visualization challenges.

2 **Places as systems of interlinked characteristics**

Places are versatile and have been treated differently: as named domains occurring in human discourse [13], sets of realized or unrealized affordances [23, 19], functional relationships between humans and locations [20], or as references to events and entities [28]. What all these definitions have in common is the concept of places as locations with meaning [25], including a locale (the material setting found in a location), and a sense of place. Each of the outlined definitions is useful for a particular aim. We argue here, however, that these aspects can also be considered together simultaneously, by following the systemic tradition of geography emphasized by Alfred Hettner and others [14]. The affordances and other properties of places are often interlinked reciprocally with how people perceive and mentally represent the geometric, temporal, and perceptual characteristics of places. We therefore suggest a combined viewpoint emphasizing that many of the previously described place dimensions are interrelated. This viewpoint largely reflects what Anderson et al. have recently called *assemblage thinking* (including both relations and things) [1]. Further, certain phenomena can only occur in places if all relevant contextual characteristics are fulfilled. The formulation of a reasonable and realistic conceptual place-based counterpart to the field of spatial analysis therefore requires a combined viewpoint instead of accounting for isolated components of places individually. The following paragraphs explicate and utilize this viewpoint.

3 **Analysing places statistically**

The subjectivity of places is in stark contrast to spatial and conventional statistics. The latter often assume *identically* distributed observations [3] through the notions of intrinsic or second-order stationarity (i.e., stable moments up to some order) [10], guaranteeing that all observations originate from the same process and in turn allowing the estimation of statistical properties. The subjectivity of perceived places runs counter to this. Different people apply idiosyncratic modes of perception, verbalize subjective opinions, and assign varying complex meanings to places. Further, because place is heavily rooted in spatially and temporally diverse context, even the *patial* expressions of only single individuals are not coherent and thus not necessarily comparable. This raises questions about suitable methodological and conceptual approaches.

3.1 *Patial* index sets and units

Mike Goodchild describes the geographic world as a “set of overlapping continua” [11, p. 36]. These continua represent spatially and temporally superimposed places [12], reflecting that different people represent and verbalize their very own subjective places in the same locations

¹ The term *patial* is used here as a complement to the term *spatial*.

and times simultaneously. Place can therefore only be treated in limited terms in the sense of spatial analysis, which is based on spatially exclusive observations. The latter is reflected by the types of spatial indexes applied: geostatistical, lattice-based, or spatially stochastic units prevail [4]. Reducing places to their spatial domain is thus insufficient, and recent results obtained this way revealed issues in terms of the reliability of spatial-statistical results and with respect to drawn conclusions [26, 27]. Platial analysis requires index sets, methods, and concepts beyond the spatial domain.

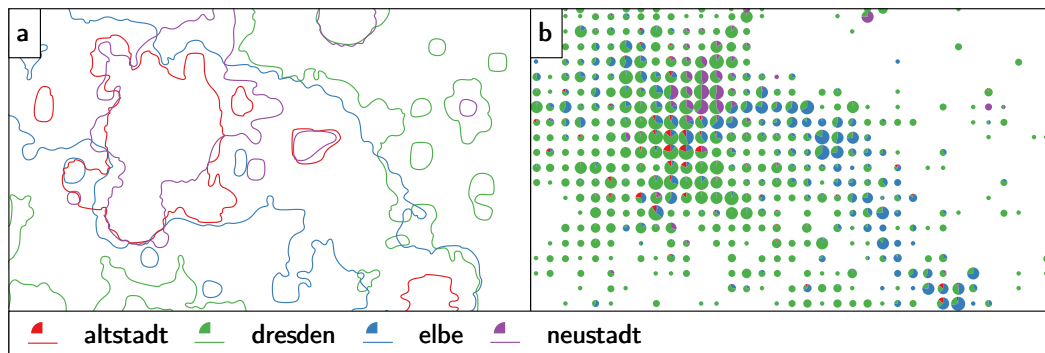
It is the context which allows certain phenomena to occur. A platial notion of index sets should thus take account of multiple contextual conditions in the combined way suggested in Section 2. The contextual dimensions should thus form part of the domain over which platial phenomena are defined. For example, a regular workplace must allow the respective work to be carried out. At the same time, different employees will attach certain idiosyncratic senses of place to the respective location, which are also linked to emotional attachments and other influences. In a platial perspective, these conditions are not treated as spatially referenced attributes, but are part of the coordinate system that allows phenomena to occur in the workplace. The phenomenon *chatting with colleagues at work* is then not defined in terms of space and time, but also in terms of coordinates reflecting the outlined contextual factors. The context thus determines platial units, which are elements of a platial index set and thus of a platial domain.

One possibility to conceptualize platial units could be the definition of regions in *conceptual spaces* [9]. A conceptual space C is spanned by so-called quality dimensions $q_1 \times \dots \times q_n$, which, following their original interpretation, represent how people judge stimuli to be similar. These dimensions are well-suited for representing saliency, which places have to fulfill in order to allow people distinguishing different places [28]. Quality dimensions further represent psychological *integral dimensions*, which are roughly interpreted as decompositions of perceptual stimuli into their base components. Thinking of places as concepts, and of their subjective properties as quality dimensions, regions in multidimensional conceptual spaces could form platial units in analogy to spatial units like administrative regions or raster cells. Such platial units would automatically meet the conjectured container property allowing objects and processes to be “in” a place [28]. Similar to measures of spatial and geographic distance, the calculus offered by conceptual spaces could further be used to define distance relationships between places. However, the framework is borrowed from cognitive science and applying it to places will require future technical adaptations.

3.2 Platial concepts and data

Operationalizing platial analysis requires the definition of mature adapted statistical concepts. For example, it is unclear what the study of *platial autocorrelation* would mean. Spatial autocorrelation is a key concept in spatial analysis and it refers to an association between correlation-based pattern within attributes with some notion of spatial distance [7]. This characteristic largely resembles the so-called *first law of geography* [24]. Gao et al. suggest a place-based counterpart to this law by stating that “every place is related to other places, but more similar places are more interlinked” [8]. Still, in the light of the different discussed available notions of place, it is yet unclear what exactly *interlinked* means in a generalizable sense. By analogy, concepts like heterogeneity, stationarity, and randomization must be coherently defined to enable a solid statistical theory of platial analysis.

One promising source of platial information is user-generated geographic information, like those extracted from geosocial media feeds. These datasets reflect peoples’ subjective impressions, which is why they have been conjectured to be of platial instead of geospatial



■ **Figure 1** Maps of the named places *Dresden*, the *Elbe* river, and the urban districts *Altstadt* and *Neustadt*, based on Flickr tags visualized as (a) isolines and (b) micro diagrams.

nature [21]. Cognitive psychology shows that meaningful thoughts and experiences are stored in the *long-term memory*, especially in the fraction called *episodic memory* [22]. Given the meaningful nature of places, geosocial media data raises the question of the extent to which the messages posted on such feeds originate from long-term memory. If large portions of the posted contents may reflect short-term (and thus non-platial) information, it would be questionable whether geosocial media is a useful source of platial information. It is instead likely that the data found on these feeds represents a mixture of platial and non-platial information, making it difficult to interpret obtained analysis results. An alternative possible source of platial information is data collected through survey techniques like the event-sampling method (ESM) [2]. This technique enables the collection of in-situ information by triggering context-based surveys. These and related methods thus allow to collect platial information in a systematic manner. Future research should clarify to what extent user-generated information and the ESM technique are useful for investigating human platial experiences.

4 Visualizing Places

Communicating results derived from platial analysis requires new techniques and strategies. Below we present ideas for visualizing places through an example using data from the photo-sharing platform *Flickr*. Places are frequently extracted and visualized from this kind of data by using selected assigned tags [15, 17]. A more sophisticated approach combining multiple, different sources into a joint classification of topics and thematic regions is found in [19]. Another frequently applied method of visualizing spatially continuous qualitative areas is kernel density estimation. The isoline method is an alternative approach to this, the results of which can be portrayed along with the underlying data points [15]. The approaches outlined present pre-processed analysis results to the viewers of maps and visualizations. Another related idea called tag maps [6] is to avoid extracting places a priori, but to instead show all contained tags in one map and to let the viewer decide about reasonable places. It is further possible to derive 2.5-dimensional pseudo surfaces known from GIS. For this, interpolation methods like inverse distance weighting (IDW) can be used in a first step to produce surfaces, from which isolines or hill shadings can then be derived. The dominance of place representations can be visualized through proportional symbol grid maps or by means of grid choropleths [16, p. 137].

The approaches presented are not optimally suited for the visualization of mental place representations. For example, kernel density maps create the wrong impression of a spatial continuum, which might be misleading for some types of places. Further, when combining information derived from multiple datasets, various individually estimated kernel density surfaces may not be one-by-one comparable, making their joint mapping problematic. Similarly, because places are characterized by multiple dimensions, it is often of interest to visualize more than just one attribute, as it is the case with the outlined techniques. We thus suggest the aggregation of point-based data through regular grids based on point counts as a viable alternative to the outlined interpolation approaches. In this way, the viewer at least does not get the wrong impression of a possibly non-existent surface.

The micro diagram method is another promising approach for mapping diverse places [18]. This method utilizes different kinds of diagrams to represent multiple types of aggregated qualitative information. We show the potential of this approach for the visualization of places using an example based on Flickr data from Dresden. Figure 1a shows the spatial extents of named places based on their occurrences in the Flickr tags. The visualization is based on isolines extracted from a statistical surface estimated by IDW. This type of visualization demonstrates the aforementioned superimposed nature of subjective platial verbalizations. In contrast, Figure 1b shows the results for the micro diagram method, which shows the detailed quantitative composition of the locations in terms of how people interpreted them as places. Other than in 1a, the Elbe river is now notable (blue), and the *Altstadt* (red) and the *Neustadt* (violet) are distinguishable. Beyond this proposed symbology, we suggest avoiding the use of background maps, classical scale bars and other cartographic elements to avoid the impression of a one-to-one mapping between space and place, which may not always exist. Such an omission, however, requires the viewer to have a certain topographical knowledge of the respective region.

5 Conclusions and future research

Investigating places is important for gaining a thorough understanding of peoples' everyday lives and to obtain insights on the perceived structures of urban areas. Current statistical approaches from spatial analysis are not suited for this. We discussed challenges and useful solution paths that may bring us closer to the long-term vision of a platial analysis framework. One major challenge is to find suitable units upon which statistical analyses of places can be conducted. Conceptual spaces have been identified as one promising way to define such units, though an in-depth harmonization of this framework with places still needs to be done in future work. Further, platial counterparts to important spatial-statistical concepts must be formulated in order to develop a valid and rigorous statistical theory of places. It is not yet clear to what extent data taken from user-generated feeds is truly platial. Since data is a crucial ingredient to achieving insights on places, this is one of the major empirical steps to be undertaken in the near future. In terms of visualizing places, the major issues with current approaches include wrong spatial impressions created through interpolation techniques, the problem of displaying multifaceted place-based information at once, and the combination of different subjective places in one map. However, the proposed example using micro diagrams has shown first promising results for the presentation of multidimensional, qualitative information together with the spatial outline of places in a conceivable way.

References

- 1 Ben Anderson, Matthew Kearnes, Colin Mcfarlane, and Dan Swanton. On Assemblages and Geography. *Dialogues in Human Geography*, 2(2):171–189, 2012. doi:10.1177/2043820612449261.
- 2 Matthias Bluemke, Clemens Lechner, Bernd Resch, René Westerholt, and Jan-Philipp Kolb. Integrating Geographic Information into Survey Research: Current Applications, Challenges, and Future Avenues. *Survey Research Methods*, 11(3):307–327, 2017. doi:10.18148/srm/2017.v11i3.6733.
- 3 Yongwan Chun and Daniel Griffith. *Spatial Statistics and Geostatistics*. SAGE, London, UK, 2013.
- 4 Noel Cressie. *Statistics for Spatial Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ, 1993. doi:10.1002/9781119115151.
- 5 Tim Cresswell. Place. In Nigel Thrift and Rob Kitchin, editors, *International Encyclopedia of Human Geography*, volume 8, pages 169–177. Elsevier, Oxford, UK, 2009. doi:10.1139/h2012-055.
- 6 Alexander Dunkel. Visualizing the perceived environment using crowdsourced photo geodata. *Landscape and Urban Planning*, 142:173–186, 2015. doi:10.1016/j.landurbplan.2015.02.022.
- 7 Manfred Fischer and Arthur Getis. Introduction. In Manfred Fischer and Arthur Getis, editors, *Handbook of Applied Spatial Analysis*, pages 1–24. Springer, Heidelberg, 2010. doi:10.1007/978-3-642-03647-7_1.
- 8 Song Gao, Krysztof Janowicz, Grant McKenzie, and Linna Li. Towards Platial Joins and Buffers in Place-Based GIS. In *Proceedings of The First ACM SIGSPATIAL International Workshop on Computational Models of Place - COMP '13*, pages 42–49, Orlando, FL, 2013. doi:10.1145/2534848.2534856.
- 9 Peter Gärdenfors and Mary-Anne Williams. Reasoning About Categories in Conceptual Spaces. In *International Joint Conference on Artificial Intelligence*, pages 385–392, Seattle, WA, 2001.
- 10 Tilmann Gneiting and Peter Guttorp. Continuous Parameter Stochastic Process Theory. In Alfred Gelfand, Peter Diggle, Montserrat Fuentes, and Peter Guttorp, editors, *Handbook of Spatial Statistics*, pages 17–28. CRC Press, Boca Raton, FL, 2010.
- 11 Michael Goodchild. Geographical Information Science. *International Journal of Geographical Information Systems*, 6(1):31–45, 1992. doi:10.1080/02693799208901893.
- 12 Michael Goodchild. Space , Place and Health. *Annals of GIS*, 21(2):97–100, 2015. doi:10.1080/19475683.2015.1007895.
- 13 Michael Goodchild and Linna Li. Formalizing Space and Place. In *Proceedings du 1er Colloque International du CIST*, pages 177–183, Paris, 2011.
- 14 Alfred Hettner. Das Wesen und die Methoden der Geographie. *Geographische Zeitschrift*, 11(10):545–564, 1905.
- 15 Livia Hollenstein and Ross Purves. Exploring place through user-generated content: Using Flickr to describe city cores. *Journal of Spatial Information Science*, 1(1), 2010. doi:10.5311/JOSIS.2010.1.3.
- 16 M. J. Kraak and Ferjan Ormeling. *Cartography: visualization of geospatial data*. Prentice Hall, Harlow ; New York, 3rd ed edition, 2010. OCLC: ocn477062041.
- 17 Linna Li and Michael F. Goodchild. Constructing Places from Spatial Footprints. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, GEOCROWD '12, pages 15–21, New York, NY, USA, 2012. ACM. doi:10.1145/2442952.2442956.

- 18 Mathias Gröbe and Dirk Burghardt. Micro Diagrams: A Multi-Scale Approach for Mapping Large Categorized Point Datasets. In *The 20th AGILE International Conference on Geographic Information Science*, Wageningen, 2017.
- 19 Grant Mckenzie and Benjamin Adams. Juxtaposing Thematic Regions Derived from Spatial and Platial User-Generated Content. In Eliseo Clementini, Marueen Donnelly, May Yuan, Christian Kray, Paolo Fogliaroni, and Andrea Ballatore, editors, *Proceedings of the 13th International Conference on Spatial Information Theory (COSIT 2017)*, L'Aquila, 2017. doi:10.4230/LIPIcs.COSIT.2017.20.
- 20 Jeremy Mennis and Michael Mason. Modeling Place as a Relationship between a Person and a Location. In *Proceedings of the 9th International Conference on GIScience*, pages 2014–2017, Montréal, CA, 2016. doi:10.21433/B3119W316472.
- 21 Teriitutea Quesnot and Stéphane Roche. Platial or Locational Data? Toward the Characterization of Social Location Sharing. *Proceedings of the Annual Hawaii International Conference on System Sciences*, pages 1973–1982, 2015. doi:10.1109/HICSS.2015.236.
- 22 Andrew Rutherford, Markopoulos Gerasimos, Davide Bruno, and Mirjam Van den Bos. Long-Term Memory. In Nick Braisby and Angus Gellatly, editors, *Cognitive Psychology*, pages 229–265. Oxford University Press, Oxford, UK, 2 edition, 2012.
- 23 Simon Scheider and Krisztof Janowicz. Place Reference Systems. *Applied Ontology*, 9(2):97–127, 2014. doi:10.3233/A0-140134.
- 24 Waldo Tobler. A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, 46:234–240, 1970. doi:10.2307/143141.
- 25 Yi-Fu Tuan. *Space and Place: The Perspective of Experience*. University of Minnesota Press, Minneapolis, MN, 1977.
- 26 René Westerholt, Bernd Resch, and Alexander Zipf. A Local Scale-Sensitive Indicator of Spatial Autocorrelation for Assessing High- and Low-Value Clusters in Multiscale Datasets. *International Journal of Geographical Information Science*, 29(5):868–887, 2015. doi:10.1080/13658816.2014.1002499.
- 27 René Westerholt, Enrico Steiger, Bernd Resch, and Alexander Zipf. Abundant Topological Outliers in Social Media Data and Their Effect on Spatial Analysis. *PLOS ONE*, 11(9):e0162360, 2016. doi:10.1371/journal.pone.0162360.
- 28 Stephan Winter and Christian Freksa. Approaching the Notion of Place by Contrast. *Journal of Spatial Information Science*, 5(5):31–50, 2012. doi:10.5311/JOSIS.2012.5.90.

An Experimental Comparison of Two Definitions for Groups of Moving Entities

Lionov Wiratma¹

Department of Information and Computing Sciences, Utrecht University
Utrecht, The Netherlands
l.wiratma@uu.nl

Department of Informatics, Parahyangan Catholic University
Bandung, Indonesia
lionov@unpar.ac.id

Maarten Löffler²

Department of Information and Computing Sciences, Utrecht University
Utrecht, The Netherlands
m.loffler@uu.nl

Frank Staals³

Department of Information and Computing Sciences, Utrecht University
Utrecht, The Netherlands
f.staals@uu.nl

Abstract

Two of the grouping definitions for trajectories that have been developed in recent years allow a continuous motion model and allow varying shape groups. One of these definitions was suggested as a refinement of the other. In this paper we perform an experimental comparison to highlight the differences in these two definitions on various data sets.

2012 ACM Subject Classification Computing methodologies → Model development and analysis, Theory of computation → Computational geometry

Keywords and phrases Trajectories, grouping algorithms, experimental comparison

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.64

Category Short Paper

1 Introduction

The presence of devices equipped with advanced tracking technologies, such as GPS-enabled mobile phones and RFID tags, makes it possible to easily record the position of moving entities over a period of time. The widespread use of such inexpensive devices leads to the availability of a vast amount of movement data. Consequently, in many research areas there is an increasing interest in analyzing such movement data [3, 11].

Typically, movement data is described as a *trajectory*: a path made by a moving entity over a period of time together with time stamps at the locations. Differently put, a trajectory is a continuous mapping from a time interval $I = [t_{start}, t_{end}]$ to the space in which the entity

¹ Supported by The Ministry of Research, Technology and Higher Education of Indonesia (No. 138.41/E4.4/2015)

² Supported by The Netherlands Organisation for Scientific Research on grant no. 614.001.504

³ Supported by The Netherlands Organisation for Scientific Research on grant no. 612.001.651



© Lionov Wiratma, Maarten Löffler, and Frank Staals;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 64; pp. 64:1–64:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

is moving. An analysis task that has been well studied is to extract collective movement patterns from such data. Some of the movement patterns considered are flocks [1], herds [5], convoys [7], moving clusters [8], mobile groups [6] and swarms [9]. Buchin et al. formalize the definition for another variation called *groups* [2]. They define a group of moving entities by taking into account three parameters: the spatial parameter (are the entities close enough?), the temporal parameter (does the togetherness last long enough?), and the size parameter (are there enough entities?). They implement the algorithm to compute groups and present experimental evaluation of their method using both generated and real-world datasets. In a recent paper [10], we refined the definition of groups by Buchin et al. We made a slight change in the condition for the spatial parameter and argued that the refined definition of groups is more intuitive and is expected to be better for finding the right groups in a dense environment. Consequently, this change leads to different algorithms to compute groups.

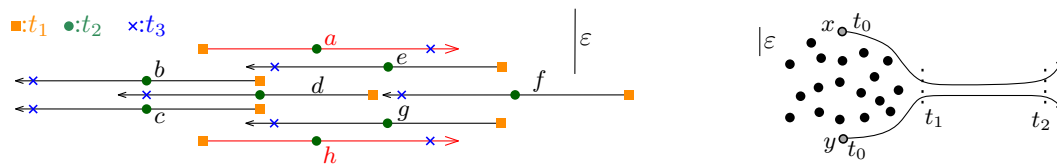
In this paper we compare the two definitions experimentally. While there are many definitions of flocks, herds, groups, etc., the last two definitions and the flocking definition are the only ones that respect the continuity of the trajectories, and do not consider only fixed time-stamped locations. We exclude the flocking definition because it uses a fixed-size circle to define closeness, which does not allow for elongated groups. To compare the two grouping definitions, we implemented the algorithm to compute groups based on the refined definition (an implementation of the other one exists) and conducted experiments on dense pedestrian data. We compare the outputs from both implementations, which is the same as comparing the two definitions of groups, since the implementations follow the definitions exactly. We analyze the claim made (by us) in [10] that the newer definition is more intuitive, especially when the environment is dense. Arguably, dense situations are especially difficult for identifying groups.

Results and Organization. In the following section, we review both definitions, and highlight their differences. Section 3 briefly describes what we expect to find in an experimental analysis where we compare the two definitions. In Section 4, we describe our experiments. We focus our evaluation on the differences of the two definitions, and thus on the maximal groups that are reported, rather than the differences between the algorithms and their implementation. Moreover, we consider only a single dataset consisting of trajectories of pedestrians walking through a narrow corridor. We conclude in Section 5 where we discuss the advantages and disadvantages of the two definitions.

2 Description and Properties of the two Definitions for Groups

The original definition of a group by Buchin et al. relies on three parameters: the number of entities in a group, the time interval in which those entities form a group and the distance between entities in the group [2]. While the first two parameters are simple to formalize, the latter needs to be described in more detail. The ε -connectivity between two entities is defined as follows: Let \mathcal{X} be a set of moving entities and consider two entities $x, y \in \mathcal{X}$. If at some time t , the Euclidean distance between x and y is at most ε ($\varepsilon > 0$), then x and y are *directly ε -connected*. Furthermore, x and y are *ε -connected in \mathcal{X}* at time t if there is a sequence $x = x_0, \dots, x_k = y$, with $x_0, \dots, x_k \in \mathcal{X}$ and for all i , x_i and x_{i+1} are directly ε -connected at time t . Then, with the maximum entity inter-distance ε , a minimum number of m entities in a group and a minimum required duration of δ , a subset $G \subset \mathcal{X}$ is a *group* during time interval I , if the following three conditions hold [2]:

- G contains at least m entities.
- I has a duration at least δ .



■ **Figure 1** (left) Entities in $G = \{a, h\}$ are ε -connected using entities not in G [10]. (right) In the original definition [2], x and y are ε -connected during $[t_0, t_2]$.

- Every pair entities $x, y \in G$ is ε -connected in \mathcal{X} during I .

Furthermore, G is a *maximal group* during time interval I if there is no time interval $I' \supset I$ for which G is also a group and there is no $G' \supset G$ that is also a group during I .

However, this definition might have a counter-intuitive effect and may not be suitable in a dense environment. In [10], we presented an example where this definition will have two entities in one group that are far apart during their entire duration as a group, see Figure 1 (left). Here, a and h are always ε -connected through different entities between t_1 and t_3 . Hence, $\{a, h\}$ form a group during the time interval $[t_1, t_3]$. Since there is no superset of $\{a, h\}$ in the same time interval I , $\{a, h\}$ is a maximal group. Intuitively, we do not view $\{a, h\}$ as a group because they are separated by other entities that move in the opposite direction. To avoid this counter-intuitive situation, we refined the definition of a group by changing the requirement on the connectivity between entities in a group:

- Every pair entities $x, y \in G$ is ε -connected in G during I .

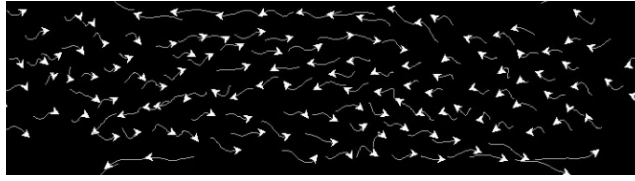
The only difference is that connectivity must happen using entities in the group G itself, and it can no longer use any entity in the whole set that is not part of the group. With this refined definition, $\{a, h\}$ is not a group because they are not ε -connected through other entities in the same group. Another example that shows the difference between the two definitions can be seen in Figure 1 (right) [10]. With the original definition, x and y are a group starting at t_0 because they are ε -connected through black entities that are standing still. However, by the refined definition, the group of $\{x, y\}$ starts only at t_1 when a and h encounter each other.

We compute all maximal groups according to the original definition using the algorithm of Buchin et al. [2]. For a set of n entities each specified using τ time-stamped locations, this algorithm runs in $O(\tau n^3 + N)$ time, where N is the output size. We use their original implementation. Computing all maximal groups according to refined definition [10] takes $O(\tau^2 n^5 \log n)$ time. We implemented the algorithm ourselves.

3 Expectations

The two definitions for groups differ only in a subtle way. We observe that any group by the refined definition is a group by the original definition, in particular, any maximal group by the refined definition is a (not necessarily maximal) group by the original definition. This implies that for any maximal group by the refined definition, there exists a maximal group by the original definition that has at least these entities and at least this duration.

We can expect that in situations that are “easy” for detecting groups, the two definitions give similar results in terms of the number of maximal groups and the duration of these groups. When the situation gets more and more complex, the detection of groups also gets more difficult. The small difference in the definitions may lead to different results now, because the accidental linking of entities through ε -closeness via entities that are not in the group is more likely to happen, which is exactly where the definitions differ. So we



■ **Figure 2** Trajectories of people walking in the corridor from the pedestrian data provided by the Jülich Supercomputing Centre.

may see maximal groups in the original definition that do not exist in the refined definition. Furthermore, maximal groups may have a longer duration by accidental linking just before the group is ε -connected or just after it.

It is not directly clear, however, that the original definition will return more maximal groups. Besides the effect just sketched above, it can also be that a maximal group in the original definition is briefly spread too much but some other entity in the neighborhood provides the linking to keep on seeing it as one maximal group. This linking would not be realized in the refined definition, which may lead to two maximal groups due to the interruption. If this happens much, the refined definition might give more maximal groups.

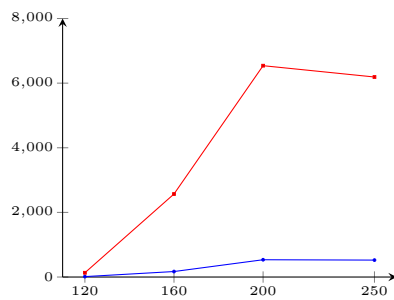
4 The Pedestrian Data

Our set of experiments uses pedestrian data collected by the Civil Security and Traffic division of the Jülich Supercomputing Centre [4] to study the dynamics of pedestrians. The data consists of trajectories extracted from video recordings of people walking in a synthetic environment. The particular datasets we use consist of two sets of people walking in opposite directions through a corridor that is 8 meters long and 3.6 meters wide [4]. The density inside the corridor is controlled by the width w , in centimeters, of the two entrances to the corridor: a larger width w means that more people can enter the corridor simultaneously. The considered widths w are taken from $\{120, 160, 200, 250\}$. Each experiment consists of 300 trajectories, each of approximately 300 vertices as well.

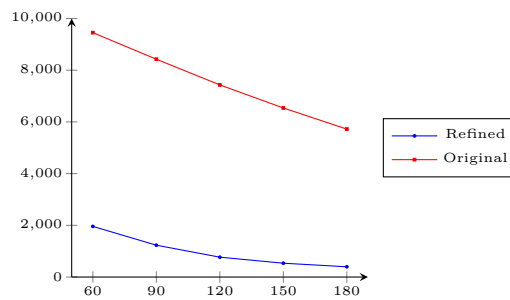
In our experiments we fix the inter-entity distance ε to 80 cm, and choose the minimum group size m from $\{3, 6, 9\}$. For the minimum required duration δ we consider values in the range $[60, 180]$. This corresponds to a minimum group duration roughly between four and twelve seconds. For comparison, the average time \bar{t} it takes a person to cross the corridor ranges from roughly twelve to twenty-three seconds.

The Number of Maximal Groups. We first consider the number of maximal groups as a function of w , and thus of the density of the environment. As Figure 3 highlights for the case $m = 6$ and $\delta = 150$, we see that up to $w = 200$, the number of reported maximal groups increase as a function of w . This applies for both the definitions of a group, although the number of maximal groups according to the original definition increases much faster than for the refined definition. For even bigger values of w , the number of maximal groups flattens off, or sometimes even decreases. These results are more apparent for larger values of δ .

The number of maximal groups reported by the refined definition is generally much smaller than the number of maximal groups reported by the original definition. This is also clearly visible in Figure 4, where we show the number of maximal groups, with $m = 6$, and $w = 200$, as a function of δ . The graphs for different settings of m and w are similar. Here, we also see that the number of maximal groups decreases as we increase the minimum required duration (which is to be expected).



■ **Figure 3** The number of maximal groups for $m = 6$ and $\delta = 150$ as a function of the width w of the corridor entrance, which influences density.



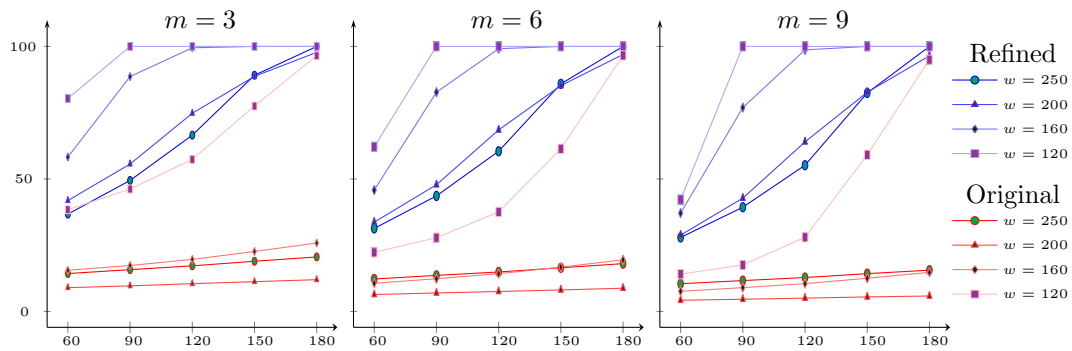
■ **Figure 4** The number of maximal groups for $m = 6$ and $w = 200$ as a function of δ . There are much fewer maximal groups according to the refined definition when compared with the original definition.

Measuring the Conformity of a Group. Since all entities (pedestrians) completely cross the corridor, we can classify each entity as type going “left to right” (type R), or “right to left” (type L). We can extend this notion to groups of entities by taking the type of the majority of its members (in case of ties we pick arbitrarily). We then define the *conformity* $c(G)$ of a group G as the percentage of its members that have the same type as the type of the group. Hence, the conformity of G is a value varying from 50, half of the members of G cross the corridor each way, to 100, all members of G go in the same direction. Intuitively, we expect that a set of people that act as a group (in the social sense) travel in the same direction, and thus we expect the conformity to be high in a good grouping definition.

We now measure the conformity of all maximal groups reported by our two definitions. Specifically, we consider the percentage of maximal groups that have conformity 100, that is, all group members travel in the same direction. We say that such a group is *uni-directional*. The results are in Figure 5. Consider the case where $m = 3$ and $w = 120$. For both definitions, we see that as the minimum required duration increases, so does the percentage of uni-directional maximal groups. However, the refined definition generally has a much higher percentage of uni-directional maximal groups. In particular, for a duration as short as 90 time units (about 5 seconds), all maximal groups are uni-directional. For the original definition this requires a minimum duration threshold of more than 180. These results are even more clearly visible as we increase the width of the corridor. For example, for $w = 160$, all maximal groups with a duration of at least $\delta = 120$ are uni-directional, whereas in the original definition less than 40% of the reported maximal groups are uni-directional, even if we increase the minimum required duration to 180. We expect that this is mostly due to the fact that the original definition reports many more maximal groups than the refined definition. We get similar results for larger minimum group size thresholds, that is, $m = 6$ and $m = 9$.

5 Conclusions

We examined two definitions for groups in trajectory data which both support continuous movement and varying shapes of groups. One definition was introduced as a refinement of the other, to obtain a more natural formalization of groups, but at the expense of a less efficient algorithm for their computation. Our comparison is based on a number of experiments where groups are computed by both definitions.



■ **Figure 5** The conformity of the maximal groups in the pedestrian data as a function of δ .

The most important finding is that the two definitions differ more and more as the density of the crowd increases. This implies that in dense situations it does matter which definition is taken, even though they seem very similar. A second observation is that the refined definition appears to be more natural, at least in some cases. The original definition reports many groups that contain entities that move in opposite directions, whereas the refined definition finds only a few of them. Moreover, such groups then often have a short duration. An other interesting observation is that the refined definition gives fewer groups. It is not clear whether this is an advantage or a disadvantage, since the nature of both definitions gives rise to groups that share entities at the same time.

References

- 1 Marc Benkert, Joachim Gudmundsson, Florian Hübner, and Thomas Wollé. Reporting flock patterns. *Comput. Geom.*, 41(3):111–125, 2008.
- 2 Kevin Buchin, Maike Buchin, Marc van Kreveld, Bettina Speckmann, and Frank Staals. Trajectory grouping structure. *Journal of Computational Geometry*, 6(1):75–98, 2015.
- 3 Joachim Gudmundsson, Patrick Laube, and Thomas Wollé. Computational movement analysis. In Wolfgang Kresse and David N. Danko, editors, *Handbook of Geographic Information*, pages 725–741. Springer, Berlin, 2012.
- 4 Stefan Holl. *Methoden für die Bemessung der Leistungsfähigkeit multidirektional genutzter Fußverkehrsanlagen*. Dr., Bergische Universität Wuppertal, Jülich, 2016. Bergische Universität Wuppertal, Diss., 2016.
- 5 Yan Huang, Cai Chen, and Pinliang Dong. *Modeling Herds and Their Evolvments from Trajectory Data*, pages 90–105. Springer, Berlin, 2008.
- 6 San-Yih Hwang, Ying-Han Liu, Jeng-Kuen Chiu, and Ee-Peng Lim. Mining mobile group patterns: A trajectory-based approach. In *Proc. 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, PAKDD’05, pages 713–718. Springer, 2005.
- 7 Hoyoung Jeung, Man Lung Yiu, Xiaofang Zhou, Christian S. Jensen, and Heng Tao Shen. Discovery of convoys in trajectory databases. *PVLDB*, 1(1):1068–1080, 2008.
- 8 Panos Kalnis, Nikos Mamoulis, and Spiridon Bakiras. *On Discovering Moving Clusters in Spatio-temporal Data*, pages 364–381. Springer, Berlin, 2005.
- 9 Zhenhui Li, Bolin Ding, Jiawei Han, and Roland Kays. Swarm: Mining relaxed temporal moving object clusters. *PVLDB*, 3(1):723–734, 2010.
- 10 Marc van Kreveld, Maarten Löffler, Frank Staals, and Lionov Wiratma. A Refined Definition for Groups of Moving Entities and its Computation. In *Proc. 27th International Symposium on Algorithms and Computation*, volume 64, pages 48:1–48:12. LIPIcs, 2016.
- 11 Yu Zheng and Xiaofang Zhou. *Computing with Spatial Trajectories*. Springer, 2011.

Extracting Geospatial Information from Social Media Data for Hazard Mitigation, Typhoon Hato as Case Study

Jibo Xie

Institute of Remote Sensing and Digital Earth, No.9 Dengzhuang South Rd, Haidian District, Beijing 100094, China
xiejb@radi.ac.cn

Tengfei Yang

Institute of Remote Sensing and Digital Earth; University of Chinese Academy of Sciences, No.9 Dengzhuang South Rd, Haidian District, Beijing 100094; Beijing 100049, China
yangtf@radi.ac.cn

Guoqing Li

Institute of Remote Sensing and Digital Earth, No.9 Dengzhuang South Rd, Haidian District, Beijing 100094, China
ligq@radi.ac.cn

Abstract

With social media widely used for interpersonal communication, it has served as one important channel for information creation and propagation especially during hazard events. Users of social media in hazard-affected area can capture and upload hazard information more timely by portable and internet-connected electric devices such as smart phones or tablet computers equipped with (Global Positioning System) GPS devices and cameras. The information from social media (e.g. Twitter, facebook, sina-weibo, WebChat, etc.) contains a lot of hazard related information including texts, pictures, and videos. Most important thing is that a fair proportion of these crowd-sourcing information is valuable for the geospatial analysis in Geographic information system (GIS) during the hazard mitigation process. The geospatial information (position of observer, hazard-affected region, status of damages, etc) can be acquired and extracted from social media data. And hazard related information could also be used as the GIS attributes. But social media data obtained from crowd-sourcing is quite complex and fragmented on format or semantics. In this paper, we introduced the method how to acquire and extract fine-grained hazard damage geospatial information. According to the need of hazard relief, we classified the extracted information into eleven hazard loss categories and we also analyzed the public's sentiment to the hazard. The 2017 typhoon "Hato" was selected as the case study to test the method introduced.

2012 ACM Subject Classification Human-centered computing → Social media, Information systems → Geographic information systems, Computing methodologies → Information extraction, Human-centered computing → Geographic visualization

Keywords and phrases Social media, hazard mitigation, GIS, information extraction, typhoon

Digital Object Identifier 10.4230/LIPICs.GIScience.2018.65

Category Short Paper

Funding The national key research and development program of China (2016YFE0122600).

Acknowledgements We want to thank Edward T.-H. Chu from National Yunlin University of Science and Technology in the collaboration project.



© Jibo Xie, Tengfei Yang, and Guoqing Li;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 65; pp. 65:1–65:6

Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Social media has been widely used in our daily information creation and propagation especially in the hazard scenario. Users of social media in hazard affected area can acquire real-time firsthand in-situ observation data and share these data by messages, short texts, pictures, or videos. And a fair proportion of these crowd-sourcing data includes geospatial information which is valuable for geospatial analysis of Geographic information system (GIS) during the hazard mitigation process. The geo-location information of social media data plays an important role in emergency detection and quick response [3]. The useful geospatial information including position, geospatial distribution, location clustering, and status of damages related with hazard, is hidden in the large number of social media data. Unlike conventional spatiotemporal data, social media data is dynamic, massive, unevenly distributed in space and time, noisy, incomplete, biased in terms of population, and represented in stream of unstructured media (e.g. texts and photos), which pose fundamental challenges for representation and computation to conventional spatio-temporal analysis [1]. Many researchers in GIS study area have noticed the importance of social media as an important source of geospatial information. In the past few years, geospatial information created by volunteers and facilitated by social networks has become a promising data source in time-critical situations [5]. And the concept of volunteering of geographic information (VGI) [4] has been introduced. While the quantity and real-time availability of VGI make it a valuable resource for disaster management applications, data volume, as well as its unstructured, heterogeneous nature, make the effective use of VGI challenging [2]. user-generated data can provide unique and highly useful information in several contexts (e.g. brand communication, market research, political communication as well as in extreme events) [6]. The social media GIS enables disaster information provided by local residents and governments to be mashed up on a GIS base map, and for the information to be classified and provided to support the utilization of the information by local residents [8]. In our study, we introduced the method to extract geospatial information and use these information to get the hazard loss categories and map the hazard-affected area. The Typhoon, a yearly happened hazard events in northwest Pacific, was selected as case study.

2 Methods

The key step of extracting hazard related geospatial information from social media data is how to understand the meanings of messages and texts. And the Natural Language Processing (NLP) is a common method for social media information extraction. In our study, we proposed a Social Media based Hazard information Recognition and Classification (SHRC) model for hazard related geospatial information extraction and analysis based on the NLP method. The workflow of the model (As shown in Fig. 1). The key steps of the SHRC model are as followed.

1. Event-driven hazard information acquisition from social media: Social media platform usually provides the interface or API for developers to retrieve and get social media data by using time-span, location and event related key words.
2. Data cleaning and store: Many messages from social media are repeated or not related, so we need to clean and filter the redundancy. After that the data are stored in the database for further analysis.
3. Definition of hazard loss categories: To evaluate the hazard loss, we proposed a hazard loss classification method of eleven categories including loss of life, interruption of water supply, building damage, business influence, forestry loss, traffic congestion, vehicle



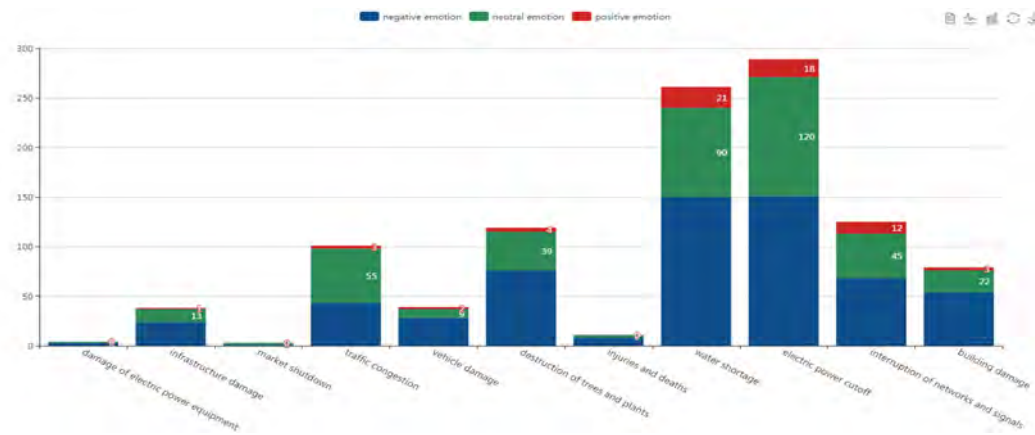
■ **Figure 1** Flowchar of extracting hazard geospatial information from social media data.

damage, power supply outage, electricity equipment broken, Communication Interrupt, infrastructure damage.

4. Creation of classification knowledge base based on feature words and lexicon: The first step is to extract some feature words from the sample micro- blog text of different disaster loss categories based on Chinese grammar rules and constructed the pairs of feature words collocation. The word vector model and existing lexicon is used to supplement and expand these pairs of feature words collocation. And the external natural language corpus is used to optimize the semantic collocation relationship between feature words.
5. Hazard information interpretation and extraction: The topics of social media messages are usually random and we use the hazard classification knowledge base in the step 4) for different types of hazard damage information extraction. A Chinese language processing and information retrieval toolkit, NLPiR (<http://ictclas.nlpir.org/downloads>), is deployed for word segmentation and part-of-speech tagging(POS). Then corresponding lexicon is used to match feature words for disaster loss information classification and sentiment analysis based on the knowledge base.
6. Sentiment analysis: The model uses sentiment words for sentiment analysis. The basic sentiment words from the text base on Chinese sentiment word table from “HowNet” (http://www.keenage.com/html/c_index.html). And the model extends the basic sentiment words by using the feature words from social media. There are three kinds of emotions, positive, neutral, and negative.
7. Spatio-temporal visualization: The hazard information from social media is geo-located by GPS position, address match by user’s position, and Identification of place names from text. Then we can use the geospatial information and hazard loss attributions for visualization and mapping of the hazard-affected area.
8. Evaluation and validation: Three parameters, precision, recall and F-Measure (F1), serve as the evaluation indexes to evaluate the experimental results.



■ **Figure 2** Typhoon “Hato” landed on the coast of Zhuhai city at 12:50, 23rd August, 2017. Map from <http://typhoon.zjwater.gov.cn>.



■ **Figure 3** Sentiment analysis of hazard damage information extracted from social media. The blue, green, and red colors refer to negative, neutral, and positive emotion, respectively.

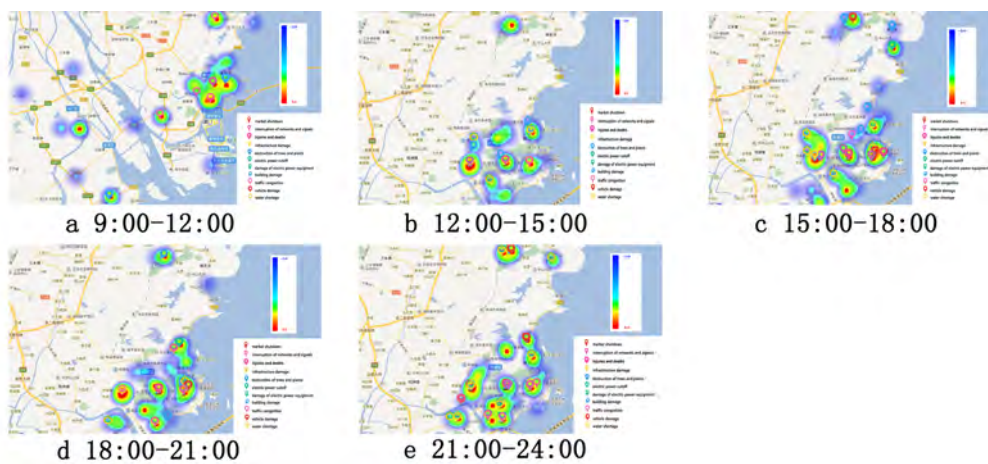
3 Case study

We used the dataset of 2017 typhoon events (about 20,000 records) [7] to train the model proposed in the paper. And typhoon “Hato” event landed on the coast of Zhuhai city at 12:50 23rd August, 2017 was selected as the case study to test the effectiveness of the model. The moving track of this typhoon is shown in Fig. 2. We selected 1600 records of the hazard related information from “sina-weibo” (<https://weibo.com/>) after cleaning and filtering redundant and irrelevant information with time span from 0:00 to 23:00 of the typhoon landfall day.

The statistics of the hazard information extraction and classification from “sina-weibo” are shown in Fig. 3. We can see that the numbers of power outage and interruption of water supply were the biggest, 289 and 261 respectively. According to the statistics, we can conclude that outage of power and water supply were the most affected hazard damages or people paid much more attention on that during the typhoon hazard. And we did further analysis on sentiment and classified the human emotion in three categories, negative, positive and neutral sentiment. Fig.3 gives us a direct illustration of people’s sentiment to different categories of hazard loss types. The blue, green, red colors refer to negative, neutral, and positive emotion respectively. For example, there is a micro-blog message saying that “The typhoon is really terrible”, we can identify the Chinese sentiment word “terrible” to put this short sentence in category of negative sentiment. But, there is a micro-blog message saying



■ **Figure 4** Distribution map of different types of hazard damage information from social media.



■ **Figure 5** Social media based spatio-temporal sentiment analysis.

that “The typhoon has great destructive power, it has no electricity until now, but thanks to the power workers who are working hard to repair it. Give them a thumbs up”. This text contains “Power supply outage”, but the emotion is positive, which shows that people were satisfied with the disaster reduction response. We can get people’s reaction to the typhoon event or know how severe of the hazard influence on the people’ life there. And we also got the geospatial information of different damage types and visualized in the map as shown in Fig. 4. This map shows the hazard damage distribution of hazard affected area. And this map can be a useful supplemental geospatial data to the official hazard mitigation. Most important thing is that, the geospatial data can be obtained in a near real-time manner which is just the insufficiency of the common geospatial data acquisition method. As Fig. 5 shown, we did a spatio-temporal sentiment analysis of typhoon “hato” during its landfall on coast of Zhuhai City. Before the landfall of the typhoon as shown in Fig. 5a, the major hazard influence was traffic congestion. And people of the hazard-affected area were in a hurry to return home. With the typhoon landfalled and moved to the northwest, more categories of hazard damages emerged along with the landfall route as illustrated in Fig. 5b,c,d,e. And outage of power and water supply was the prominent influence types of the hazard by analysis of the number and sentiment of social media records. And as the typhoon passed by, the negative emotion decreased and positive emotion increased.

We evaluated the experimental results with precision, recall and F-1. The comprehensive evaluation index of different disaster loss categories was greater than 0.74. And the comprehensive evaluation index of different sentiment categories was greater than 0.83.

4 Methods

Social media data contains a lot of valuable near-real time geospatial information during the hazard events. This paper introduced the method to extract hazard related geospatial information for evaluation of hazard loss. And we proposed a Social Media based Hazard information Recognition and Classification model for hazard related geospatial information extraction and analysis based on the nature language processing and sentiment analysis. Typhoon “Hato” (landed on the coast of Zhuhai city at 12:50, 23rd August, 2017) was selected as the case study. Firstly, social media data of hazard event were collected and cleaned for further hazard information extraction. Nature language processing and semantic interpretation was done to understand the content of the text of social media data. A hazard damaged evaluation standard with eleven categories was proposed for the information classification. And these hazard loss categories were geo-located and mapped to show the distribution of hazard loss. Also sentiment analysis was done to extracted people’s reaction to the Typhoon hazard. The geospatial time-serial map of sentiment analysis was generated. In our recent research, We are developing a near real-time social media based hazard information acquisition and analysis system. And we will use more hazard events to train the model before it can be used in practical hazard mitigation.

References


- 1 Valentina Cerutti, Georg Fuchs, Gennady Andrienko, NataliaAndrienko, and Frank Ostermann. Identification of disaster-affected areas using exploratory visual analysis of georeferenced tweets: application to a flood event. In *16th Annual Symposium on Foundations of Computer Science, Berkeley, California, USA, October 13-15, 1975*, pages 1–5, 2016.
- 2 P. Thakuriah et al. *Using Social Media and Satellite Data for Damage Assessment in Urban Areas During Emergencies*. Springer Geography, 2016.
- 3 Xu et al. Participatory sensing-based semantic and spatial analysis of urban emergency events using mobile social media. *EURASIP Journal on Wireless Communications and Networking*, 2016(4):1–9, 2016. doi:10.1186/s13638-016-0553-0.
- 4 Michael F. Goodchild. *Citizens as sensors: web 2.0 and the volunteering of geographic information*. GeoFocus, 2007.
- 5 Linna Li and Michael F. Goodchild. The Role of Social Networks in Emergency Management: A Research Agenda. *International Journal of Information Systems for Crisis Response and Management*, 2(4):49–59, 2010. doi:DOI:10.4018/jiscrm.2010100104.
- 6 Milad Mirbabaie, Stefan Stieglitz, and Stephan Volkeri. Volunteered geographic information and its implications for disaster management. In *HICSS '16 Proceedings of the 2016 49th Hawaii International Conference on System Sciences (HICSS), Washington, DC, USA ,January 05 - 08, 2016*, pages 207–216, 2016. doi:10.1109/HICSS.2016.33.
- 7 Jibo Xie Tengfei Yang and Guoqing Li. A social media-based dataset of typhoon disasters. *China Scientific Data*, 2018(3), 2018. doi:10.11922/scdata.2017.0014.en.
- 8 Kayoko YAMAMOTO and Shun FUJITA. Development of Social Media GIS to Support Information Utilization from Normal Times to Disaster Outbreak Times. *International Journal of Advanced Computer Science and Applications*, 6(9):1–14, 2015.

Propagation of Uncertainty for Volunteered Geographic Information in Machine Learning

Jin Xing

Centre for Research in Geomatics, Laval University, Quebec City, Canada

jin.xing.1@ulaval.ca

 <https://orcid.org/0000-0001-5693-3414>

Renee E. Sieber

Department of Geography, McGill University, Montreal, Canada

renee.sieber@mcgill.ca

Abstract

Although crowdsourcing drives much of the interest in Machine Learning (ML) in Geographic Information Science (GIScience), the impact of uncertainty of Volunteered Geographic Information (VGI) on ML has been insufficiently studied. This significantly hampers the application of ML in GIScience. In this paper, we briefly delineate five common stages of employing VGI in ML processes, introduce some examples, and then describe propagation of uncertainty of VGI.

2012 ACM Subject Classification Information systems → Uncertainty

Keywords and phrases Uncertainty, Machine Learning, Volunteered Geographic Information, Uncertainty Propagation

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.66

Category Short Paper

1 Background of VGI in Machine Learning

Machine Learning (ML) represents a set of methods that automatically learn from “experience” or training data with respect to given tasks. The learning can be implemented via a large body of models and algorithms, such as heuristic rules [32], decision trees [27], and cellular automata [31]. In Geographic Information Science (GIScience), ML has attracted considerable interest due to its wide applications in place recognition [34], ecology models [25], remote sensing image classification [33], transportation pattern discovery [22], and gazetteer analysis [9]. The rapid grow of ML has intensified due to the increasing ‘bigness’ of geospatial data, which describes the exaflood of geographic information at unprecedented volume, velocity, and variety, as well as challenges to veracity.

Among the diverse sources of big data, Volunteered Geographic Information (VGI) is considered a main provider of input data/services [12]. For example, OpenStreetMap OSM, in which individuals have crowdsourced editable web mapping services and content, has become a powerful platform for building, training, and evaluating ML algorithms and models in GIScience [15]. VGI describes the process of obtaining geographic data or services (e.g., rating accuracy of feature labels) from large groups of users in an open call that is self-organizing via the Internet [10]. Uncertainty is innate within VGI, which means data is noisy, containing redundancies, irrelevant content, errors and biases contributed by users, who are often non-experts [26]. VGI also is disorderly, in which data may be unstructured, incorrectly ordered, mis-formatted (e.g., lacking a header), and possibly poorly geo-registered. Finally, users may be unreliable in providing consistent input and inputting within the appropriate



© Jin Xing and Renee E. Sieber;

licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 66; pp. 66:1–66:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

■ **Table 1** Uncertainty Issues in Applying VGI for ML

ML Process	Uncertainty Type		Examples in VGI
Data Collection, Annotation, and Cleaning	Data Uncertainty		Inaccurate geolocation; spatial unevenness in data contributions; redundancies; gender, culture, and race bias in training data
Data Distribution	Operation	Uncertainty	Boundary Vagueness (e.g., artificial boundaries introduced by data splitting); aggregation errors (e.g., heaping error in determining the existence of a traffic jam, binning of VGI point data)
Feature/Topic Detection	Representation	Uncertainty	Interpreting location from place (from a well-defined to a poorly defined object)
Model/Algorithm Selection and Training	Decision Uncertainty		Simpler/alternate models than ML may be better like linear regression
Evaluation and Tuning	Service Uncertainty		Biased classification; Inconsistency in grading

time periods. Noisy, disordered, and unreliable data and service can significantly lower the value of VGI in ML.

Previous work in VGI's uncertainty largely concentrates on the data quality. Researchers focused, for example, on uncertainty regarding the non-expert (e.g., skill levels and motivation), the thematic diversity of input (scattered focus relative to analysis needs), and the spatial unevenness of contributions (e.g., popularity of places relative to others) [11]. In ML, VGI is viewed primarily for its ability to provide data for ML, either as training data or general input data. It also has been employed for result evaluation and tuning of ML [18]. A worrying trend in GIScience inquiry into ML is its treatment as a big black box, where issues of data uncertainty are treated as I/O problems. We break down the black box of ML into a collection of workflow processes to briefly identify uncertainty from VGI that can occur within the ML as well as in its parameterization and refinement.

Other taxonomies tend to focus on classifying ML methods (e.g., supervised, unsupervised, and reinforcement learning) and application areas (e.g., computer vision, natural language processing, and speech recognition)[16]. The importance of uncertainty and its propagation have not been highlighted. We view the interaction between VGI and ML as five stages throughout the processing of VGI: data collection and cleaning, data distribution, feature/topic detection, model/algorithm selection and training, and evaluation and tuning.

2 A General Framework for Integrating Geospatial Crowdsourcing and ML

Our framework (Table 1) follows the standard ML workflow (data collection and cleaning, splitting of training from testing data, model training, evaluation, parameter tuning) [28] and adds components from big data handling [21] and ML computation [4] for de-/re-composition. Since the five stages may occur iteratively (e.g., the evaluation result could be fed back to the training process to improve accuracy), uncertainty also can propagate if we fail to attend to the origin of the uncertainty.

2.1 Data Collection and Cleaning

The primary utility of VGI in ML is for training and, more generally, input data. Training refers to data used by ML to calculate its parameters/weights so that input data generates expected outputs. Geospatial content is available across a wide range of VGI. It can be raster (landscape photographs) and vector (social checkins, binned aggregations of points); structured (Twitter metadata) and unstructured (Twitter text), explicit (x,y's, placenames in hashtags) and implicit (colloquial names for neighborhood), absolute (latitude/longitude) and relative (concepts of home), passive (geo-fencing) and active (Amazon Mechanical Turk-AMT). It can be static or dynamic (harvesting of Flickr geotags at point in time or movement data), compensated or voluntary (AMT or VGI) [19]. Considerable research has been conducted to assess uncertainty with various VGI (cf., [14]).

Like other crowdsourced content, VGI data contains considerable error, vagueness, and ambiguity, and is vulnerable to malicious contributions (e.g., via GPS spoofing). As suggested above, this is the richest area of current research so this section is admittedly brief. Most research on the negative impact of ML focuses on the issue of algorithmic bias due to input data [26]. Location often serves as a proxy for race so one needs to debias on the basis of primary variable as well as data which functions as its surrogate [1]. Often debiasing requires human intervention (cf., gendered word2vec example in [2]) so this stage also can utilize crowdsourcing. Geographic unevenness in data contributions can further distort ML output, for example the low OSM participation in Africa or the differential accuracy of OSM in urban areas versus rural regions [29]. Privacy protections, like the EU's General Data Protection Regulation, will increase distortions in VGI as whole swaths of data are removed or masked [6]. Lastly, much of VGI is streamed, which requires new sampling techniques (e.g., reservoir sampling) to normalize temporal spikes or redundancies.

2.2 Data Distribution

The attraction of VGI to ML is both in its source (geosocial media) and its potential as big data. The latter likely requires de-/re-composition to distribute the computing. Data distribution may suffer from disorder in VGI because geographic data has its own internal topology and geometry that can be destroyed by arbitrary decomposition or splitting. For example, rectangular decomposition can distort the boundary of geographic objects and increase output uncertainty [5]. Most VGI is point-based and may need to be binned. A more sophisticated feature type, a polygon like a hexagon, does not easily alleviate the problem and any aggregation is subject to modifiable areal unit problems [24] that can alter ML output.

ML can be employed to reduce uncertainty in data distribution. Felzenszwalb et al. [7] employed latent support vector machine to decompose the original raster data into multiple object-based rectangles to lower boundary distortions. Temporal disorder in VGI, such as burstiness of reporting of natural disasters, could be addressed by decomposition with parallel processing.

2.3 Feature/Topic Detection

ML is designed in large part to recognize patterns, generate rules, approximate functions, and classify data sets. An important use of VGI in ML can be for feature or topic detection (e.g., forest, alternate route to avoid traffic jam). We lack explicit control over the feature representation in VGI. Users may not provide feature identification as planned or neural networks may fail to extract useful features from noisy VGI. For example, uncertainty in

placename makes it difficult to infer locations; “downtown nearby” could be interpreted as multiple locations [8]. Although iterative feature/object detection in ML can reduce uncertainty, there is no easy way to clean data to better disambiguate place to a location and location to a place. This resembles the challenge of NLP regarding semantic modeling to disambiguate slang (e.g., “bad”, “hot”, “sick”) in ML. Aggregation (pattern detection) is a likely outcome of ML that is based on VGI and therefore is subject to Sorites paradox and modifiable areal unit problems here as well (e.g., how many cars constitute a jam; how many trees constitute a forest).

The temptation for users new to ML is to treat it as a blackbox, an algorithm amongst many in a software library. Treating ML as a black box means that ML cannot necessarily accommodate the geography of VGI. For example, max pooling, which is a widely used method to pass features from one layer of neural network to another, is considered problematic in convolutional neural network by Sabour et al. [30] because max pooling lacks topology. In another example, a word embedding algorithm may produce very different vectors to represent “pub” and “bar” due to the surrounding content, which may then require multiple detection iterations.

2.4 Model/Algorithm Selection and Training

Which ML model or algorithm achieves the highest accuracy with a given input dataset and features? What is the best way to calculate the weights or parameters of the ML model/algorithm? Should we rely on a single ML model/algorithm or combine several ones together? These questions are difficult in ML and there are no clear answers. VGI can potentially assist this selection process with existing knowledge about model/algorithm selection and training strategies (think a wiki of appropriate ML) [23]. However, knowledge contributed via VGI may be unreliable because of a “follow the crowd” mentality with little investigation into alternate approaches [17]. Deep neural network is increasingly popular in ML research but a linear regression may be more appropriate, considering the quality of the data at hand and the ease of an ML implementation.

2.5 Evaluation and Tuning

Performance of ML algorithms needs to be evaluated with datasets different from the training process. VGI plays a pivotal role in collecting evaluation datasets and crowdsourcing can play a role in the evaluation process. To avoid overfitting (i.e., model is too closely fitted to the training data), ML scientists usually employ cross-validation, which can reduce the influence of uncertainty from VGI training data. Evaluation can be conducted with crowdsourcing services, such as the translation validation within the Google Translate Community [20] or Captcha [3]. Here, issues similar to data collection re-emerge, with potential biases introduced by the evaluators, who may be drawn from a particular gender, race, class, or skill level. These issues resemble the social approach to assessing spatial data accuracy in [13], in which the focus shifts from the uncertainty of the contribution to that of the contributor. One may wish to implement ranking or rating systems to improve confidence in the validators.

3 Propagation of Uncertainty in ML and Conclusion

In this paper, we propose a general framework to explore VGI uncertainty in ML. This includes the concrete importance of VGI for training data as well as the use of crowdsourcing for model/algorithm selection and performance evaluation in ML.

Uncertainty also can propagate across the ML workflow. Uncertainty in data collection can make data distribution more difficult because we do not know the appropriate aggregation size or scale. Without adequate cleaning, noisy data can generate messy features or false positives that will invalidate the chosen ML models and algorithms. Crowdsourcers bring their own bias to the evaluation of ML, which can influence the training of ML for parameter tuning. Disagreements during the cross validations may generate inconsistency in iterations of ML and force us to re-run the process. Where possible, it is critical to identify uncertainty at each stage to minimize the propagation of uncertainty. However, the cost (e.g., human intervention) of reducing the uncertainty in the early stages of ML (e.g., data collection and cleaning) is generally less than later stages (e.g., evaluation and tuning), so it is useful for us to consider at which stages it is appropriate to insert geographic crowdsourcing and crowdsourcers.

References

- 1 Julia Angwin. Make algorithms accountable. *The New York Times*, 1, 2016.
- 2 Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357, 2016.
- 3 Kumar Chellapilla and Patrice Y Simard. Using machine learning to break visual human interaction proofs (hips). In *Advances in neural information processing systems*, pages 265–272, 2005.
- 4 Cheng-Tao Chu, Sang K Kim, Yi-An Lin, YuanYuan Yu, Gary Bradski, Kunle Olukotun, and Andrew Y Ng. Map-reduce for machine learning on multicore. In *Advances in neural information processing systems*, pages 281–288, 2007.
- 5 Joao Porto De Albuquerque, Benjamin Herfort, Alexander Brenning, and Alexander Zipf. A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *International Journal of Geographical Information Science*, 29(4):667–689, 2015.
- 6 Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003, 2015.
- 7 Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.
- 8 Andrew J Flanagin and Miriam J Metzger. The credibility of volunteered geographic information. *GeoJournal*, 72(3-4):137–148, 2008.
- 9 Noah W Garfinkle, Lucas Selig, Timothy K Perkins, and George W Calfas. Geoparsing text for characterizing urban operational environments through machine learning techniques. In *Geospatial Informatics, Fusion, and Motion Video Analytics VII*, volume 10199, page 101990C. International Society for Optics and Photonics, 2017.
- 10 Michael F Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221, 2007.
- 11 Michael F Goodchild. Commentary: whither vgi? *GeoJournal*, 72(3-4):239–244, 2008.
- 12 Michael F Goodchild and J Alan Glennon. Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*, 3(3):231–241, 2010.
- 13 Michael F Goodchild and Linna Li. Assuring the quality of volunteered geographic information. *Spatial statistics*, 1:110–120, 2012.
- 14 Joel Grira, Yvan Bédard, and Stéphane Roche. Spatial data uncertainty in the vgi world: Going from consumer to producer. *Geomatica*, 64(1):61–72, 2010.

- 15 Mordechai Haklay and Patrick Weber. Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18, 2008.
- 16 Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.
- 17 Ece Kamar, Severin Hacker, and Eric Horvitz. Combining human and machine intelligence in large-scale crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 467–474. International Foundation for Autonomous Agents and Multiagent Systems, 2012.
- 18 M Kanevski, A Pozdnukhov, and V Timonin. Machine learning algorithms for geospatial data. applications and software tools. In *Proceedings of the 4th International Congress on Environmental Modelling and Software*, 2008.
- 19 Leyla Kazemi and Cyrus Shahabi. Geocrowd: enabling query answering with spatial crowdsourcing. In *Proceedings of the 20th international conference on advances in geographic information systems*, pages 189–198. ACM, 2012.
- 20 Somesh Kumar. *Methods for community participation: a complete guide for practitioners*. New Delhi (India) Vistaar Pub., 2002.
- 21 Jae-Gil Lee and Minseo Kang. Geospatial big data: challenges and opportunities. *Big Data Research*, 2(2):74–81, 2015.
- 22 Jin Liu, Xiao Yu, Zheng Xu, Kim-Kwang Raymond Choo, Liang Hong, and Xiaohui Cui. A cloud-based taxi trace mining framework for smart city. *Software: Practice and Experience*, 47(8):1081–1094, 2017.
- 23 Gary Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.
- 24 Amelia McNamara and Aran Lunzer. Exploring the effects of spatial aggregation, 2016.
- 25 Julian D Olden, Joshua J Lawler, and N LeRoy Poff. Machine learning methods without tears: a primer for ecologists. *The Quarterly review of biology*, 83(2):171–193, 2008.
- 26 Cathy O’Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2017.
- 27 Georgios Paliouras, Vangelis Karkaletsis, Georgios Petasis, and Constantine D Spyropoulos. Learning decision trees for named-entity recognition and classification. In *ECAI Workshop on Machine Learning for Information Extraction*, 2000.
- 28 Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.
- 29 Chris Perkins. Plotting practices and politics:(im) mutable narratives in openstreetmap. *Transactions of the Institute of British Geographers*, 39(2):304–317, 2014.
- 30 Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, pages 3859–3869, 2017.
- 31 Hossein Shafizadeh-Moghadam, Ali Asghari, Mohammad Taleai, Marco Helbich, and Amin Tayyebi. Sensitivity analysis and accuracy assessment of the land transformation model using cellular automata. *GIScience & Remote Sensing*, 54(5):639–656, 2017.
- 32 Anna L Swan, Dov J Stekel, Charlie Hodgman, David Allaway, Mohammed H Alqahtani, Ali Mobasher, and Jaume Bacardit. A machine learning heuristic to identify biologically relevant and minimal biomarker panels from omics data. *BMC genomics*, 16(1):S2, 2015.
- 33 Liangpei Zhang, Lefei Zhang, and Bo Du. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 4(2):22–40, 2016.
- 34 Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.

Satellite Image Spoofing: Creating Remote Sensing Dataset with Generative Adversarial Networks

Chunxue Xu

College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Oregon, USA
xuch@oregonstate.edu

Bo Zhao

College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Oregon, USA
zhao2@oregonstate.edu

Abstract

The rise of Artificial Intelligence (AI) has brought up both opportunities and challenges for today's evolving GIScience. Its ability in image classification, object detection and feature extraction has been frequently praised. However, it may also apply for falsifying geospatial data. To demonstrate the thrilling power of AI, this research explored the potentials of deep learning algorithms in capturing geographic features and creating fake satellite images according to the learned 'sense'. Specifically, Generative Adversarial Networks (GANs) is used to capture geographic features of a certain place from a group of web maps and satellite images, and transfer the features to another place. Corvallis is selected as the study area, and fake datasets with 'learned' style from three big cities (i.e. New York City, Seattle and Beijing) are generated through CycleGAN. The empirical results show that GANs can 'remember' a certain 'sense of place' and further apply that 'sense' to another place. With this paper, we would like to raise both public and GIScientists' awareness in the potential occurrence of fake satellite images, and its impacts on various geospatial applications, such as environmental monitoring, urban planning, and land use development.

2012 ACM Subject Classification Human-centered computing → Geographic visualization

Keywords and phrases Deep Learning and AI, GANs, Fake Satellite Image, Geographic Feature

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.67

Category Short Paper

1 Introduction

Deep learning and Artificial intelligence (AI) techniques are attracting more and more attentions from geographers and spatial scientists. Big data with geospatial information like satellite images and crowdsourcing data are perfect input for computer-based learning algorithms, especially considering the huge success of machine learning and deep learning methods in computer vision problems, such as image classification, object detection and feature extraction[1]. Deep neural networks are developed so powerful that they are applied to enable geospatial system to capture underlying features and patterns at a near-human perception level[6]. For example, recent studies indicate that convolutional neural networks (CNNs) are highly effective in perceiving features in large-scale image recognition and semantic segmentation[5, 7]. Deep learning algorithm can facilitate the research in the domain of geoscience and remote sensing, and it is encouraged to be used with expertise from



© Chun X. Xu and Bo Zhao;

licensed under Creative Commons License CC-BY

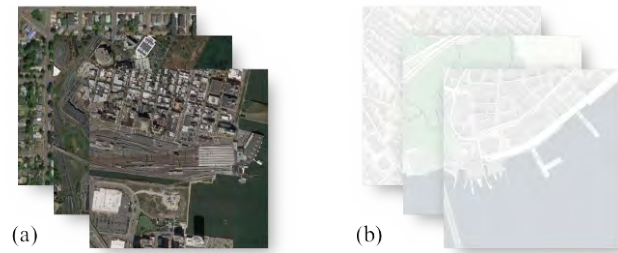
10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 67; pp. 67:1–67:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



■ **Figure 1** Input Datasets:(a) Google satellite images; (b) CartoDB basemap.

geospatial and remote-sensing scientists as an implicit general model to tackle large-scale and unprecedented challenges like climate change and urbanization[8].

But can AI be trusted to take care of our location information? How can we deal with fake information if deep learning and AI act as a ‘bad’ agency? ‘Fake’ is a big concern because of the popularity of fake news and post-truth politics. In geography, the falsification and spoofing of geospatial data have become a heated topics[9]. And ‘fake’ has also been a highly popular word in AI since 2014, when Generative Adversarial Networks (GANs) were introduced by Ian Goodfellow as a kind of artificial intelligence algorithm with huge potential to mimic various data distribution[2]. GANs have been used in hyperspectral image classification as a semi-supervised learning algorithm[3]. It provides us insight into how the machine can ‘remember’ what it saw in the past, and then generate fake data in any fields like image, speech or music. Instead of being used as a potential tool to deal with pragmatic mapping issue in cartography, will it be meaningful and operable to create fake satellite images with a certain kind of patterns or features?

To answer this question, this study aims to explore the potentials of GANs in capturing geographic features and whether the machine can generate a ‘sense of place’ like humans. Previous studies have indicated that it is possible to generate satellite images from street map through a mapping from image to image (or from pixel to pixel)[4, 10]. In this study, Corvallis is used as the place of study, and fake satellite images with a certain style will be generated through GANs to discuss this problem.

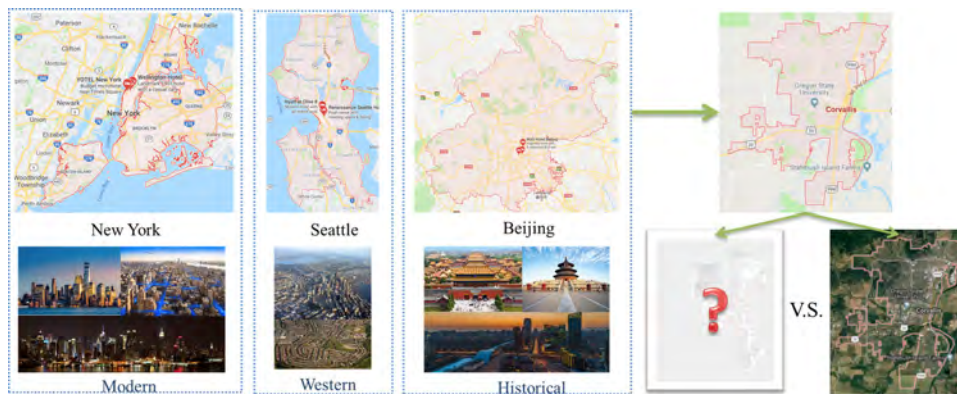
2 Data and method

2.1 Data

Satellite images from Google Earth and positron (no label) basemap from CartoDB are used as input datasets in this study. Positron basemaps from CartoDB are developed based on data from OpenStreetMap. The basemap is designed with latest data and has limited color schemes, which gives users freedom to customize according to their visualization use. Satellite images and maps are collected as multi-level raster tiles through a script based on Google maps API and Qtile in QGIS. The image tile is 512*512.

2.2 Method

Cycle-Consistent Adversarial Networks (CycleGAN)[10] were implemented in this study to train the model from training set and generate fake satellite images. CycleGAN is a kind of image-to-image translation algorithm that can work without paired examples



■ **Figure 2** Satellite images and basemaps of three cities (i.e. New York City, Seattle and Beijing) were used as input dataset for CycleGAN to extract city styles. Corvallis in Oregon is used as a sample for fake satellite images generation.

of transformation from source to target domain. Compared to previous image-to-image translation algorithm Pix2Pix, CycleGAN can learn the transformation without one to one mapping. Adversarial losses and cycle consistency losses are applied to mapping functions as follows[10]:

$$\mathcal{L}_{cyc}(G, F) = \mathbb{E}_{x \sim p_{data}(x)} [\|F(G(x)) - x\|_1] + \mathbb{E}_{y \sim p_{data}(y)} [\|G(F(y)) - y\|_1]. \quad (1)$$

$$\mathcal{L}_{GAN} = \mathbb{E}_{y \sim p_{data}(y)} [\log D_Y(y)] + \mathbb{E}_{x \sim p_{data}(x)} [\log(1 - D_Y(G(x)))]. \quad (2)$$

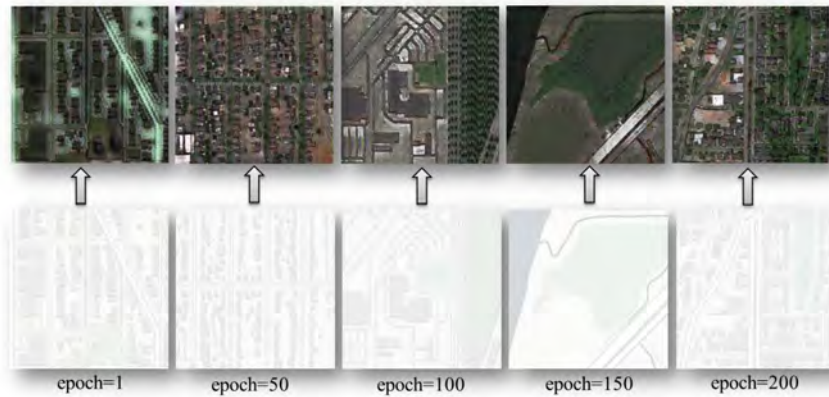
In this study, CycleGAN was used to reconstruct the satellite imagery from its basemap tiles for three typical cities in the world: New York City, Seattle and Beijing. It is assumed in this study that these three cities have different urban styles. As the center of the New York metropolitan area, New York City is the most populous urban agglomerations with a modern style, while buildings in Seattle have relatively low heights even in the urban center. Beijing is a hybrid of historical and modern style, with most urban area has been covered with tall buildings and new architectures. The geographic features are shown on the satellite images with different textures, such as shade, color scheme or spectral signatures, which are significant information for image interpretation and supervised classification. Despite of the spatial heterogeneity in internal urban area, these three cities in general give people different sense of place, which is a little bit obscure and elusive. Although hard to describe and quantify, we can still differentiate cities with their features from satellite images.

To examine whether it is possible for CycleGAN to extract city styles and transfer the potentially learned style to another place, Corvallis, where Oregon State University locates, was then used as a sample for generating satellite images. For the training dataset, there are 1066, 1196 and 1122 pairs of images for New York City, Seattle and Beijing respectively, and there are 360 map tiles covering Corvallis.

3 Result

3.1 Training process

Taking the training process with input data from Seattle as an example, CycleGAN training process was set with 200 epochs. Figure 3 shows some results during training process. After the first epoch, the training model can learn and generate some geometric features. More



■ **Figure 3** Some results during training process using training dataset of Seattle.

details can be captured and created as the number of epoch increases. As the basemap has limited color schemes, CycleGAN seems very sensitive in capturing green land.

3.2 Fake image generated for a single tile

It is difficult to quantify geographic features with a certain character or pattern, especially taking spatial variability and heterogeneity into account. Landscape exhibits various patterns and processes in different scales. However, it seems that CycleGAN is able to extract some general features from spatial distribution of city structures.

Take a specific location in northwest of Corvallis as an illustration. In general, three models are performing well in generating green land in the park area. But the details differ.

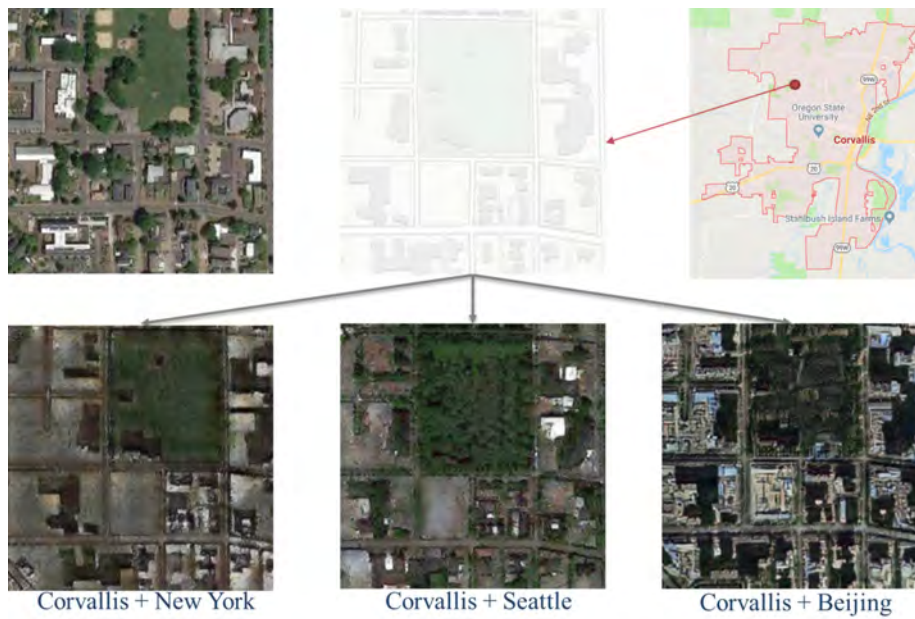
Although the fake satellite image with New York's style failed to generate details in some community block, some structures like tall building are created in the left corner of the image. And the one with Seattle's style has more detailed information, including low houses and buildings created around the park. What's more, the fake image with Beijing's style has most detailed information of tall buildings and shadow around them.

3.3 Fake images generated in a large scale

To examine the stability of the training model, fake satellite images are generated in a relatively large-scale area north to Oregon State University. Nine tiles covering this area are put together using mosaic method.

As New York model was not well-trained and failed to generate detailed information, fake satellite images with Seattle and Beijing are generated to examine the stability of the model in a large scale. Due to the blank in the northeast area of the basemap, less details are generated there compared to other areas in the fake satellite images with Seattle's style. Generally speaking, the images created from the model have similar pattern and style with Corvallis in terms of geographic characteristics.

But we can easily differentiate the fake images with transferred style from Beijing with that of Seattle. As is shown in figure 5, almost all building created with Beijing's style are tall with shadows. The fake city is really crowded considering the small block. As the real satellite images from Beijing have various temporal resolutions, the shadows looks unreasonable as the sunshine looks from various directions. However, the style of the imaginary city looks pretty unify across the whole area. In terms of the mosaic, no obvious edge could be observed



■ **Figure 4** Fake satellite image with transferred style from different cities for a single tile.



■ **Figure 5** (a) CartoDB basemap of large-scale study area in Corvallis; (b) Fake satellite image with transferred style from Seattle; (c) Fake satellite image with transferred style from Beijing.

from the image edges, which suggests that the model performs well and generates stable results for this area in Corvallis.

4 Conclusion and discussion

CycleGAN and Pix2Pix networks have the ability to generate non-existent but realistic images. This study explored whether it can capture complicated geographic features and patterns from a large dataset of a certain city, and transfer the style to another place in a large scale. The results show that CycleGAN can learn general styles from input dataset. And it seems that the more specific features the dataset share, the better CycleGAN model performs. There are big potentials to use CycleGAN or Pix2Pix mapping to transfer city styles and geographic features from one place to another, which provides us insight into its future use in urban planning or visualization.

To some extent, Geographic feature is a very elusive concept to perceive. We tend to develop sense of place based on our impression and experience subconsciously. However, it is very difficult to quantify our feelings. CycleGAN may provide a new perspective on how we think of a certain place with a vivid way.

There are many potential directions worth further exploration. Firstly, CartoDB Positron basemap has limited color schemes and symbols. Although this can give GANs a great deal of latitude to generate fake images, it also ‘confuses’ the machine. Therefore other map style may improve the performance of the model efficiently. Second, GANs are often regarded as an unsupervised learning algorithm, but more expertise from geographers or spatial scientists is necessary to develop the training set and the network architectures. Moreover, instead of taking the whole city as the input, using certain type of landscape or building as training set may generate much better images.

References

- 1 Jiaoyan Chen and Alexander Zipf. Deepvgi: Deep learning with volunteered geographic information. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 771–772, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee. doi:10.1145/3041021.3054250.
- 2 Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Advances in Neural Information Processing Systems*, 3:2672–2680, 2014.
- 3 Zhi He, Han Liu, Yiwen Wang, and Jie Hu. Generative adversarial networks-based semi-supervised learning for hyperspectral image classification. *Remote Sensing*, 9(10):1042, 2017. doi:10.3390/rs9101042.
- 4 Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016. arXiv:1611.07004.
- 5 Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc. URL: <http://dl.acm.org/citation.cfm?id=2999134.2999257>.
- 6 Huina Mao, Yingjie Hu, Bandana Kar, Song Gao, and Grant McKenzie. Geoai 2017 workshop report. *The 1st ACM SIGSPATIAL International Workshop on GeoAI: @AI and Deep Learning for Geographic Knowledge Discovery: Redondo Beach, CA, USA - November 7, 2016*, 9:25–25, 01 2018.
- 7 Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. arXiv:1409.1556.
- 8 Liangpei Zhang, Lefei Zhang, and Bo Du. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 4(2):22–40, June 2016. doi:10.1109/MGRS.2016.2540798.
- 9 Bo Zhao and Daniel Z. Sui. True lies in geospatial big data: detecting location spoofing in social media. *Annals of GIS*, 23(1):1–14, 2017. doi:10.1080/19475683.2017.1280536.
- 10 Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, volume 00, pages 2242–2251, Oct. 2018. doi:10.1109/ICCV.2017.244.

A Safety Evaluation Method of Evacuation Routes in Urban Areas in Case of Earthquake Disasters Using Ant Colony Optimization and Geographic Information Systems

Kayoko Yamamoto

Graduate School of Informatics and Engineering, National University of
Electro-Communications, Tokyo, Japan
k-yamamoto@is.uec.ac.jp

Ximing Li

Graduate School of Information Systems, National University of Electro-Communications,
Tokyo, Japan
liximing007@gmail.com

Abstract

The present study aims to propose the method for the quantitative evaluation of safety concerning evacuation routes in case of earthquake disasters in urban areas using Ant Colony Optimization (ACO) algorithm and Geographic Information Systems (GIS). Regarding the safety evaluation method, firstly, the similarity in safety was focused on while taking into consideration road blockage probability, and after classifying roads by means of the hierarchical cluster analysis, the congestion rates of evacuation routes using ACO simulations were estimated. Based on these results, the multiple evacuation routes extracted were visualized on digital maps by means of GIS, and its safety was evaluated. Furthermore, the selection of safe evacuation routes between evacuation sites, for cases when the possibility of large-scale evacuation after an earthquake disaster is high, is made possible. As the safety evaluation method is based on public information, by obtaining the same geographic information as the present study, it is effective in other areas regardless of whether the information is of the past and future. Therefore, in addition to spatial reproducibility, the safety evaluation method also has high temporal reproducibility. Because safety evaluations are conducted on evacuation routes based on quantified data, highly safe evacuation routes that are selected have been quantitatively evaluated, and thus serve as an effective indicator when selecting evacuation routes.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases Large-Scale Evacuation, Evacuation Route, Safety Evaluation, Earthquake Disaster, ACO (Ant Colony Optimization), GIS (Geographic Information Systems)

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.68

Category Short Paper

1 Introduction

From the experiences gained through the Great Hanshin earthquake (1995) as well as the Great East Japan earthquake (2011), in recent years, Japan has been focusing on disaster reduction by means of self and mutual help. In case of earthquake disasters, especially in crowded urban areas, many road blockages are likely to occur due to secondary disasters



© Kayoko Yamamoto and Ximing Li;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 68; pp. 68:1–68:7

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

which include the collapse and combustion of buildings. Additionally, if an earthquake disaster occurs during a big event such as the Tokyo Olympics and Paralympics, which will be held in Japan in 2020, busy urban areas especially around the stadiums are expected to be crowded with evacuees. Therefore, in such cases, it is necessary to put emphasis on how to reduce damage for risk management in central Tokyo.

In order to make a quick and safe “escape” from disasters, a clear evacuation plan, or more specifically, an evacuation route must be developed. Current efforts to develop such evacuation plans by means of self and mutual help include activities such as walking the streets for disaster prevention in addition to disaster drills using maps, and an example of the latter is Disaster Imagination Game (DIG). However, as the main purpose to create evacuation routes using such activities is the promotion of disaster prevention awareness and disaster prevention education, they are not quantitatively evaluated. As a result, the evacuation plans developed in the way described may be influenced by the developers’ subjective thinking, and its practicability is left uncertain.

Based on the backdrop mentioned before, assuming a large-scale evacuation in case of an earthquake disaster, the present study aims to propose the method for the quantitative evaluation of safety concerning evacuation routes in urban areas using Ant Colony Optimization (ACO) algorithm and Geographic Information Systems (GIS). More specifically, the multiple evacuation routes extracted using ACO algorithm will be visualized on digital maps by means of GIS, and its safety will be evaluated. Based on the evaluation results, the present study will provide effective information concerning disaster reduction through self and mutual help, namely, the development of evacuation plans by individuals and voluntary disaster prevention organizations in regional communities.

2 Related Work

The present study is related to (1) evacuation routes; (2) road blockages; (3) Spatial analysis using geographic information; and (4) application of ACO algorithm for route searches. The present study comes under the first category of studies, and proposes the safety evaluation method of evacuation routes referring to the results of studies related to (2), (3), and (4). In studies related to (1), though evacuation routes were extracted, its safety was not yet evaluated. Additionally, comparing the related studies, the present study particularly targets crowded urban areas as well as densely populated areas, and in order to provide effective information concerning the development of evacuation plans for earthquake disasters, originality and usefulness will be displayed when proposing the method for the quantitative evaluation of safety concerning evacuation routes. More specifically, assuming a large-scale evacuation in case of an earthquake disaster, the present study will propose the method for the quantitative evaluation of safety concerning evacuation routes in urban areas using ACO algorithm for risk management, in order to provide effective information for the development of evacuation plans using self-help and mutual help methods, which are more direct forms of natural disaster reduction help.

3 Safety Evaluation Method of Evacuation Routes

3.1 Evaluation Method and Framework

In the present study, firstly in this section, after selecting the evaluation target area, the evacuation rules, which is the most important factor for the safety evaluation method of

evacuation routes, will be introduced. In Section 4, a road network data using GIS will be made. Additionally, Road blockage probability per road and the number of estimated populations of evacuees in roads will be calculated, and will be added to the road network data as attributes. A hierarchy cluster analysis using the above data will be conducted, and roads will be classified by focusing on the similarity in safety.

In Section 5, an evaluation experiment of ACO algorithm concerning the evaluation target area will be conducted. More specifically, a simulation of congestion conditions of evacuation routes in the target area will be conducted, the parameter of the most valid evaluation results will be applied to ACO, and the congestion rate of each evacuation route will be calculated. ACO is a search method which mimics the path generation process when the ant is on the hunt for food, and it has been applied to combinatorial optimization problems such as the Travelling Salesman Problem (TSP), and has had many effective research results. Ants use pheromones to communicate while they move as a group, resulting in a type of organized system. ACO uses this system formation process for searches.

As the present study assumes a large-scale evacuation, in section 6, evacuation routes with high congestion rates will be extracted and visualized on a digital map by means of GIS, and its safety will also be evaluated. In the present study, the ArcGIS Ver.10.1 of ESRI will be used as GIS, and making of road network data, spatial analysis and the visualization of evacuation route choices will be conducted.

3.2 Selection of Evaluation Target Area

Large-scale evacuations are when many people around venues of large events have to evacuate. As the Olympics and Paralympics will be held in Tokyo Metropolis in 2020, an increase in tourists around the stadiums and other venues is expected. Therefore, if an earthquake disaster occurs in such areas, the probability of a large-scale evacuation is high. Consequently, Hongo district of Bunkyo Ward, an area close to the Tokyo Dome Stadium where a lot of sports games and entertainment events are frequently held, is chosen as the evaluation target area.

3.3 Evacuation Rules in Evaluation Target Area

3.3.1 Order of Evacuation

If an earthquake disaster occurs in Tokyo Metropolis, especially in crowded urban areas, many road blockages are likely to occur due to secondary disasters which include the collapse and combustion of buildings. Therefore, in Tokyo Metropolis, the principal of a two-step evacuation on foot must be kept as an evacuation rule in case of an earthquake disaster. In this two-step evacuation, as a first step in the evacuation from the occurrence of a disaster, evacuees will evacuate to a temporary gathering sites and the damage situation will be confirmed. After checking the damage situation, if the temporary gathering site is seen as dangerous, an evacuation to a wide-area evacuation site (more than 10 hectares) will be made as part of the second step. On the other hand, if the damage caused by fires following the disaster extends significantly and the evacuation to temporary evacuation site is seen as dangerous, evacuees will be instructed to go directly to wide-area evacuation sites (direct evacuation). In the present study, in order to evaluate the safety of evacuation route assuming a large-scale evacuation, the travel between evacuation sites will be assumed and the safety of route between evacuation sites will be evaluated.

3.3.2 Assignment of Evacuation Sites

As with the target area in the present study, University of Tokyo and an open space near the Dome Stadium are assigned as a wide-area evacuation site by Bunkyo Ward. Additionally, nine locations are also assigned as temporary evacuation sites. Therefore, in the present study, a total of ten evacuation sites have been set as both the evacuation starting points as well as destination points, and searches for evacuation routes will be conducted.

3.3.3 Evacuation Distance and Time for Evacuation Routes

Based on material from the Urban Disaster Prevention Office of the City Bureau of the Ministry of Construction and the Director-General for Disaster Management in the Cabinet Office, taking into consideration walking speed and evacuation time in times of a disaster, locations that are within 2 km of walking distance are assigned as wide-area evacuation sites. Regarding walking speed, although it is generally said to be about 4 km/h, taking into consideration the elderly and children as well as the fact that this is in case of a disaster, walking speed is considered to be 2 km/h which is half the normal speed (speed can drop to 1 km/h if in the dark).

Regarding evacuation time, from the fatality occurrences according to the cause in the Great Kanto Earthquake in 1923, it became evident that the fatalities caused by fire rapidly increased 3 hours after the earthquake, and that the first hour after earthquake quickly passed by with the transportation of injured people, first-aid firefighting, and situation grasping. Therefore, although this leaves 2 hours for evacuation, if another hour is allowed as a margin, only an hour is left for the actual evacuation time. Hence, the evacuation distance for an hour of evacuation time is around 2 km.

This shows that for evacuation route in case of wide-area evacuation, the evacuation distance must be within 2 km and a route that is within an hour of evacuation time is desirable. As mentioned in the previous section, because the present study sets all evacuation sites including temporary evacuation sites as targets, evacuation routes within 1 km of evacuation distance and 0.5 hours of evacuation time are considered desirable.

4 Creation of Road Information and Road Classification

4.1 Creation of Road Information Using GIS

In the present study, using the land use data and building use data developed by the Tokyo Metropolitan Office, the road blockage probability will be calculated from the relationship between the debris width buffer of wooden buildings along the road and the road's centerline located in front of the buildings. Therefore, the road centerlines will be extracted from the above land use data to make the node data with intersection and junctions as well as link data with roads connecting the intersections and junctions. By integrating the node data and link data, a road network data will be made. The the road blockage probability was added to the road network data as an attribute. Additionally, using the population data provided by the Ministry of Internal Affairs and Communications, and the above building use data, the estimated populations of evacuees were calculated in roads at the time of a disaster outbreak in both cases of daytime (AM 7:00–PM 6:00) and nighttime (PM 6:00–AM 7:00). The estimated populations of evacuees were also added to the road network data as attributes, and were used in the ACO when calculating evacuation time for a route.

4.2 Road Classification

Focusing on the similarity in safety and setting the road blockage probability calculated on the basis of building collapse risk, all roads in the evaluation target area were classified by the hierarchy cluster analysis using MATLAB. The Euclidean distance as the distance between objectives, and Ward's method for calculation of the distance between clusters are used. Cluster 1 has the lowest safety, and safety levels are classified into 5 levels. The road classification results will be referred to when estimating the congestion rate using ACO.

5 Evaluation Experiment of the ACO Algorithm

5.1 Method

5.1.1 Congestion Estimation Using the Back-Track Method

In this section, an evaluation experiment will be conducted to demonstrate the effectiveness of the ACO algorithm. Using the information of large earthquakes obtained from a rescue simulator, the ever-changing disaster conditions will be replayed. Additionally, route searches using ACO, which is a crowd flow model and metaheuristic solution, will be conducted to reduce the amount of calculations and to estimate real-time congestion in evacuation routes. In the evaluation experiment in this section, evacuation starting points and destination points are all considered as evacuation sites, and the congestion condition of evacuation route between evacuation sites will be estimated.

In the present study, when trying to obtain the number of evacuation routes from geographical information, a simulation by the back-track method using MATLAB will be conducted. The back-track method is a type of search that prioritizes depth, and when searching for a solution using this method, a potential process will be tested according to order. If it becomes evident that a solution cannot be found in a certain process, it will go back to the previous step and a different process will be tested.

5.1.2 Simulation Process

According to the two-step procedures, the simulation will be conducted:

- (1) 1st tour:
 1. All ants will select an evacuation route at random with the same probability;
 2. The evaluation function of the evacuation routes selected by each ant will be calculated;
 3. Based on the evaluation function value selected in ii), pheromones will be attached to the evacuation routes;
- (2) From the 2nd tour on:
 4. All ants will select evacuation routes following the pheromones;
 5. The evaluation function of the evacuation routes selected by each ant will be calculated;
 6. Based on the evaluation function value of each ant, the pheromones on the evacuation routes will be updated.

5.2 Simulation and Evaluation Experiments Concerning the Target Area

In order to verify the effectiveness of the original simulation of the present study, a comparison of three different simulation results will be conducted, by changing the number of ants in correspondence with the number of evacuees and the amount of pheromones adjusted according to the process in the previous section. By doing these, the congestion rate of each road will be estimated. In order to compare the difference in results depending on

the number of ants, all parameters except for the number of ants (100, 200, 300) is made the same, and simulations for three different cases were conducted. Because there are ten evacuation sites within the target area, the number of solutions where pheromones can be updated is also seven. Additionally, the number of evacuees is represented in the number of ants. Taking into consideration the daytime and nighttime populations within the target area, the number of ants is 300 for daytime population and 100 for nighttime population.

6 Safety Evaluation of Evacuation Routes Assuming a Large-scale Evacuation

In this section, the simulation results of the previous section will be applied to the evaluation target area, and evacuation routes with high congestion rates will be extracted. From these evacuation routes, those that meet the three conditions will be extracted with reference to section 3.3.2. The extracted route will be visualized on digital maps by means of GIS and its safety will be evaluated: (i) Distance between evacuation sites is less than 1 km; (ii) Evacuation time between evacuation sites is under 0.5 hours; (iii) Congestion rate is more than 80 percent. Considering the congestion rate that represents the difficulty in activity in times of a disaster, each evacuation site will be selected as evacuation starting points. In the present study, in order to estimate the congestion rate of evacuation route between evacuation sites, the closest evacuation site to each evacuation site will be selected as evacuation destinations.

The simulations are conducted, and ACO parameters set in the cases of daytime evacuation (300 ants) and nighttime evacuation (100 ants). Since two evacuation sites which are close to the Tokyo Dome Stadium, the areas around these sites require attention especially when large-scale events are being held. Additionally, it is evident that evacuation routes with high congestion rates differ according to whether it is daytime or nighttime.

7 Discussion and Conclusion

In the present study, the multiple evacuation routes extracted were visualized on digital maps by means of GIS, and its safety was evaluated. Furthermore, the selection of safe evacuation routes between evacuation sites, for cases when the possibility of large-scale evacuation after an earthquake disaster is high, is made possible. Additionally, using the safety evaluation method in the present study, if all data and research information is updated with the future technology developments and advances in related fields, it will be possible to update and provide even more accurate information.

From land use data and building data, road network information which would be an evacuation hindrance in disasters situations was created in the present study. Because the safety evaluation of evacuation routes based on current building and road layout conditions is made possible by estimating evacuation routes with high congestion rates based on such information, it can be said that the safety evaluation method has high spatial reproducibility. Furthermore, as the safety evaluation method is based on public information, by obtaining the same geographic information as the present study, it is effective in other areas regardless of whether the information is of the past and future. Therefore, in addition to spatial reproducibility, the safety evaluation method also has high temporal reproducibility.

However, regarding spatial reproducibility, it is essential to modify the safety evaluation method, paying attention to the differences concerning types of disasters and secondary disasters between regions. In the present study, assuming a large-scale evacuation in case of


an earthquake disaster and targeting a fireproofed area to conduct the safety evaluation of evacuation routes, it is required to just consider building collapse risks. On the other hand, on the same assumption as the present study in the area has a high concentration of wooden dwellings, it is necessary to consider fire risk in addition to building collapse risk.

Analysis of Irregular Spatial Data with Machine Learning: Classification of Building Patterns with a Graph Convolutional Neural Network

Xiongfeng Yan

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


xiongfeng.yan@whu.edu.cn

 <https://orcid.org/0000-0003-4748-464X>

Tinghua Ai

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

tinghuaai@whu.edu.cn

 <https://orcid.org/0000-0002-6581-9872>

Abstract

Machine learning methods such as Convolutional Neural Network (CNN) are becoming an integral part of scientific research in many disciplines, the analysis of spatial data often failed to these powerful methods because of its irregularity. By using the graph Fourier transform and convolution theorem, we try to convert the convolution operation into a point-wise product in Fourier domain and build a learning architecture of graph CNN for the classification of building patterns. Experiments showed that this method has achieved outstanding results in identifying regular and irregular patterns, and has significantly improved in comparing with other methods.

2012 ACM Subject Classification Computing methodologies → Machine learning

Keywords and phrases Building pattern, Graph CNN, Spatial analysis, Machine learning

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.69

Category Short Paper

Acknowledgements The first author would like to express thankfulness to Shuaixiong Rao for providing the computing resources, which is very helpful to this paper.

1 Introduction

With the improvement of computing power and the advent of data era, machine learning methods are becoming an integral part of scientific research in many disciplines. As a supervised learning method, CNN has excellent performance in many fields, such as computer vision and natural language processing. These successes are mainly attributed to its two important properties: first, inspired by neuronal processing, CNN focuses on local structures (Local Receptive Fields), and combines them into a whole, which can be applied to classification or identification. Second, local structure of different regions can be detected by the same convolution kernel, that is, weights sharing. The former accords with the decomposability of object and hierarchy of cognition, the latter reduces complexity and improves learnability.

However, it should be noted that both the local connection and weight sharing properties require that the local structure of data is fixed, normative, and can be clearly defined. For example, the images in visual analysis are organized in a grid of pixels as a processing unit, and sentences in natural language processing are organized in a linear arrangement of words as a processing unit. But, for most of spatial graphics data in GIS fields, the arrangement,



© Xiongfeng Yan and Tinghua Ai;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 69; pp. 69:1–69:7

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

combination, or connection between objects may be more diversified, and it is often difficult to satisfy this requirements of specification, for example, the group relationships between plane buildings. Therefore, this kinds of data cannot directly use these powerful learning methods for pattern recognition and knowledge discovery.

Although spatial data is irregular and cannot be organized according to grid or linear structure, it is still possible to represent by graph structure. The graph cannot define a convolution operation in the vertex domain directly, but in virtue of graph Fourier transform and convolution theorem, the operation can be transformed into a point-wise product in the Fourier domain, which is similar to the transformation of spatial domain convolution into frequency domain convolution in image processing. Based on this idea, we propose a graph CNN for identifying patterns and mining knowledge of irregular spatial data.

In this study, we focus on using machine learning to solve the problem of building group pattern recognition, which can be used in many fields, such as urban morphology and environmental analysis. Although the related researches have been carried out for decades, there are still some problems such as incomplete taxonomy and inconsistent recognition rules. The introduction of machine learning method is an effective attempt and supplement to solve such classical problems in spatial analysis. In the following sections, we will describe detailed methods, then conduct experiments and compare with other similar methods, and finally discuss and conclude this study.

2 Methodology

2.1 Definition of Building Pattern Classification

Building patterns refer to visually salient structures exhibited collectively by a group of buildings[4]. Traditional patterns detection methods are to predefine some specific perceptual rules according to the characteristics of azimuth angle, direction difference and proximity, and then to inquire whether there is a local group that satisfy such rules[3][8][10]. But these rules are difficult to formalize and too rigid, which inevitably lead to an unsatisfactory result[6].

Similar to image processing, determining which pattern a building group visually belongs to is essentially an issue of classification. A building group is an analogy to a picture, and each building is analogy to the pixel and its shape features are analogy to color channels, in spite of the relationship between them is an irregular graph structure, not a fixed grid.

2.1.1 Features of single building

Single building has spatial features that describe its graphical structures and semantic features that describe its attributes, which in combination can effectively reflect its basic form. For the description of these features, dozens of indices have been proposed[8]. In this study, we mainly consider area A_b , main direction α , and three shape indices including length-width ratio R_{lw} , area ratio R_A , and compactness C . These indices illustrated in figure 1.

2.1.2 Graph representation of building group

Graph is an ideal tool to describe the relationships between multiple objects. Delaunay triangulation (DT) and Minimum Spanning Tree (MST) are the two most commonly used ways due to they can take spatial constraints and other contextual constraints into consideration, such as proximity.

Regardless of whether DT or MST, they can be defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$, where \mathcal{V} and \mathcal{E} is a finite set of $|\mathcal{V}| = n$ vertices and edges, respectively, $\mathcal{W} \in \mathbb{R}^{n \times n}$ is a adjacency matrix

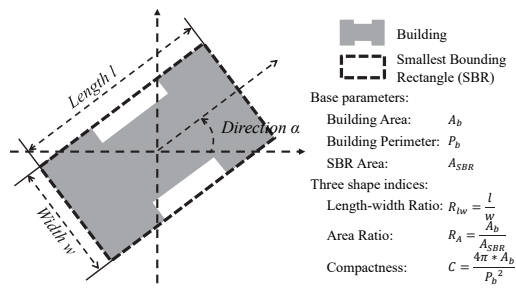


Figure 1 Input feature indices.

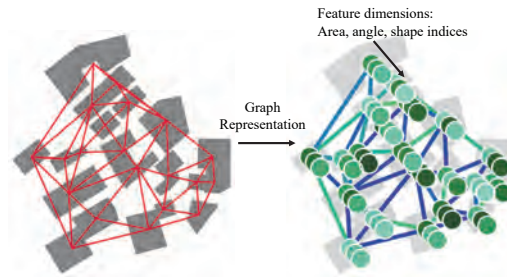


Figure 2 Graph construction.

encoding the weight between two vertices, and each vertex also contains one or several features, as seen in figure 2.

2.2 Graph Convolutional Neural Network

2.2.1 Graph Fourier transform

Fourier transform is an effective tool in signal analysis and image processing, it decomposes an original function (e.g., a signal or image) into the frequencies that make it up. The process is essentially a linear transformation by using given orthogonal basics $\langle f, e^{i\omega t} \rangle$.

For the graph structure, we utilize the eigenvectors χ_ℓ of Laplacian as the decomposition basics instead of complex exponentials, then define the graph Fourier transform as:

$$\hat{f}(\lambda_\ell) = \langle f, \chi_\ell \rangle = \int_{n=1}^N \chi_\ell^T(n) f(n) \quad (1)$$

Where, λ_ℓ is the eigenvalue and the inverse Fourier transform as:

$$f(n) = \int_{n=1}^N \hat{f}(\lambda_\ell) \chi_\ell(n) \quad (2)$$

This definition is precise analogy to the classical one, and it can be interpreted as an expansion of f in terms of the eigenvectors of the Laplacian[5][7].

2.2.2 Convolution on graph

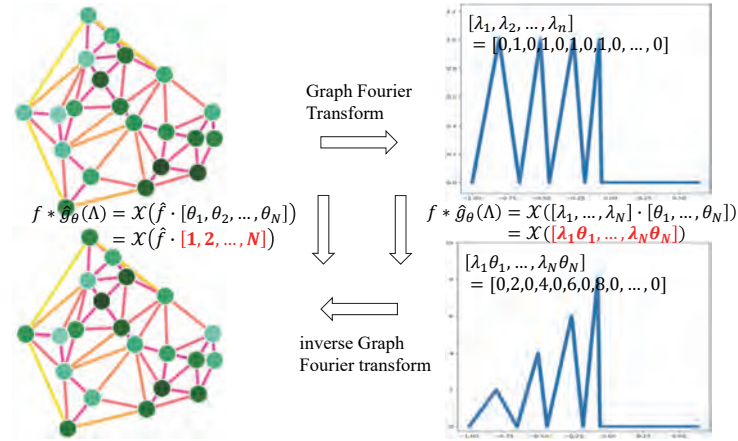
As we cannot conduct the convolution in vertex domain directly, we can try to convert this operation into a point-wise product in Fourier domain by means of graph Fourier transform and convolution theorem, and it can be defined as:

$$(f * g)(n) = \int_{n=1}^N \hat{f}(\lambda_\ell) \hat{g}(\lambda_\ell) \chi_\ell(n) \quad (3)$$

Using notation from the matrix theory, the convolution also can be written as:

$$f * g = \mathcal{X}((\mathcal{X}^T f) \bullet (\mathcal{X}^T g)) = \mathcal{X}diag(\hat{g}(\lambda_1), \dots, \hat{g}(\lambda_N)) \mathcal{X}^T f = \mathcal{X}\hat{g}_\theta(\Lambda) \mathcal{X}^T f \quad (4)$$

Where, $diag(\hat{g}(\lambda_1), \dots, \hat{g}(\lambda_N))$ can be understood as a set of free parameters θ in the Fourier domain (the Eigenspaces of Laplacian), or a function of the eigenvalues $\hat{g}(\Lambda)$.



■ **Figure 3** The convolution of a graph f and a kernel of free parameters.

2.2.3 Polynomial approximation for fast localized convolution

The above definition of convolution operation on graph still has two limitations: 1) in each operation, the Eigen decomposition need to be performed, which will bring lots of computational cost; 2) without considering the locality in space, the features of a vertex may be related to global vertices after this operation, it is not consistent with the local connection property of classical CNN[1][2].

In response to these problems, Hammond[5] proposed a fast localized convolution based on low-order polynomial approximation that represent $\hat{g}_\theta(\Lambda)$ as a polynomial function of eigenvalues:

$$\hat{g}_\theta(\Lambda) = \int_{k=0}^{K-1} \theta_k \Lambda^k \quad (5)$$

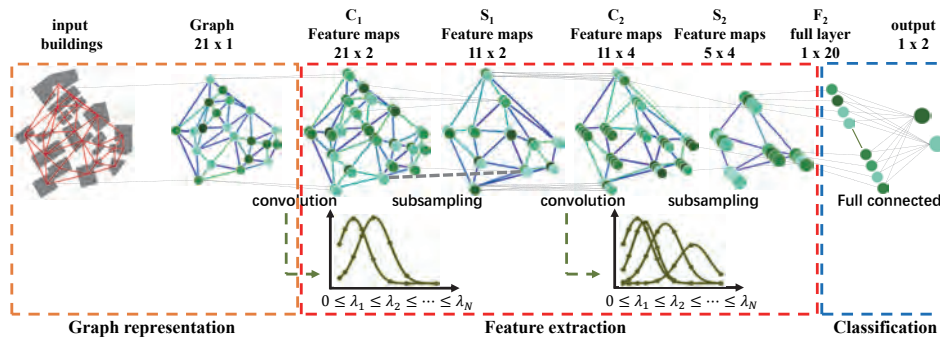
Then, the Equation (4) can be rewritten as:

$$f * g = \mathcal{X} \left(\int_{k=0}^{K-1} \theta_k \Lambda^k \right) \mathcal{X}^T f = \left(\int_{k=0}^{K-1} \theta_k (\mathcal{X} \Lambda^k \mathcal{X}^T) \right) f = \left(\int_{k=0}^{K-1} \theta_k \mathcal{L}^k \right) f \quad (6)$$

As we can see, no need to perform the Eigen decomposition anymore, and the feature values of vertex are related only to its K-order neighboring vertices, which satisfies the locality in space.

2.2.4 Architecture of convolutional neural network on graph

Based on the above-defined graph convolution, we propose a learning architecture of CNN on graph for the classification of building patterns, as seen in figure 4. This architecture includes convolutional, subsampling, and full connected layers, where subsampling layer is optional and the full connected layer is the same as the classical CNN. We input a building group that has already been represented as graph to this architecture, after the steps of feature extraction and classification, we can get the probability that it belongs to each class and choose the class with maximum probability as the final classification result.



■ **Figure 4** The architecture of convolutional neural network on graph.

■ **Table 1** Accuracies of the proposed method and other methods.

Method	SVM	Random Forest	Graph CNN
Accuracy	90.2%	93.4%	98.04%

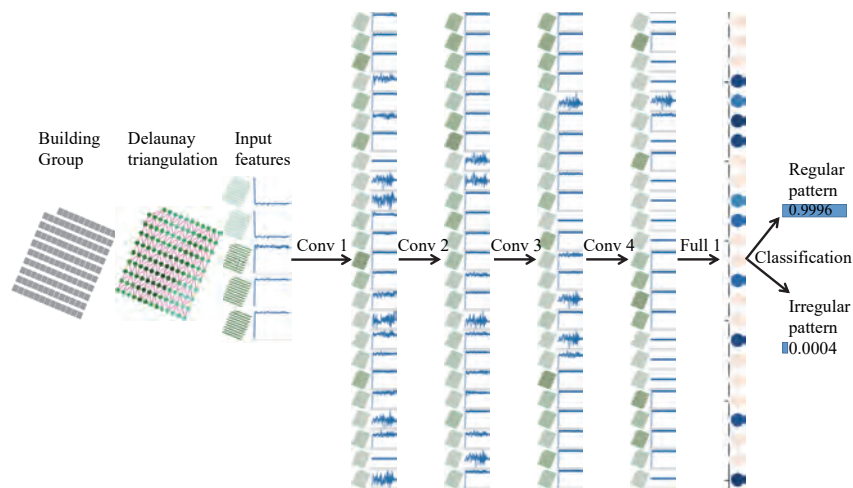
3 Experiments

The experimental buildings were extracted from a large-scale 1:2000 topological map of the city of Guangzhou, China. We divided them into separated groups by using road network division and clustering techniques, and each group contains 20-128 buildings. Then, we manually identified the two patterns of regular and irregular from all groups. Each group was estimated by at least three participants to ensure the correctness, and these ambiguous groups were discarded. At last, there are 2647 and 2646 available groups for regular and irregular pattern respectively and contain a total of 318 598 buildings. Each group can serve as a sample for the graph CNN, all samples were split into training, validation and test sets by 6:2:2, and input features of all data were standardized using training set.

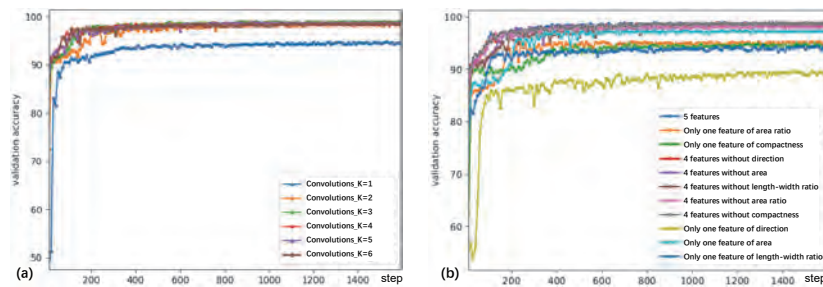
We used a shallow graph CNN architecture with four convolutional layers and one full connected layer to test the datasets, each convolutional layer contains 24 third-order polynomial convolution kernels. The more convolutional layers, the more complex the model is and the more samples are required. In addition, regularization and dropout techniques are also used to control the complexity, and their parameters are referenced from empirical values and fine-tuned. The accuracy is 98.04%, which is better than that of SVM[9] and random forest[6] methods, the comparison results are shown in table 1.

The activation of a sample is shown in figure 5 and the input volume stores the graph of building group (left) and the last volume holds the scores for each class (right).

In this model, the order K of the polynomial is one of the important parameters. We tested the values of 1, 2, 3, 4, 5, and 6 respectively, the performances on the validation set are shown in figure 6a. The comparison found that it achieved the best performance when $K=3$. The larger of K , the more complex of the training and the longer it takes. We further tested the effect of input features of individual building on the classification of group patterns. We tried to train and learn by using 4 features of them or only one, these results are shown in figure 6b and we found that the area was one of the important features and the accuracy could also reach 96.34% when only area feature was used. This may be due to the fact that areas of buildings in a regular pattern are more homogeneous.



■ **Figure 5** The activations of an example graph CNN architecture.



■ **Figure 6** Performances when taking different K values or inputting different features.

4 Discussion and Conclusion

As a classical problem in the analysis of irregular spatial data, the traditional building pattern classification method needs to manually extract features and design rules for specific patterns. In this paper, we propose a graph CNN that represent the building groups by graph and convert the convolution of vertex domain into a point-wise product in Fourier domain. It can directly extract patterns characteristics based on the training and learning of sample data. Experiments showed that proposed method has achieved outstanding results in identifying regular and irregular patterns, and has significantly improved in comparing with other methods. Meanwhile, it has great potential to extend to other analyses of irregular spatial data, such as classification of road patterns and identification of point clouds.

The difficulties of this method lie in the selection of input features and the training process. We have selected five features for experiments, but there are still many other descriptive indices that can be selected. Determining which indices can better describe building patterns and applying them to training and learning still requires more experiments, and the principal component analysis may be a worthwhile approach to try. The training of graph CNN requires more samples, otherwise it will easily lead to overfitting, especially for deep networks with many convolutional layers. In the follow-up work, Volunteer Geographic Information (VGI) is a desirable and feasible data source.

References

- 1 Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. *CoRR*, 2013. URL: <http://arxiv.org/abs/1312.6203>.
- 2 Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *CoRR*, 2016. URL: <http://arxiv.org/abs/1606.09375>.
- 3 Shihong Du, Liqun Luo, Kai Cao, and Mi Shu. Extracting building patterns with multilevel graph partition and building grouping. *ISPRS Journal of Photogrammetry and Remote Sensing*, 122:81–96, 2016. doi:10.1016/j.isprsjprs.2016.10.001.
- 4 Shihong Du, Mi Shu, and Chen Chieh Feng. Representation and discovery of building patterns: a three-level relational approach. *International Journal of Geographical Information Science*, 30(6):1161–1186, 2015. doi:10.1080/13658816.2015.1108421.
- 5 David K. Hammond, Pierre Vandergheynst, and Rémi Gribonval. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150, 2009. doi:10.1016/j.acha.2010.04.005.
- 6 Xianjin He, Xinchang Zhang, and Qinchuan Xin. Recognition of building group patterns in topographic maps based on graph partitioning and random forest. *ISPRS Journal of Photogrammetry and Remote Sensing*, 136:26–40, 2018. doi:10.1016/j.isprsjprs.2017.12.001.
- 7 David I Shuman, Benjamin Ricaud, and Pierre Vandergheynst. Vertexfrequency analysis on graphs. *Applied and Computational Harmonic Analysis*, 40(2):260–291, 2016. doi:10.1016/j.acha.2015.02.005.
- 8 Zhiwei Wei, Qingsheng Guo, Lin Wang, and Fen Yan. On the spatial distribution of buildings for map generalization. *Cartography and Geographic Information Science*, pages 1–17, 2018. doi:10.1080/15230406.2018.1433068.
- 9 Liqiang Zhang, Hao Deng, Dong Chen, and Zhen Wang. A spatial cognition-based urban building clustering approach and its applications. *International Journal of Geographical Information Science*, 27(4):721–740, 2013. doi:10.1080/13658816.2012.700518.
- 10 Xiang Zhang, Tinghua Ai, Stoter Jantien, Kraak Menno Jan, and Molenaar Martien. Building pattern recognition in topographic data: examples on collinear and curvilinear alignments. *GeoInformatica*, 17(1):1–33, 2013. doi:10.1007/s10707-011-0146-3.

Assessing Neighborhood Conditions using Geographic Object-Based Image Analysis and Spatial Analysis

Chi-Feng Yen

Center for Human Dynamics in the Mobile Age, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182, USA.

cyen@sdsu.edu

Ming-Hsiang Tsou

Center for Human Dynamics in the Mobile Age, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182, USA.

mtsou@sdsu.edu

Chris Allen

Center for Human Dynamics in the Mobile Age, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182, USA.

ccchris.allen@gmail.com

Abstract

Traditionally, understanding urban neighborhood conditions heavily relies on time-consuming and labor-intensive surveying. As the growing development of computer vision and GIScience technology, neighborhood conditions assessment can be more cost-effective and time-efficient. This study utilized Google Earth Engine (GEE) to acquire 1m aerial imagery from the National Agriculture Image Program (NAIP). The features within two main categories: (i) aesthetics and (ii) street morphology that have been selected to reflect neighborhood socio-economic (SE) and demographic (DG) conditions were subsequently extracted through geographic object-based image analysis (GEOBIA) routine. Finally, coefficient analysis was performed to validate the relationship between selected SE indicators, generated via spatial analysis, as well as actual SE and DG data within region of interests (ROIs). We hope this pilot study can be leveraged to perform cost- and time- effective neighborhood conditions assessment in support of community data assessment on both demographics and health issues.

2012 ACM Subject Classification Computing methodologies → Computer vision

Keywords and phrases neighborhood conditions assessment, geographic object-based image analysis, spatial analysis

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.70

Category Short Paper

1 Introduction

Socio-economic and demographic data are fundamental components of understanding neighborhood makeup and health condition [13]. Conventional approach to investigate the socio-economic and demographic condition within urban areas relies heavily on time-consuming and labor-intensive surveying that usually causes the lag of socio-economic and demographic changes (e.g. American Census Survey) [3]. With the growing development of computer vision, remote sensing and GIS technology, the lag of socio-economic and demographic data in urban areas can be effectively compromised to some degree. This pilot study aims to



© Chi-Feng Yen, Ming-Hsiang Tsou, and Chris Allen;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 70; pp. 70:1–70:7

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

examine the utility of RS imagery in urban neighborhood conditions assessment at census block within given zip codes, specifically using a popular image classification paradigm: GEOBIA. The research focus is to find SE indicators, generated from the extracted features using spatial analysis, that can represent certain SE or DG conditions in given census-block neighborhoods.

2 Remote Sensing on Neighborhood Condition Assessment

Considerable researches have demonstrated the feasibility of RS satellite imagery in estimation of DG data. [2] exploited both high- and medium- resolution satellite imagery to estimate population distribution for areas lacking census data in support of disaster resilience in Haiti. [1] filled DG data gap within developing countries using per-pixel population estimates generated by a classification and regression trees (CART) and multi-resolution satellite imagery. More recently, [3] identified the significant associations between the presence of specific vehicle models and voter preferences across 200 cities in the US. using convolutional neural network (CNNs) and the Google Street View image dataset.

In terms of SE condition, the contextual features, such as the size of building structures, the abundance of vegetation cover and etc., can reflect the different socio-economic status (SES) in urban areas, particular in residential areas [8, 9]. Specifically, lower SES usually accompanies less vegetation cover and swimming pools but high density of residential buildings [7, 11, 12]. In addition, SES information plays an important role to understand neighborhood health conditions.[4] indicated that neighborhood characteristics (e.g. SE and built environments) impact cancer incidence or outcomes. [6] found that women in high-SES neighborhoods have higher breast cancer-specific survival than in low-SES neighborhoods. [13] illustrated that lower SES is highly related to poorer health condition in developing countries because of the close proximity of people living and insanitary settlement. These characteristics could directly or indirectly ascend disease spread within the neighborhoods.

Although RS has been widely-applied for neighborhood condition assessment, providing more detailed perspectives of neighborhood characteristics is still on demand. Here, we take advantages of ortho high-resolution aerial imagery and GEOBIA to provide detailed and accurate neighborhood conditions assessment. The methodology will be elaborated in the following paragraphs.

3 Methodology

3.1 ROIs

Three ROIs: (i) 92130 Carmel Valley, (ii) 92120 Del Cerro and (iii) 92113 Logan Heights were selected to represent high, medium and low SES, respectively. The selection of ROIs was based on household income data, derived from city-data.com.

3.2 GEOBIA

The term of GEOBIA is specifically for GIScience because of requiring the knowledge in geographic information (GI) to segment and classify RS imagery. Moreover, the objects of GEOBIA are usually associated with natural features (e.g., grassland) or artificial features (e.g., building) [10]. These unique emphases set apart GEOBIA from object-based image analysis (OBIA), which is more used in other disciplines (e.g., computer vision and biomedical imaging) [5]. This pilot study applied GEOBIA to extract selected features

from NAIP imagery. The features are within two main categories, aesthetics and street morphology, that have potential to reflect SE and DG conditions within neighborhoods.

3.3 Feature Extraction and Building SE Indicators

Vegetation cover and swimming pool were two primary features extracted from NAIP imagery via GEOBIA. The following SE indicators were built from these two features via spatial analysis:

(i) Aesthetics:

- Percentage of vegetation cover (veg_percent)
- Swimming pool density(sp_den)
- Percentage of swimming pool area (sp_percent)
- Number of swimming pool (sp_num)

where: percentage of vegetation cover = the area of vegetation within each census block / each area of census block; swimming pool density = the number of swimming pool within each census block / each area of census block; percentage of swimming pool area = the area of swimming pool within each census block / each area of census block; number of swimming pool = total number of swimming pools within each census block.

(ii) Street Morphology:

- Road density (rd_den)
- Road junction density (rd_junction_den)

where: road density = total road length within each census block / each area of census block; road junction density = total number of road junctions within each census block / each area of census block.

3.4 Selection of SE and DG variables

To validate whether the given SE indicators can represent certain SES or DG makeup, we selected the following SE and DG variables from 2015 American Census Survey data.

- T1115_INCOM: Median income; total household income
- F1115_MHV: Median housing value
- T1115_PROF: Total professional, scientific, management, administrative employed civilians age 16 and older
- P1115_I75: Percent individuals with income below / over \$75,000

Coefficient analysis was subsequently performed to assess the association of the SE indicators with surveyed SE and DG data.

4 Results and Discussion

Due to the limitation of pages, this paper will only show the results of high and medium SES in this section.

4.1 Feature Extractions based on GEOBIA

Figure 1a-e shows a subset of NAIP image, swimming pool and vegetation cover in high SES and medium SES, respectively.



■ **Figure 1** The NAIP image with the detection of vegetation cover and swimming pools in high SES and medium SES in San Diego County.

4.2 The correlation of SE Indicators as well as surveyed SE and DG variables

The coefficient outcomes of high SES and medium SES were demonstrated in Table 1 and Table 2, respectively. The highest positive / lowest negative correlation between each SE indicator and individual surveyed variable was highlighted by red / blue.

Table 1 shows that swimming pool density and percentage of swimming pool area have the highest positive correlation with total household income. Number of swimming pool yields the highest positive correlation with total professional, scientific, management, administrative employed civilians age 16 and older. Percentage of vegetation cover yield the highest positive correlation with percent individuals with income below / over \$75,000, while road density has the highest negative correlation with this SE variable.

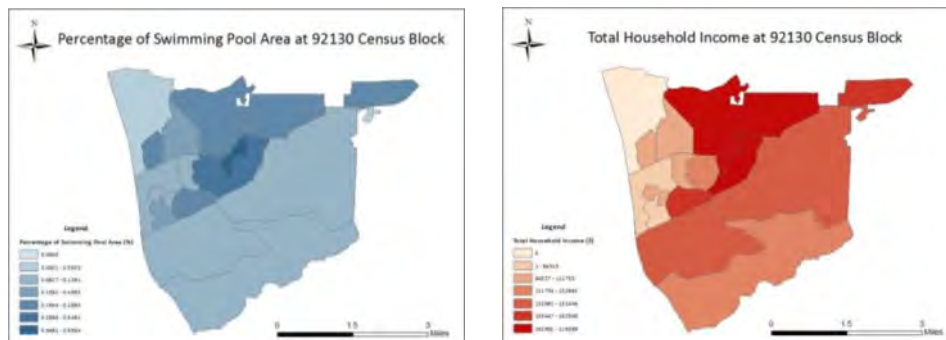
Table 2 demonstrates that three swimming pool-related indicators have the highest positive correlation with median housing value. Percentage of vegetation cover yield the highest positive correlation with total professional, scientific, management, administrative employed civilians age 16 and older. Two road-related indicators have the greatest negative correlation with percent individuals with income below / over \$75,000.

■ **Table 1** The coefficient analysis of high SES neighborhood (92130 Carmel Valley).

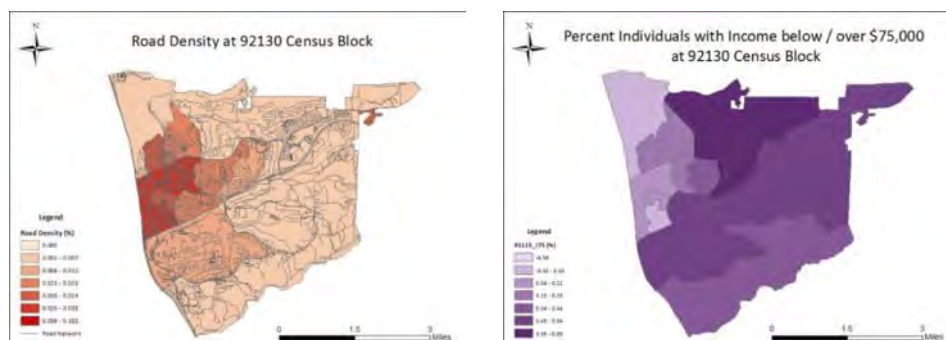
Features / SE or DG variables	T1115_INCOME	T1115_PROF	T1115_MHV	P1115_I75
sp_den	0.60	-0.18	0.42	0.49
sp_percent	0.62	-0.18	0.44	0.51
sp_num	0.45	0.53	0.45	0.45
veg_percent	0.25	0.28	-0.05	0.37
rd_den	-0.35	-0.27	-0.43	-0.62
rd_junction_den	-0.13	-0.21	0.13	-0.11

■ **Table 2** The coefficient analysis of medium SES neighborhood (92120 Del Cerro).

Features / SE or DG variables	T1115_INCOME	T1115_PROF	T1115_MHV	P1115_I75
sp_den	0.31	-0.16	0.44	0.42
sp_percent	0.29	-0.09	0.47	0.40
sp_num	0.10	-0.01	0.49	0.42
veg_percent	0.29	-0.20	0.26	0.44
rd_den	-0.18	0.08	0.13	-0.22
rd_junction_den	-0.39	0.29	-0.12	-0.57

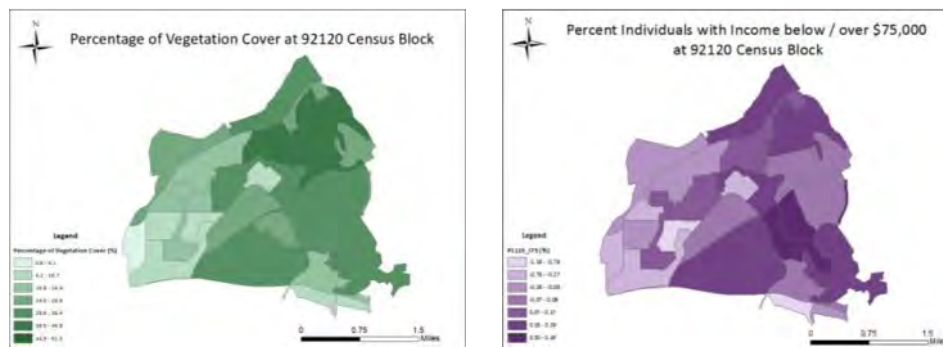


■ **Figure 2** Coefficient=0.62



■ **Figure 3** Coefficient=-0.62

Here we highlighted few pairs of SE indicators and the surveyed variables via Geovisualization (Figure 2-4).



■ **Figure 4** Coefficient=0.44

5 Conclusion

Although all highest positive correlation between SE indicators and surveyed variables are not significant, some certain SE indicators show the potential to assess specific SE or DG conditions within ROIs. Specifically, swimming pool-associated indicators have the greatest correlation with total household income at high SES and with median housing value at medium SES. Vegetation indicator yields the highest correlation with percent individuals with income below / over \$75,000 at both high and medium SES. In terms of negative correlation, road density has the greatest negative correlation with percent individuals with income below / over \$75,000 at high SES, while road junction density meets the greatest negative correlation with this DG variable. In the near future, we plan to incorporate Google Street View into our framework to provide different angle of features that have potential to represent neighborhood conditions.

References

- 1 Derek Azar, Ryan Engstrom, Jordan Graesser, and Joshua Comenetz. Generation of fine-scale population layers using multi-resolution satellite imagery and geospatial data. *Remote Sensing of Environment*, 130:219–232, 2013.
- 2 Derek Azar, Jordan Graesser, Ryan Engstrom, Joshua Comenetz, Robert M Leddy Jr, Nancy G Schechtman, and Theresa Andrews. Spatial refinement of census population distribution using remotely sensed estimates of impervious surfaces in haiti. *International Journal of Remote Sensing*, 31(21):5635–5655, 2010.
- 3 Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences*, 114(50):13108, 2017. URL: <http://www.pnas.org/content/114/50/13108.abstract>.
- 4 Scarlett Lin Gomez, Sally L Glaser, Laura A McClure, Sarah J Shema, Melissa Kealey, Theresa HM Keegan, and William A Satariano. The california neighborhoods data system: a new resource for examining the impact of neighborhood characteristics on cancer incidence and outcomes in populations. *Cancer Causes & Control*, 22(4):631–647, 2011.
- 5 Geoffrey J. Hay and G. Castilla. *Geographic Object-Based Image Analysis (GEOBIA): A new name for a new discipline*, pages 75–89. Springer, 2008.
- 6 Theresa HM Keegan, Salma Shariff-Marco, Meera Sangaramoorthy, Jocelyn Koo, Andrew Hertz, Clayton W Schupp, Juan Yang, Esther M John, and Scarlett L Gomez. Neighbor-

- hood influences on recreational physical activity and survival after breast cancer. *Cancer Causes & Control*, 25(10):1295–1308, 2014.
- 7 Xiaojiang Li, Chuanrong Zhang, Weidong Li, Yulia A. Kuzovkina, and Daniel Weiner. Who lives in greener neighborhoods? the distribution of street greenery and its association with residents' socioeconomic conditions in hartford, connecticut, usa. *Urban Forestry & Urban Greening*, 14(4):751–759, 2015. doi:10.1016/j.ufug.2015.07.006.
 - 8 D Stow, A Lopez, C Lippitt, S Hinton, and J Weeks. Object-based classification of residential land use within accra, ghana based on quickbird satellite data. *International journal of remote sensing*, 28(22):5167–5173, 2007.
 - 9 Douglas Stow. *Geographic object-based image change analysis*, pages 565–582. Springer, 2010.
 - 10 Douglas A Stow, Christopher D Lippitt, and John R Weeks. Geographic object-based delineation of neighborhoods of accra, ghana using quickbird satellite imagery. *Photogrammetric Engineering & Remote Sensing*, 76(8):907–914, 2010.
 - 11 Douglas A Stow, John R Weeks, Sory Toure, Lloyd L Coulter, Christopher D Lippitt, and Eric Ashcroft. Urban vegetation cover and vegetation change in accra, ghana: Connection to housing quality. *The Professional Geographer*, 65(3):451–465, 2013.
 - 12 Francisco J Tapiador, Sylvania Avelar, Carlos Tavares-Corrêa, and Rainer Zah. Deriving fine-scale socioeconomic information of urban areas using very high-resolution satellite imagery. *International journal of remote sensing*, 32(21):6437–6456, 2011.
 - 13 John R Weeks, Arthur Getis, Douglas A Stow, Allan G Hill, David Rain, Ryan Engstrom, Justin Stoler, Christopher Lippitt, Marta Jankowska, and Anna Carla Lopez-Carr. Connecting the dots between health, poverty and place in accra, ghana. *Annals of the Association of American Geographers*, 102(5):932–941, 2012.

Spatial Information Extraction from Text Using Spatio-Ontological Reasoning

Madiha Yousaf

University of Bamberg, Germany
madiha.yousaf@uni-bamberg.de

Diedrich Wolter

University of Bamberg, Germany
diedrich.wolter@uni-bamberg.de

Abstract

This paper is involved with extracting spatial information from text. We seek to geo-reference all spatial entities mentioned in a piece of text. The focus of this paper is to investigate the contribution of spatial and ontological reasoning to spatial interpretation of text. A preliminary study considering descriptions of cities and geographical regions from English Wikipedia suggests that spatial and ontological reasoning can be more effective to resolve ambiguities in text than a classical text understanding pipeline relying on parsing.

2012 ACM Subject Classification Computing methodologies → Information extraction

Keywords and phrases spatial information extraction, geo-referencing, spatial reasoning

Digital Object Identifier 10.4230/LIPICs.GIScience.2018.71

Category Short Paper

Funding This work is supported by the Deutsche Forschungsgemeinschaft (DFG), priority program Volunteered Geographic Information. Financial support is gratefully acknowledged.

1 Introduction

We are involved with a project that aims to develop an automated system capable of interpreting spatial language for resolving place descriptions. While ‘place’ is an inherently complex and elusive concept (cp. [3, 7]), we take a pragmatic approach here: Given a natural language description like “the campground south of Bamberg, near the river”, we seek to identify geographic entities in the OpenStreetMap (OSM) data base¹ that match noun phrases occurring in the sentence that refer to real-world entities, in our example thus identifying a campground, an entity named Bamberg, and a river. Automated interpretation of text is still a challenging task, mainly because of language parsing and the ambiguity of names and human conceptualization. While ambiguity of named entities can be tackled by considering any entity with a matching name found in the database and then applying ranking techniques based on geographic scope [1], there are no easy solutions to tackle failed attempts to parse a piece of text. Ambiguity resolution can be regarded as a task of reasoning since the goal is to identify a single interpretation from a set of candidates which is jointly agreeable with all information given. We are motivated to investigate to which extent reasoning about spatial and ontological properties is also capable of overcoming problems with natural language

¹ <http://www.openstreetmap.org>



parsing. Therefore, we investigate a radical attempt that only performs a very simple analysis of the input text, generating many interpretation candidates. We then apply reasoning to single out the interpretation that is most agreeable. This paper presents our method and a preliminary study that suggests spatio-ontological reasoning is offering powerful means for resolving ambiguous interpretations that can outperform classic interpretation pipelines built around natural language parsing.

2 Approach and Discussion of Related Work

We seek to identify named and unnamed entities in a piece of text. While geo-referencing named entities considers names and spatial relations to other entities [1], dealing with unnamed entities presents a special case that can only exploit spatial constraints and maybe type information. We can thus regard both cases jointly as tasks of ambiguity resolution. One approach is geographic scope resolution which allows potential interpretations to be restricted to within a known scope. Whereas Andogah et al. propose a method based on a set of pre-defined geographic scopes [1], Richter et al. [5] consider granularity effects caused by object types. They state that knowing the finest possible level of granularity with respect to a general ontology of spatial entities is helpful for resolving place descriptions. Both ideas can be integrated by attuning the semantics of relations and queries to focus on results that fit a scope indicated by type and location of geographic entities appearing in the same text. For example, the semantics of “near” can be set according to the granularity of objects and their geographic scope. Exploiting such context information presents a chicken-and-egg problem, though: information obtained by resolving entities is to be employed simultaneously to resolving the entities. This motivates an approach using logic programming techniques since dependencies can be expressed in a declarative manner, abstracting from algorithmic realization. The declarative representation can be regarded as a constraint satisfaction problem (CSP) in which variables correspond to spatial entities. A solution to the CSP is obtained when all variables are geo-referenced by matching them to a spatial database.

Interpretation of a place description can be regarded as a simple cognitive simulation of language interpretation, similar to Tschander et al. [6] who describe an artificial agent capable of following route instruction, using a conceptual-level instruction language. In contrast to that agent, we are mostly interested in interpreting describing statements like “campground south of Bamberg”, rather than processing incremental instructions like “take road R123 south”. Therefore, we are not incrementally interpreting a place description, but aim to build a single declarative description from a single description.

Formal ontologies have been claimed to offer adequate means to represent semantic commitments of a spatial language phrase [2]. Likewise, we employ an ontology-like representation to augment the semantic representation of words (the lexicon). However, we have chosen not to employ formal ontology techniques for two reasons. First, truth semantics of classical ontology languages is binary, i.e., entities belong to a certain class or they do not. In the case of spatial entities and concepts, such crisp classification may be hard to achieve and concepts may vary across individuals. Instead, concepts or relations like ‘near’ may be more adequately represented using a semantic capturing vagueness, e.g., using Fuzzy or probabilistic models. Second, existing ontology languages do not support the spatial domain and manifold spatial relations to the extent required to empower spatial reasoning for computing likely interpretations of a locative phrase.

Since we are involved with natural language texts, application of natural language parsing techniques appears reasonable. Several works make use of different parser and their

input	Bamberg	is	a	town	north	of	Nuremberg.
1. POS tagging	Bamberg:NE			town:N	north:R		Nuremberg:NE
2. named entity resolution	{ID0, ID1, ...}			town:N	north:R		{ID8, ID9, ...}
3. ontological annotation	{ID0, ID1, ...}			settlement	north:R		{ID8, ID9, ...}
4. logic program generation	$(\text{isa}(\text{ID0}, 'settlement') \wedge \text{northOf}('settlement', \text{ID9})) \vee$ $(\text{isa}(\text{ID1}, 'settlement') \wedge \text{northOf}('settlement', \text{ID9})) \vee$ $\dots \text{northOf}(\text{ID0}, \text{ID8}) \wedge \dots) \vee \dots$						

■ **Figure 1** Example of processing steps in generation phase of information extraction (NE: named entity, N: noun, R:relation, ID:denotes reference to objects in OSM database).

modules (like Stanford NLTK chunking or the dependency parser²) for relating the objects and relations between them. However, as we are only interested in spatial entities which correspond to nouns in the input text and the relations holding between them, a parser which has to take the verb of a sentence as starting point may not be necessary. Moreover, we found that no freely available parser was able to resolve references in the text correctly. A wrongly identified reference can easily inhibit correct interpretation of a sentence. As our experiments discussed further below reveal, wrongly identified references are a common problem. By contrast, an unidentified reference can often be inferred from context. The basic idea of our approach is thus to generate all candidate interpretations of references and then apply reasoning to single out the most likely interpretation.

3 Processing Pipeline

The basic idea of our method is to use a set of logical statements as an intermediate representation that over-generalizes information expressed in a sentence. Then, spatio-ontological reasoning is applied to prune off implausible interpretations. The method can therefore be regarded as exhaustive search consisting of a *generation* and *pruning* phase. Both phases rely on the same sources of information:

- an ontology of geographic entity types
- a geographic data base (OpenStreetMap) providing information about entity names, their type with respect to the ontology, and associated geographical information
- a lexicon comprising all nouns that represent geographic entity types and all spatial relations

In the *generation phase* (see Fig. 1 for an example), we process a sentence as follows:

1. Perform part-of-speech (POS) tagging by applying named entity recognition using the geographic database and checking for occurrence in the lexicon. All recognized words are labeled with their category, all other words are discarded. To handle composite expressions of several nouns, (e.g., “art gallery”), nouns immediately following one another get joined and treated as a single noun. In case of ambiguities at this or any later point, all possible options are stored.
2. For all named entities, possible interpretations from the geographic database are retrieved. For example, in case of Bamberg, we would obtain an OpenStreetMap entity referring to the city of Bamberg, depicted as ID0 in Fig. 1, and to the corresponding district

² <https://nlp.stanford.edu/software/>

of Bamberg, ID1, both for Bamberg, Germany and for Bamberg, SC, USA (creating ambiguity in the extracted information).

3. For all nouns ontological type information is obtained from the lexicon. Nouns are then replaced by their ontological type. Every noun is assumed to either represent an unnamed entity (e.g., “**park** in the town of Bamberg”) or a type designator for another noun or named entity (e.g., “park in the **town** of Bamberg”).
4. Possible interpretations are determined as disjunctions by compiling interpretations of words and references of relations:
 - For every relation, a term is constructed combining any word (noun or named entity) appearing before the relation with any word occurring after the relation. The designator for each relation is retrieved from the lexicon.
 - For every noun an ontological “is-a” relation is generated in reference to any other noun or named entity, e.g., is-a(Bamberg, town).

In the *pruning phase* every conjunctive term generated in the generation phase is processed individually, see also Fig. 2 for illustration. A term gets discarded if

- a single noun occurs simultaneously in a “is-a” and a spatial relation, i.e., it would represent ontological information and an unnamed entity simultaneously,
- a noun or named entity in the input is not contained in at least one relation,
- or the grouping of relations violates word order in the input sentence. We disallow for relational statements $r(w_a, w_b) \wedge r'(w_c, w_d)$ if the position in the sentence (denoted $\text{Pos}()$) is in crossed order, i.e., it violates $\text{Pos}(w_a) < \text{Pos}(w_c) \Rightarrow \text{Pos}(w_b) \leq \text{Pos}(w_d)$. For example, in “Bamberg is a town north of Nuremberg, on the river Regnitz” interpretations containing $\text{isa}('Bamberg', 'river') \wedge \text{isa}('town', 'Regnitz')$ get discarded.

After the pruning phase, we search for the conjunctive term which can best be satisfied. This means, for unreferenced nouns a suitable instantiation from the geographic database is searched that agrees with the relations—agreement is measured gradually and summed up. Also, relations between named entities and/or referenced nouns are tested. In case of the example in Fig. 1 we would only find for the entity representing Bamberg in Germany a matching entity Nuremberg such that Bamberg is located north of Nuremberg. The ontological constraint saying Bamberg is a settlement would only be fulfilled for the city of Bamberg, not the administrative region. We thus arrive at the desired interpretation.

4 First Findings

We collected a corpus of place descriptions from English Wikipedia by selecting sentences which present mainly spatial information. We have used the summary part from Wikipedia articles about geographical entities to collect the corpus and have applied our approach to 50 sentences. We also test natural language parsing using the Stanford NTLK parser on the corpus and investigate parser output. One aim of the study is to compare the amount of ambiguities introduced by our over-generalizing method of information extraction to wrongly identified references by the parser. Also, we are interested to learn what kind of spatial and ontological reasoning is required to interpret the output of our approach.

Let us start by considering a first example shown in Fig. 2 showing relations extracted by the parser and by the generation method. For clarity of presentation, no entities were replaced by OpenStreetMap references and no nouns were replaced by ontological types. We write $r(\{n_1, n_2\}, \{n_3, n_4\})$ as shorthand notation to denote that all four interpretations $r(n_1, n_3), r(n_2, n_3), \dots$ are considered. As can be seen, the parser identifies that the town mentioned is located in Upper Franconia, but it does not make the relation between the

	$\overbrace{\text{Bamberg}}^B \text{ is a } \overbrace{\text{town}}^T \text{ in } \overbrace{\text{Upper Franconia}}^{UF}, \overbrace{\text{Germany}}^G, \text{ on the } \overbrace{\text{river Regnitz}}^R$ $\text{close to its } \overbrace{\text{confluence}}^C \text{ with the } \overbrace{\text{river Main}}^M.$
parser output	$\text{in}(T, UF), \text{on}(G, R)$
generation phase	$\text{is-a}(\{B, UF, G, R, M\}, C), \text{is-a}(\{B, UF, G, R, M\}, T),$ $\text{in}(\{T, B\}, \{M, C, R, G, UF\}), \text{on}(\{G, UF, T, B\}, \{M, C, R\}),$ $\text{close}(\{B, T, UF, G, R\}, \{C, M\})$
pruning phase	
ontological	$\text{is-a}(\{B, UF, G, R, M\}, C), \text{is-a}(\{B, UF, G, R, M\}, T),$
spatial	$\text{in}(\{T, B\}, \{M, C, R, G, UF\}), \text{on}(\{G, UF, T, B\}, \{M, C, R\}),$
ordering	$\text{close}(\{B, T, UF, G, R\}, \{C, M\})$

■ **Figure 2** Example of pruning using spatio-ontological reasoning.

named entity ‘Bamberg’ and town explicit. Also, the parser commits wrongly to claiming Germany would be located on the river Regnitz. By contrast, exhaustive search contains all correct interpretations by construction, but also several statements not following from the input text. Applying ontological reasoning one immediately identifies that only named entity Bamberg is of type town. Spatial reasoning reveals, for example, that Upper Franconia is neither located on river Regnitz nor Main. As our approach is not yet prepared to handle geographic names like “Upper Franconia, Germany”, we miss this important piece of information.

In addition to above example, some candidate interpretations generated by exhaustive search that are not valid interpretations of the input text take more effort to reject. In case of *The Historical Museum of Bamberg is a museum located in the Alte Hofhaltung next to the city cathedral*, the interpretation $\text{in}(\text{'museum'}, \text{'city cathedral'})$ cannot easily be rejected if the geographic database also includes a museum in the city cathedral. If, during search for the most likely interpretation, the unintended reference is accepted, then order constraints inhibit any further connection to the named spatial entity “Alte Hofhaltung” (Old Court). So in this case we are relying on preferring the larger set of jointly possible interpretations that involves $\text{in}(\text{'museum'}, \text{'Alte Hofhaltung'})$ and $\text{next_to}(\text{'Alte Hofhaltung'}, \text{'city cathedral'})$ over just $\text{in}(\text{'museum'}, \text{'city cathedral'})$. We have tested our corpus and in initial testing we have found out that in 50% sentences it is providing us information that is not present in the sentence but generated by the algorithm. In many examples, these facts are not incorrect like $\text{in}(\text{'Bamberg'}, \text{'Germany'})$ from the example in Fig. 2. While these unintended but correct interpretation candidates did not inhibit a correct manual interpretation, it remains an open question whether this will hold for automated interpretation on a larger corpus.

Now looking at the parser outputs, we can clearly see that it provides us with limited information. In particular relations from complex language constructs are missing. In case of the output of Fig. 2, the relations apply to different entities which inhibits any chaining by means of reasoning. All in all, the parser is not able to provide a densely connected set of facts that would make spatial or ontological reasoning effective. Carrying out spatial and ontological reasoning manually and comparing residual errors after processing the output of exhaustive search with facts extracted from the parser, we cannot rule out all ambiguous interpretations in 25% of the sentences, but we are facing wrong outputs from the parser in 50% of the cases.

5 Conclusion and Next Steps

This paper outlines an approach to information extraction from text which does not rely on natural language parsing, but employs a simple part-of-speech tagging and applies spatial and ontological reasoning for interpretation. Making spatio-ontological reasoning an explicit step in the interpretation also enables consideration of contextual dependencies. Clearly, exhaustive search does not tackle the fundamental problem of language understanding, but it relies on the assumption that the largest set of statements that can be matched to a geographic database corresponds to the intended interpretation. While our approach is unable to deal with negation or complex language structures, it may indeed be sufficient for typical descriptive texts. In a manual comparison using sentences from English Wikipedia that describe geographic entities we see that reasoning is able to prune off most invalid interpretations, whereas natural language parsing results in some wrong commitments one is unable to recognize in a later processing step.

Before embarking on a comprehensive study to analyze the new method, a comprehensive lexicon and knowledge base have to be prepared and reasoning methods to be automated. Information required to build lexicon and knowledge base are readily available using data sources such as WordNet[4] and OpenStreetMap, yet these need to be linked on a semantic level. We are currently working on implementing the automated reasoning method using these sources in order to arrive at a spatial interpretation of the constraints. To make the approach efficient, a query strategy will be required to avoid costly queries by serializing queries and by focusing on reasonable candidate locations.

References

- 1 Geoffrey Andogah, Gosse Bouma, John Nerbonne, and Elwin Koster. Placename ambiguity resolution. In *Proceedings of Language Resources and Evaluation Conference (LREC)*, pages 4–10, 2008.
- 2 John A. Bateman, Joana Hois, Robert Ross, and Thora Tenbrink. A linguistic ontology of space for natural language processing. *Artificial Intelligence*, 174:1027–1071, 2010.
- 3 Brandon Bennett and Pragya Agarwal. Semantic categories underlying the meaning of ‘place’. In Stephan Winter, Matt Duckham, Lars Kulik, and Ben Kuipers, editors, *Spatial Information Theory: Proceedings of the 8th International Conference, (COSIT-2007), Melbourne, Australia*, volume 4736 of *Lecture Notes in Computer Science*, pages 78–95, Berlin, Heidelberg, 2007. Springer-Verlag.
- 4 Christiane Fellbaum, editor. *Wordnet: An Electronic Lexical Database*. MIT Press, 1998.
- 5 Daniela Richter, Stephan Winter, Kai-Florian Richter, and Lesly Stirling. Granularity of locations referred to by place descriptions. In *Proceedings of Workshops and Posters at the 13th International Conference of Spatial Information Theory (COSIT 2017)*, computer environments and urban systems, pages 88–99. Elsevier, 2012.
- 6 Ladina B. Tschander, Hedda R. Schmidtke, Carola Eschenbach, Christopher Habel, and Lars Kulik. A geometric agent following route instructions. In Christian Freksa, Wilfried Brauer, Christopher Habel, and Karl F. Wender, editors, *Spatial Cognition III*, volume 2685 of *Lecture Notes in Artificial Intelligence (LNAI)*, pages 99–111, Berlin, 2003. Springer.
- 7 Stephan Winter and Christian Freksa. Approaching the notion of place by contrast. *Journal of Spatial and Information Science (JOSIS)*, 5:31–50, 2012.

Scalable Spatial Join for WFS Clients

Tian Zhao

University of Wisconsin – Milwaukee, Milwaukee, WI, USA
tzhao@uwm.edu

Chuanrong Zhang

University of Connecticut, Storrs, CT, USA
chuanrong.zhang@uconn.edu

Zhijie Zhang

University of Connecticut, Storrs, CT, USA
zhijie.zhang@uconn.edu

Abstract

Web Feature Service (WFS) is a popular Web service for geospatial data, which is represented as sets of features that can be queried using the *GetFeature* request protocol. However, queries involving spatial joins are not efficiently supported by WFS server implementations such as GeoServer. Performing spatial join at client side is unfortunately expensive and not scalable. In this paper, we propose a simple and yet scalable strategy for performing spatial joins at client side after querying WFS data. Our approach is based on the fact that Web clients of WFS data are often used for query-based visual exploration. In visual exploration, the queried spatial objects can be filtered for a particular zoom level and spatial extent and be simplified before spatial join and still serve their purpose. This way, we can drastically reduce the number of spatial objects retrieved from WFS servers and reduce the computation cost of spatial join, so that even a simple plane-sweep algorithm can yield acceptable performance for interactive applications.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases WFS, SPARQL, Spatial Join

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.72

Category Short Paper

1 Introduction

OGC Web services such as Web Feature Service (WFS) and Web Map Service (WMS) provide standard Web-based protocols for querying geospatial features. WMS clients can use *GetMap* request to retrieve map images for a specified area and use *GetFeatureInfo* request to query the attributes of specified features. WFS clients can use *GetFeature* request to retrieve the feature instances including the geometries and other feature attributes. The retrieved features can be used by clients for computations such as spatial joins.

In query-based visual exploration, users rely on an interactive client application to locate data of interests, where spatial join is a commonly used operation to discover spatial relations between features on a map. Spatial join is a computationally intensive operation that is usually executed in a server such as PostGIS database. Previous studies have focused on improving response time at server side [6] while very few research is on improving performance at client side [5]. However, in some cases, it is preferable to perform spatial joins at client side. For example, to join two or more types of features located in different WFS servers, it is inefficient to retrieve one set of features from one server and send them to the second server for spatial join. Moreover, WFS servers may not even provide efficient implementation of



© Tian Zhao, Chuanrong Zhang, and Zhijie Zhang;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 72; pp. 72:1–72:6

Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

■ **Table 1** Selected features of high-definition hydrology dataset for HUC0204 – Delaware-Mid Atlantic Coastal sub-region in the Mid-Atlantic Water Resource Region.

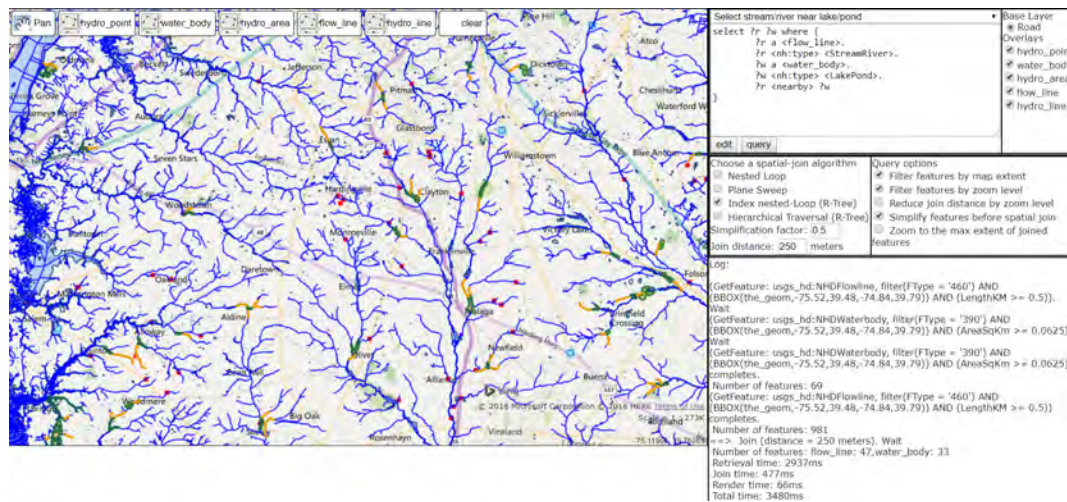
Feature Type	Geometry Type	Number of Features	Shapefile Size
NHDFlowline	MultiLineString	310835	312 MB
NHDWaterbody	MultiPolygon	57641	151 MB
NHDLine	MultiLineString	8090	3.7 MB
NHDArea	MultiPolygon	2592	113 MB
NHDPoint	Point	514	0.023 MB

spatial joins. For example, GeoServer implements spatial joins of two layers as a *GetFeature* request to the first layer where the join operation with the second layer is encoded in the filter of the request. This is similar to the *nested-loop* join [4], which loads all features in the server memory and performs spatial join on each pair of features in the two layers. This is inefficient. For example, to avoid using too much server memory, GeoServer restricts the number of features in the filter to be 1000 or less by default, which limits its ability of handling big spatial datasets.

Implementing efficient spatial join at WFS clients requires special care. WFS *GetFeature* requests can overwhelm both the server and the client when involving in big spatial datasets. For example, the hydrology dataset shown in Table 1 contains features (e.g. flowline) over 300 megabytes (MB). If we make a request to retrieve all feature instances of the flowline layer, then either the WFS server will fail to respond or the browser that runs the WFS client will quit working due to memory exhaustion. Note that while WMS can build and return maps of a large number of features such as the aforementioned flowline, WFS needs to encode all feature data in a *GetFeature* response, which can severely strain the memory capacity of the server. The WFS client, which often is the Web or mobile browser, will also become overwhelmed by the amount of memory and computation workload that are required to decode the response, store the spatial objects, and display them on a map. In addition, transmitting hundreds of MB of data across network consumes time and bandwidth. Finally, even if the server and client can process the *GetFeature* requests without crashing, spatial join can still take exceedingly long time, which is unsuitable for an interactive application.

While it is possible to improve the runtime of spatial join by implementing more efficient spatial join algorithms, there is a limit on how much improvement one can make. The query response time for a WFS *GetFeature* request includes the query processing time at WFS server, network transmission time of the query response, decoding time of the response, and computation time of spatial join. Improving performance on spatial join alone will not be sufficient if the network time and server time are still significant. In addition, WFS clients are often implemented as dynamic Web pages running in browsers, where the join operation is implemented in JavaScript that runs as a single-threaded program. There is a limited opportunity to improve spatial join performance through parallelization.

Although many spatial join algorithms have been proposed in literature, very few studies investigated performance of spatial join query on the client side over the Web or mobile browser. Based on our best knowledge, there is no study to compare performance of different spatial join algorithms for WFS, not to say in the context of Geospatial Semantic Web. To address the above efficiency problem with spatial join at WFS clients, we propose an approach that leverages the fact that users of the WFS clients are mostly interested in features of the current map extent and zoom level by not retrieving irrelevant features before performing spatial joins. In addition, retrieved features are cached to reduce network traffic



■ **Figure 1** WFS query interface.

and server load and geometry simplification is used to improve the efficiency of evaluating spatial relations. Alternative spatial join algorithms are evaluated. For index-based join, spatial indices generated online are cached to reduce runtime costs.

For the rest of the paper, we first explain our approach in Section 2, then we evaluate its performance in Section 3, and we discuss the result in Section 4.

2 Approach

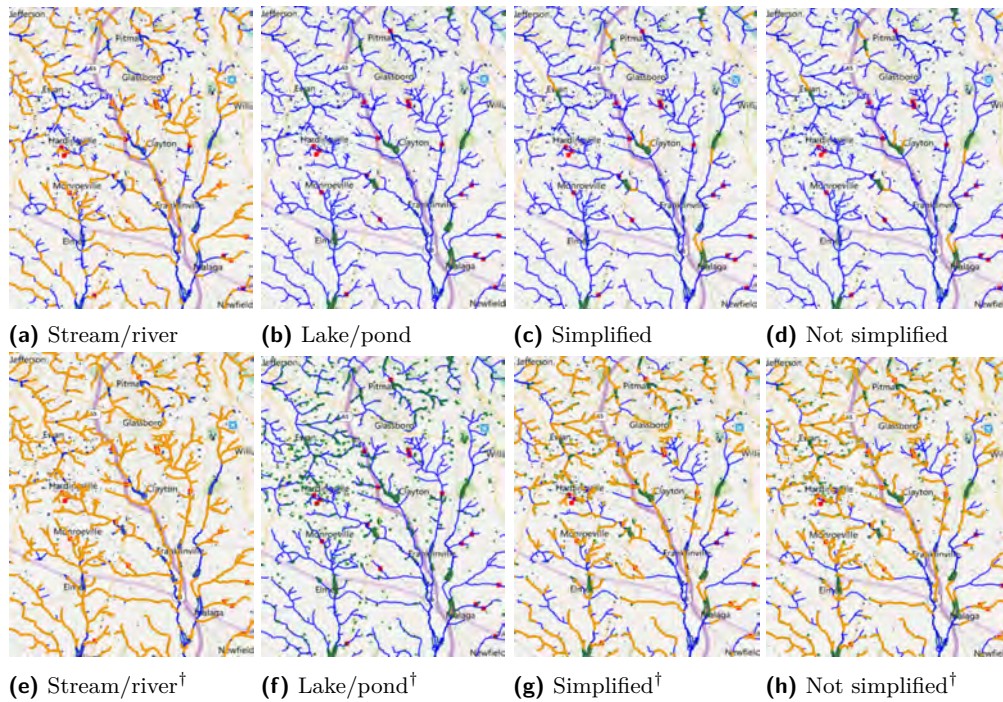
This WFS query client is an extension of our prior work on RDF query interface for WFS data [8, 7]. The spatial query is written in SPARQL-like syntax, which is translated to WFS requests and spatial join operations. A configuration file is used to map the WFS feature types and attributes to RDF classes and properties. Furthermore, certain attribute values are mapped to more recognizable constants for the convenience of writing queries. The query interface is shown in Figure 1 and the application is available at <http://tianpar.cs.uwm.edu:8080/usgs>.

The interface has an option to automatically insert spatial filters based on the current map extent so that features beyond the current map extent will not be retrieved. It also has an option to insert spatial filters based on the current zoom level so that features with sizes that are smaller than a threshold will not be retrieved. The threshold is calculated based on an adjustable scale proportional to the current zoom level. In order to perform spatial filtering based on feature size, we encode the attribute information of a feature type used to represent sizes in the configuration file.

For example, the query (Q1) below retrieves streams/ivers near lakes/ponds.

```
select ?r ?w where {
  ?r a <flow_line>.
  ?r <nh:type> <StreamRiver>.
  ?w a <water_body>.
  ?w <nh:type> <LakePond>.
  ?r <nearby> ?w
}                                     (Q1)
```

In this query, the predicate `<nearby>` specifies a spatial join between the variables `?r` and `?w`, which refer to features of streams/ivers and lakes/ponds respectively. The adjustable distance of `<nearby>` is specified separately in the query interface. This query is translated to the following concrete actions.



■ **Figure 2** Part of selected streams/ivers (in yellow) and lakes/ponds (in green) with or without size filters (marked with †) and the join result with or without simplification.

```
(GetFeature: usgs_hd:NHDFlowline,
  filter(FType = '460') AND (BBOX(the_geom, -75.52, 39.48, -74.84, 39.79))
  AND (LengthKM >= 0.5)).
(GetFeature: usgs_hd:NHDWaterbody,
  filter(FType = '390') AND (BBOX(the_geom, -75.52, 39.48, -74.84, 39.79))
  AND (AreaSqKm >= 0.0625)).
Join (distance = 250 meters).
```

The extent and size filters are inserted automatically by the query client into the generated GetFeature requests. The extent filter `BBOX(the_geom, -75.52, 39.48, -74.84, 39.79)` is derived from the current map extent (as in Figure 1). The size filter is derived from the current zoom-level and related to size attribute of each feature type. For streams/ivers, size filter is $\text{LengthKM} \geq 0.5$ and for lakes/ponds, the size filter is $\text{AreaSqKm} \geq 0.0625$.

We have implemented four spatial join algorithms: nested-loop join [4], plane sweep [1], index nested-loop join [3], and hierarchical traversal [2], where the latter two use R-tree indexing. Before spatial join, complex geometries (multi-line-strings and multi-polygons) are simplified based on a tolerance value proportional to the join distance. The implementation of nested-loop join is optimized by filtering candidate pairs using their bounding boxes.

3 Evaluation

We evaluated the performance of the four spatial join algorithms implemented in JavaScript running in Chrome browser. We used the query (Q1) to select the streams/ivers (NHDFlowline) that are within 250 meters of lakes/ponds (NHDWaterbody). Figure 2 shows some of the streams/ivers and lakes/ponds with or without size filters and the corresponding join results. The right two maps of each row are the join results with or without simplification. The maps on the second row (without size filters) are cluttered with features (especially

■ **Table 2** Numbers of features after join, written as (# of streams/ivers, # of lakes/ponds).

Algorithm	Nested loop	Plane sweep	Index nested-loop	Hierarchical traversal
Filter by extent	(3192, 1764)	(3192, 1764)	(3192, 1764)	(3192, 1764)
Filter by extent & Simplify	(2865, 1514)	(2911, 1649)	(2865, 1514)	(2865, 1514)
Filter by extent/size	(51, 36)	(51, 36)	(51, 36)	(51, 36)
Filter by extent/size & Simplify	(47, 33)	(46, 33)	(47, 33)	(47, 33)

■ **Table 3** Number of retrieved features and runtime (in seconds) for data retrieval, rendering results, and computing geometry bounds (included in the runtime of spatial join).

	Stream/River	Lake/Pond	Retrieval	Render	Bounds
Filter by extent only	8291	2757	11.6 s	0.85 s	0.99 s
Filter by extent and size	981	69	2.9 s	0.07 s	0.32 s

lakes/ponds) that are too small for visual exploration. From the figure it can be seen that, the join results with or without simplification ((c) vs. (d) and (e) vs. (f)) do not show obvious visual differences.

To measure the accuracy of various query options, we report in Table 2 the number of joined features that are with or without size filters and with or without simplification (with the tolerance of 125 meters). From Table 2, it can be seen that all four algorithms report similar results. The only exception is the *plane sweep* algorithm when the feature geometries are simplified. This difference is due to the combined effect of the simplification and the order of comparison of the join algorithms. Without simplification, all four algorithms report the same results. Simplification also reduces the number of joined features moderately.

Table 3 shows the number of retrieved features with or without size filters and the corresponding runtime for data retrieval, rendering results, and calculating geometry bounds. Table 4 shows the runtime of the four join algorithms for query (Q1) that are with or without size filters and with or without simplification. The runtime includes one-time costs such as calculating geometry bounds, simplification, spatial indexing (for index nested-loop join and hierarchical traversal), and sorting (for plane sweep). The costs are one-time since the bounds or indices are stored with the cached features and if the next user query uses cached data, such costs will not be repeated. Since these one-time costs are significant portions of the join time, for queries that can find data in the cache, the join time is much lower.

Caching reduces runtime cost even for queries that share some of the data. For example, if we first run the below query (Q2) and then run (Q1) with the same extent and size filters, the execution of (Q1) will be much faster because the features of streams/ivers will be in cache where spatial indices and geometry bounds have already been computed.

```
select ?r ?p where {                                     (Q2)
  ?r a <flow_line>.
  ?r <nh:type> <StreamRiver>.
  ?p a <hydro_point>.
  ?r <nearby> ?p }
```

In this case, the query (Q1) only needs to send a WFS request to retrieve lakes/ponds features and to perform spatial join. The runtime cost of (Q1) (with simplification) reduces from 3.5 s to about 1.7 s (1.5 s for data retrieval, 0.16 s for index nested-loop join – 0.12 s of which is for computing geometry bounds of lakes/ponds, while rendering is still 0.07 s).

■ **Table 4** Runtime (in seconds) of spatial join (including one-time costs such as calculating geometry bounds, simplification, indexing, and sorting).

Algorithm	Nested loop	Plane sweep	Index nested-loop	Hierarchical traversal
Filter by extent	73.2 s	2.87 s	2.5 s	2.74 s
Filter by extent & Simplify	49.7 s	1.4 s	1.4 s	1.37 s
Filter by extent/size	3.9 s	0.76 s	0.75 s	0.72 s
Filter by extent/size & Simplify	0.65 s	0.42 s	0.46 s	0.456 s

4 Discussion and Conclusion

This work evaluates optimization strategies for spatial join queries on client browser from distributed WFS servers. Our strategy is to automatically apply spatial filters based on map extent and feature size. The extent filter removes features beyond the currently viewed map while size filters remove features too small for the current zoom level. This kind of filters are suitable for the purpose of visual exploration. The results show the importance of spatial filtering in achieving acceptable query performance. As shown in Tables 3 and 4, the time for feature retrieval (11.6 s) dominates the time for spatial join and rendering if size filters are not applied. Even with size filters, the feature retrieval time (2.9 s) is still the largest component of the query time but at least it is within an acceptable range (3.5 s), which can be much lower if some or all of the queried data is cached. The results also show that naive implementation of spatial join (e.g. nested loop) scales poorly with the large number of features. Plane sweep, index nested-loop, and hierarchical traversal have similar performance, which makes plane-sweep a better choice due to its simplicity. Finally, the results show that geometry simplification can greatly reduce spatial join time, especially for features such as waterbody that can have tens of thousands of points in a geometry.


References

- 1 L. Arge, O. Procopiuc, S. Ramaswamy, T. Suel, and J. S. Vitter. Scalable sweeping based spatial join. In *Proceedings of the 24th International Conference on Very Large Data Bases (VLDB)*, pages 570–581, New York, August 1998.
- 2 T. Brinkho, H.-P. Kriegel, and B. Seeger. Efficient processing of spatial joins using rtrees. In *Proceedings of the ACM SIGMOD Conference*, pages 237–246, May 1993.
- 3 R. Elmasri and S. B. Navathe. *Fundamentals of Database Systems*. Addison-Wesley, Reading, MA, third edition edition, 2000.
- 4 P. Mishra and M. H. Eich. Join processing in relational databases. *ACM Computing Surveys*, 24(1):63–113, March 1992.
- 5 Cyrus Shahabi, Mohammad R. Kolahdouzan, and Maytham Safar. Alternative strategies for performing spatial joins on web sources. *Knowl. Inf. Syst.*, 6(3):290–314, May 2004.
- 6 J. Zhang, S. You, and L. Gruenwald. Towards GPU-accelerated Web-GIS for query-driven visual exploration. In *Proceedings of the 15th International Symposium on Web and Wireless Geographical Information Systems*, pages 119–136, Shanghai, China, May 2017.
- 7 T. Zhao, C. Zhang, and W. Li. Accessing distributed WFS data through a RDF query interface. In *Proceedings of GIScience*, 2016.
- 8 T. Zhao, C. Zhang, and W. Li. Adaptive and optimized RDF query interface for distributed WFS data. *International Journal of Geo-Information*, 6(4), 2017.

Modelling Spatial Patterns Using Graph Convolutional Networks


Di Zhu

Institute of Remote Sensing and Geographical Information Systems, Peking University, 5th Yiheyuan Road, Beijing, China
patrick.zhu@pku.edu.cn

 <https://orcid.org/0000-0002-3237-6032>

Yu Liu¹

Institute of Remote Sensing and Geographical Information Systems, Peking University, 5th Yiheyuan Road, Beijing, China
liuyu@urban.pku.edu.cn

 <https://orcid.org/0000-0002-0016-2902>

Abstract

The understanding of geographical reality is a process of data representation and pattern discovery. Former studies mainly adopted continuous-field models to represent spatial variables and to investigate the underlying spatial continuity/heterogeneity in a regular spatial domain. In this article, we introduce a more generalized model based on graph convolutional neural networks that can capture the complex parameters of spatial patterns underlying graph-structured spatial data, which generally contain both Euclidean spatial information and non-Euclidean feature information. A trainable site-selection framework is proposed to demonstrate the feasibility of our model in geographic decision problems.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases Spatial pattern, Graph convolution, Big geo-data, Deep neural networks, Urban configuration

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.73

Category Short Paper

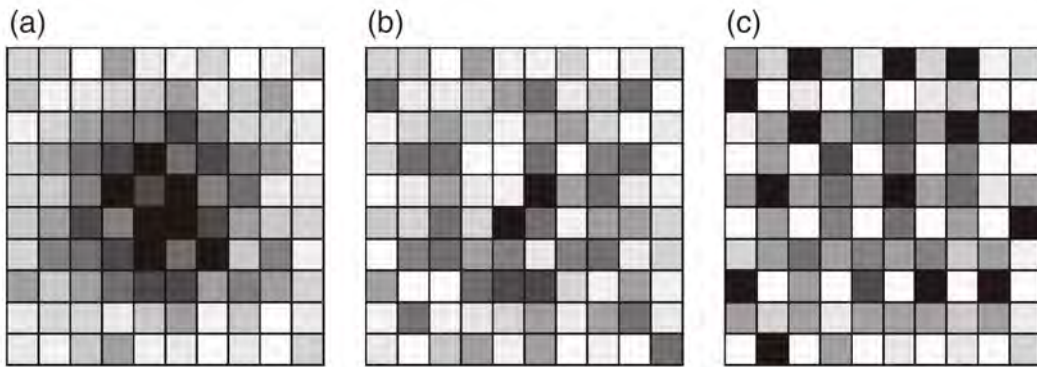
Funding This research was supported by the National Key Research and Development Program of China [Grant Number: 2017YFB0503602] and the National Natural Science Foundation of China [Grant Number: 41625003].

1 Introduction

The continuous-field model, which can be seen as a process of reducing the number of spatial variables required to represent reality to a finite set (a field) [6], is a fundamental perspective in modelling the complex geographical world. The variation of attributes in a field model represents the spatial pattern of certain geographical phenomenon at the conceptual level of abstraction [12, 7], as is shown in Figure 1. The analysis of spatial patterns based on field models has been studied extensively in traditional geography applications [2, 17]. Methods can be roughly divided into two types: autoregressive methods that adopt a spatial lag term

¹ Corresponding author





■ **Figure 1** Spatial patterns represented in a regular grid [5]. (a) Positive spatial autocorrelation. (b) Spatial randomness. (c) Negative spatial autocorrelation.

to consider the autocorrelation of local neighborhoods [1] and geostatistical methods that use semi-variograms to characterize the spatial heterogeneity [15, 2].

To uncover the deep features of spatial patterns, convolutional neural networks (CNNs) have been introduced from computer science to investigate local stationary properties of the input data by allowing long range interactions in terms of shorter, localized interactions [11]. However, the use of CNNs becomes problematic when the data is not structured in the regular spatial domain (e.g. raster model in GIS), since the local kernel filter can no longer be defined via the Euclidean metric of the grid. Graph convolutional networks (GCNs) is a generalization of CNNs to deal with graph-structured data in the irregular spatial domain (i.e., vector model in GIS), where the input data is represented as objects and their connections. The convolutional filter in GCNs can be extended to be localized in the spectral domain of the objects' features [3, 9], thus enable us to investigate both short range interactions and long range interactions in the spatial domain. We think that GCNs are suitable for modelling the complex spatial patterns in geographical data that generally contain both Euclidean spatial information and non-Euclidean feature information [13].

In this article, we will introduce a way to model the spatial patterns in geographical data by constructing a graph neural network with both spatial information and feature information embedded and by designing a localized feature filter on graph that considers spatial constraints. A layer-wise neural network framework is proposed to make the model trainable. In addition, we have applied the proposed framework in an intra-urban site-selection cases based on a POI check-in dataset in Beijing, China to demonstrate the feasibility of our model.

2 Embedding spatial patterns in graphs

2.1 Graph Fourier transformation

To enable the formulation of fundamental operations such as filtering on a graph, the Graph Fourier transform is needed first, which is defined via a generalization of the Laplacian operator on the grid to the graph Laplacian [4]. In graph $G = (V, E, W)$, V is a finite set of $|V| = n$ nodes, E is a set of edges among nodes and $W \in \mathbb{R}^{n \times n}$ is a weighted adjacency matrix representing the weights of edges. An input vector $x \in \mathbb{R}^n$ is seen as a signal defined on G with x_i denotes the spectral information of node i .

► **Definition 1** (Graph Laplacian). Let $L = \Delta - W$ be the graph Laplacian of G , where $\Delta \in \mathbb{R}^{n \times n}$ is a diagonal matrix with $\Delta_{ii} = \sum_j W_{ij}$, and the normalized definition is $L^s = I_n - \Delta^{-1/2}W\Delta^{-1/2}$ where I_n is the identity matrix.

As L^s is a real symmetric positive semidefinite matrix, it has a complete set of orthonormal eigenvectors $U = (u_1, \dots, u_n)$, and their associated nonnegative eigenvalues $\lambda = (\lambda_1, \dots, \lambda_n)$. The Laplacian is diagonalized by U such that $L^s = U\Lambda U^T$ where $\Lambda = \text{diag}([\lambda_1, \dots, \lambda_n]) \in \mathbb{R}^{n \times n}$. The graph Fourier transform of $x \in \mathbb{R}^n$ is then defined as $\hat{x} = U^T x \in \mathbb{R}^n$.

2.2 Convolutions on graphs

► **Definition 2** (Graph convolutions). The convolution operators on graphs are defined as the multiplication of x with a filter $g_\theta = \text{diag}(\theta)$ parameterized by $\theta \in \mathbb{R}^n$ in the Fourier domain, i.e.:

$$g_\theta \star x = g_\theta(L^s)x = g_\theta(U\Lambda U^T)x = Ug_\theta(\Lambda)U^T x. \tag{1}$$

We can understand $g_\theta(\Lambda)$ as a function of the eigenvalues of L^s , a non-parametric filter whose parameters are all free and can be trained.

However, the evaluation of Eq. 1 is computationally expensive, as the multiplication with eigenvector matrix U is $\mathcal{O}(n^2)$. To overcome this problem, [8] suggested the Chebyshev polynomials $T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$ up to K^{th} order to approximate $g_\theta(\Lambda)$:

$$g_{\theta'}(\Lambda) \approx \sum_{k=0}^K \theta'_k T_k(\tilde{\Lambda}), \tag{2}$$

with a rescaled $\tilde{\Lambda} = \frac{2}{\lambda_{max}}\Lambda - I_n$, $\theta' \in \mathbb{R}^K$ is a vector of polynomial coefficients, $T_0(x) = 1$ and $T_1(x) = x$.

Furthermore, by assuming $K = 1$ and $\lambda_{max} = 2$ in Eq. 2 and some renormalization tricks, [10] proposed an expression with a single parameter $\theta = \theta'_0 = -\theta'_1$ to compute:

$$g_\theta \star x \approx \theta(I_n + \Delta^{-1/2}W\Delta^{-1/2})x = \theta\tilde{\Delta}^{-1/2}\tilde{W}\tilde{\Delta}^{-1/2}x, \tag{3}$$

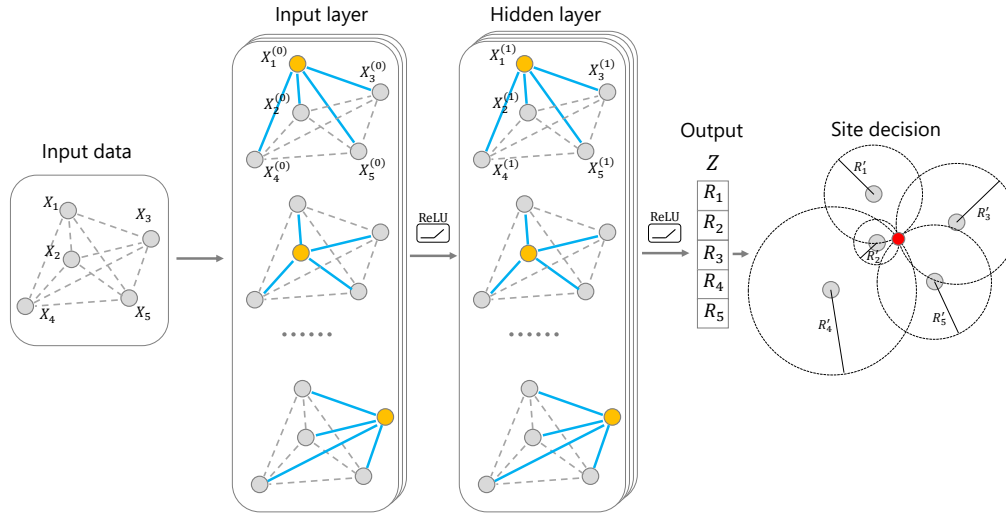
where $\tilde{W} = W + I_n$ and $\tilde{\Delta}_{ii} = \sum_j \tilde{W}_{ij}$. Eq. 3 has complexity $\mathcal{O}(|E|)$ because $\tilde{W}x$ can be efficiently implemented as a product of a sparse matrix with a dense vector.

2.3 Spatial-enriched graph construction

Different from state-of-the-art graph constructions in many recognition tasks, where the adjacency matrix W are often defined by calculating the similarity among nodes, we try to enable the constructed graph to capture the relationships between the feature similarity and the spatial displacement of node pairs, i.e., to construct a spatial-enriched graph.

Given the input features $X \in \mathbb{R}^{N \times C}$ of nodes V , where $N = |V|$ is the number of locations and $C \in \mathbb{R}$ is the number of features for each node, we define the adjacency matrix W according to spatial displacement of N locations. The distance matrix for locations can be considered a prior knowledge for the graph construction process and we can introduce the distance decay effect in geography to represent the spatial dependence of features in X . Derived from the gravity model, there many functions that could be used to express the spatial weighting function, such as the power function, the exponential function, and the Gaussian function [19]. Here, we consider a variant of the self-tuning Gaussian diffusion kernel [9]:

$$W_{ij} = \exp\left(-\frac{d(i,j)}{\sigma_i \sigma_j}\right), \tag{4}$$



■ **Figure 2** Illustration of the site-selection framework based on graph convolutional networks.

where $d(i, j)$ is the Euclidean distance between node i and j and σ_i is computed as the distance $d(i, i_k)$ corresponding to the k -th nearest neighbor i_k of node i . Eq. 4 gives a normalized measurement of spatial displacement in a graph whose variance is locally adapted around each location.

Compared to traditional geographical studies that choose arbitrary models to capture the effect of distance, our GCN-based model is a more universal way to model the relationship underlying spatial data. We treat the feature information and the spatial information separately, and leave the graph to learn the spatial pattern given certain training objective. The details of learned spatial pattern are restored in the layer-wise parameters of the deep graph convolutional network and can be adopted in various applications.

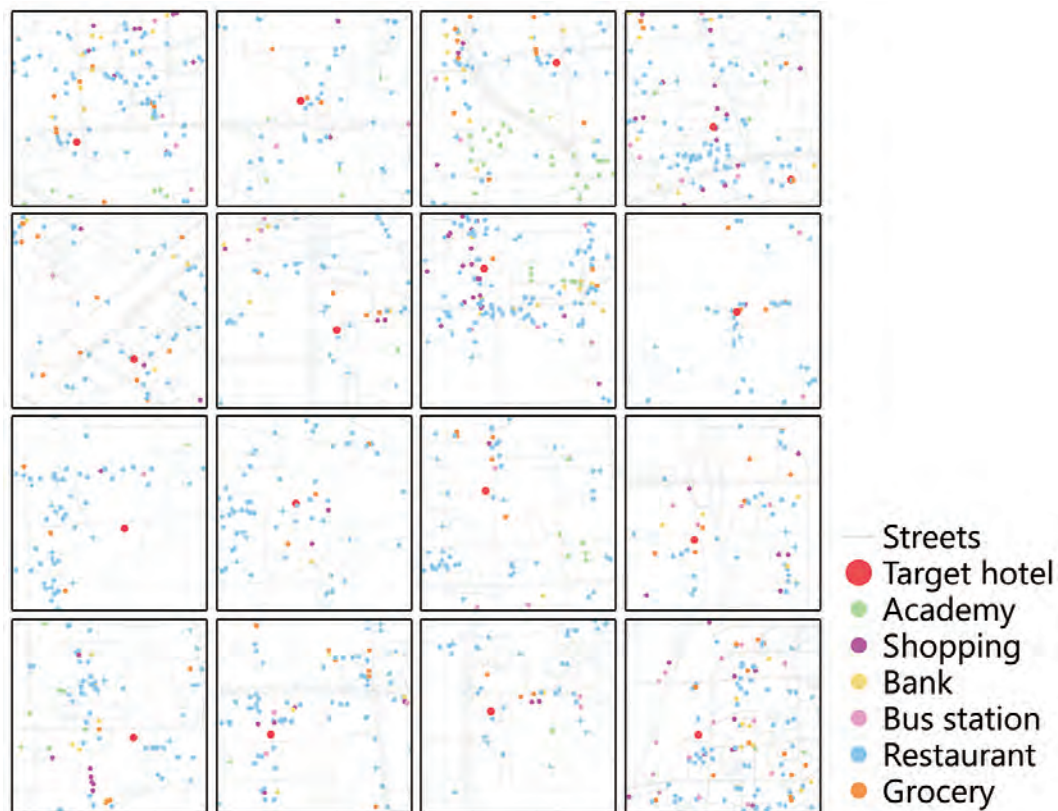
3 Example: site-selection tasks

One of the most common applications that implicitly consider spatial patterns is to find the best location to site a facility given the urban configurations. Traditionally, there are lots of studies that tried to solve this kind of site-selection problem through an spatial optimization model that considers some predefined spatial constraints [18]. However, if the model is simple and easy to compute, the optimization may be arbitrary to some extent; while if the model is too specific about the complex spatial relationships, the optimization are always difficult to compute.

Based on the graph convolutional model proposed in Section 2 that can learn the heterogeneity pattern underlying spatial data, we design a trainable neural network framework for the site-selection problem, illustrated in Figure 2. The site-selection framework is an example to show how our graph convolutional model can be adopted in geographic decision problems.

In Figure 2, the goal of the neural networks is to learn a complex function of spatial pattern on a graph $G = (V, E)$, which takes as input:

- A feature matrix $X \in \mathbb{R}^{N \times C}$ that contains the features x_i for every observed location i , where N is the number of given locations and C is the number of input feature types



■ **Figure 3** Illustration of some input training samples with only six POI types visualized. There are actually 242 POI types in total, and the multi-channel features contained in our dataset are not shown in this figure, such as the check-in number of each facility, the area of each facility, the number of photos took at each location.

- A fully-connected spatial distance matrix $W \in \mathbb{R}^{N \times N}$ summarized using Eq. 4 that represents the spatial structure of observed locations and outputs a decision vector $Z = [R_1, \dots, R_N] \in \mathbb{R}^N$ that contains the distances between the optimal site and all given locations. By calculating the virtual decision vector $Z' = [R'_1, \dots, R'_N] \in \mathbb{R}^N$ for all potential locations in the area, we can find an optimal site that minimize $\|Z - Z'\|$ or we can reject a proposal of site-selection given a distance threshold.

For simplicity, we display a simple two-layer GCN to capture the spatial dependence among urban locations and make prediction. Recalling the convolutional filter introduced in Eq. 3, let $\widehat{W} = \tilde{\Delta}^{-1/2} \tilde{W} \tilde{\Delta}^{-1/2}$, the forward propagation then takes the simple form:

$$Z = \text{ReLU} \left(\widehat{W} \text{ReLU} \left(\widehat{W} X \Theta^{(0)} \right) \Theta^{(1)} \right), \quad (5)$$

where $\Theta^{(0)} \in \mathbb{R}^{C \times H}$ is the input-to-hidden parameters for a hidden layer with H feature maps. $\Theta^{(1)} \in \mathbb{R}^{H \times 1}$ is the hidden-to-output parameters for an output decision vector Z .

Assuming all the existed facilities in urban areas are successful samples of site-selection given their circumstances, we then backpropagate the model with the mean square error loss function (MSELoss) between the output decision vector Z and the real location vector Z^* . Computational skills such as stochastic gradient descent, batch normalization and activation functions are all adopted in our work to train the model.

We utilized a dataset collected from Sina Weibo in 2014 that contains 868 million check-in records for 143,576 points of interest (POIs) in Beijing [14]. The dataset contains multiple features to form the multi-channel enriched feature matrix X as our model input. By randomly capturing 28,000 snapshots ($3km \times 3km$) that contain at least one built-up hotel as our input training samples and 7000 snapshots with the same settings as our validation set, we can adopt the framework in Figure 3 to train a network that tries to learn the function of spatial configurations between hotels and their complex urban environment. The original input training samples are shown in Figure 3. The trained network can thus be used to evaluate the built-up environment and decide where to build a potential hotel. In practice, methods of patch extraction and normalization are applied to make the input training samples comparable and combinable [16]. We formalized the comparable training graphs into minibatches without the information of target hotels, but record the ground truth decision vectors Z^* of each input sample for the calculation of MSE Loss.

Currently, we are still optimizing the experiment for this site-selection task. After more than 200 epochs of training, the average prediction accuracy on the validation set (7,000 samples) can reach around 50 meters, but the result is not very stable due to the abnormally complex POI configurations in Beijing, China. However, we believe the simple framework proposed in this section casts light on the applications of graph convolutions in geographic decision systems.

4 Conclusion and Discussion

In this article, we introduced a generalized model that can capture the spatial pattern in geographical data using graph convolutional networks. By embedding the feature information and the spatial information separately into the graph network, and designing a feature-based localized filter on the graph, our model can learn both short and long range interactions among space and approximate the high-dimensional parameters of spatial patterns according to certain training objectives. Based upon that, we proposed a trainable site-selection framework using spatial-enriched graph convolutional neural networks to demonstrate the feasibility of our model to be adopted in various geographic problems.

Important open questions remain: How about universality of the graph convolutional networks, how could it be transferred to other applications directly? How to evaluate the model's parameters in a way that is both quantitative, interpretable and intuitive for geographical analysis? How to incorporate more understanding of spatial interactions into the graph-based model except for the distance decay? In addition, this initial work has only focused on the multi-features in a single dataset; a promising area is to integrate the features of multi-sourced geo-data such as street networks, remote sensing spectra and other social sensing datasets. An improved version of our model is needed to characterize and explain the intertwined spatial variation pattern in our complex geographic world. We plan to address these questions in on-going works.

References

- 1 Luc Anselin. Spatial data analysis with gis: An introduction to application in the social sciences. *Ncgia Technical Reports*, 1992.
- 2 Noel Cressie. The origins of kriging. *Mathematical geology*, 22:239–252, 1990.
- 3 Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *arXiv preprint*, page arXiv:1606.09375, 2016.

- 4 R. K. Chung Fan. *Spectral graph theory*. American Mathematical Society, 1997.
- 5 A Stewart Fotheringham and Peter A. Rogerson. *The SAGE handbook of spatial analysis*. SAGE, 2008.
- 6 Michael F. Goodchild. Geographical data modeling. *Computers & Geosciences*, 18(4):401–408, 1992.
- 7 Michael F. Goodchild, May Yuan, and Thomas J. Cova. Towards a general theory of geographic representation in gis. *International Journal of Geographical Information Science*, 21(3):239–260, 2007.
- 8 David K. Hammond, Pierre Vandergheynst, and Rémi Gribonval. Wavelets on graphs via spectral graph theory. *Applied & Computational Harmonic Analysis*, 30(2):129–150, 2009.
- 9 Mikael Henaff, Joan Bruna, and Yann Lecun. Deep convolutional networks on graph-structured data. *arXiv preprint*, page arXiv:1506.05163, 2015.
- 10 Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint*, page arXiv:1609.02907, 2017.
- 11 Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
- 12 Yu Liu, Michael F Goodchild, Qinghua Guo, Yuan Tian, and Lun Wu. Towards a general field model and its order in gis. *International Journal of Geographical Information Science*, 22(6):623–643, 2008.
- 13 Yu Liu, Xi Liu, Song Gao, Li Gong, Chaogui Kang, Ye Zhi, Guanghua Chi, and Li Shi. Social sensing: A new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers*, 105(3):512–530, 2015.
- 14 Ying Long and Xingjian Liu. How mixed is beijing, china? a visual exploration of mixed land use. *Environment & Planning A*, 45(12):2797–2798, 2013.
- 15 Georges Matheron. Principles of geostatistics. *Economic Geology*, 58(8):1246–1266, 1963.
- 16 Giannis Nikolentzos, Polykarpos Meladianos, Jean Pierre Tixier, Konstantinos Skianis, and Michalis Vazirgiannis. Kernel graph convolutional neural networks. *arXiv preprint*, page arXiv:1710.10689, 2017.
- 17 J. K. Ord and Arthur Getis. Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis*, 27(4):286–306, 1995.
- 18 Daoqin Tong and Alan T. Murray. Spatial optimization in geography. *Annals of the Association of American Geographers*, 102(6):1290–1309, 2012.
- 19 Di Zhu, Zhou Huang, Li Shi, Lun Wu, and Yu Liu. Inferring spatial interaction patterns from sequential snapshots of spatial distributions. *International Journal of Geographical Information Science*, 32(4):783–805, 2018.

